ENCODING AND NAVIGATING LINGUISTIC REPRESENTATIONS IN MEMORY

EDITED BY: Claudia Felser, Colin Phillips and Matthew Wagers
PUBLISHED IN: Frontiers in Psychology

frontiers Research Topics

frontiers

Frontiers Copyright Statement

© Copyright 2007-2017 Frontiers Media SA. All rights reserved.

All content included on this site, such as text, graphics, logos, button icons, images, video/audio clips, downloads, data compilations and software, is the property of or is licensed to Frontiers Media SA ("Frontiers") or its licensees and/or subcontractors. The copyright in the text of individual articles is the property of their respective authors, subject to a license granted to Frontiers.

The compilation of articles constituting this e-book, wherever published, as well as the compilation of all other content on this site, is the exclusive property of Frontiers. For the conditions for downloading and copying of e-books from Frontiers' website, please see the Terms for Website Use. If purchasing Frontiers e-books from other websites or sources, the conditions of the website concerned apply.

Images and graphics not forming part of user-contributed materials may not be downloaded or copied without permission.

Individual articles may be downloaded and reproduced in accordance with the principles of the CC-BY licence subject to any copyright or other notices. They may not be re-sold as an e-book.

As author or other contributor you grant a CC-BY licence to others to reproduce your articles, including any graphics and third-party materials supplied by you, in accordance with the Conditions for Website Use and subject to any copyright notices which you include in connection with your articles and materials.

All copyright, and all rights therein, are protected by national and international copyright laws.

The above represents a summary only. For the full conditions see the Conditions for Authors and the Conditions for Website Use.

ISSN 1664-8714
ISBN 978-2-88945-132-6
DOI 10.3389/978-2-88945-132-6

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view.
By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: **researchtopics@frontiersin.org**

ENCODING AND NAVIGATING LINGUISTIC REPRESENTATIONS IN MEMORY

Topic Editors:
Claudia Felser, University of Potsdam, Germany
Colin Phillips, University of Maryland, USA
Matthew Wagers, University of California, Santa Cruz, USA

Compass on map

Image available at: https://www.pexels.com/photo/compass-map-navigation-navigation-device-269689/ under the CC0 license

Successful speaking and understanding requires mechanisms for reliably encoding structured linguistic representations in memory and for effectively accessing information in those representations later. Studying the time-course of real-time linguistic dependency formation provides a valuable tool for uncovering the cognitive and neural basis of these mechanisms. This volume draws together multiple perspectives on encoding and navigating structured linguistic representations, to highlight important empirical insights, and to identify key priorities for new research in this area.

Citation: Felser, C., Phillips, C., Wagers, M., eds. (2017). Encoding and Navigating Linguistic Representations in Memory. Lausanne: Frontiers Media. doi: 10.3389/978-2-88945-132-6

Table of Contents

07 ***Editorial: Encoding and navigating linguistic representations in memory***
Claudia Felser, Colin Phillips and Matthew Wagers

Section 1: Anaphor resolution

13 ***Reflexive anaphor resolution in spoken language comprehension: structural constraints and beyond***
Kaili Clackson and Vera Heyer

22 ***Retrieval interference in syntactic processing: the case of reflexive binding in English***
Umesh Patil, Shravan Vasishth and Richard L. Lewis

40 ***Retrieval interference in reflexive processing: experimental evidence from Mandarin, and computational modeling***
Lena A. Jäger, Felix Engelmann and Shravan Vasishth

64 ***Local anaphor licensing in an SOV language: ilmplications for retrieval strategies***
Dave Kush and Colin Phillips

76 ***Teasing apart retrieval and encoding interference in the processing of anaphors***
Lena A. Jäger, Lena Benz, Jens Roeser, Brian W. Dillon and Shravan Vasishth

94 ***The structure-sensitivity of memory access: evidence from Mandarin Chinese***
Brian Dillon, Wing-Yee Chow, Matthew Wagers, Taomei Guo, Fengqin Liu and Colin Phillips

110 ***The relationship between anaphor features and antecedent retrieval: comparing Mandarin* ziji *and* ta-ziji**
Brian Dillon, Wing-Yee Chow and Ming Xiang

126 ***Processing the Chinese reflexive "ziji": effects of featural constraints on anaphor resolution***
Xiao He and Elsi Kaiser

145 ***Wh-filler-gap dependency formation guides reflexive antecedent search***
Michael Frazier, Lauren Ackerman, Peter Baumann, David Potter and Masaya Yoshida

164 ***A fan effect in anaphor processing: effects of multiple distractors***
Kevin S. Autry and William H. Levine

179 ***Backward- and forward-looking potential of anaphors***
Petra B. Schumacher, Jana Backhaus and Manuel Dangl

193 ***Immediate sensitivity to structural constraints in pronoun resolution***
Wing-Yee Chow, Shevaun Lewis and Colin Phillips

209 ***The online application of binding condition B in native and non-native pronoun resolution***
Clare Patterson, Helena Trompelt and Claudia Felser

225 ***Structural constraints on pronoun binding and coreference: evidence from eye movements during reading***
Ian Cunnings, Clare Patterson and Claudia Felser

243 ***Active search for antecedents in cataphoric pronoun resolution***
Leticia Pablos, Jenny Doetjes, Bobby Ruijgrok and Lisa L.-S. Cheng

256 ***Interference in the processing of adjunct control***
Dan Parker, Sol Lago and Colin Phillips

269 ***On the shallow processing (dis)advantage: grammar and economy***
Arnout Koornneef and Eric Reuland

Section 2: Filler-gap dependencies

286 ***The localization of long-distance dependency components: integrating the focal-lesion and neuroimaging record***
Maria M. Piñango, Emily Finn, Cheryl Lacadie and R. Todd Constable

309 ***Cross-linguistic evidence for memory storage costs in filler-gap dependencies with wh-adjuncts***
Arthur Stepanov and Penka Stateva

328 ***An fMRI study dissociating distance measures computed by Broca's area in movement processing: clause boundary vs. identity***
Andrea Santi, Angela D. Friederici, Michiru Makuuchi and Yosef Grodzinsky

340 ***Discourse accessibility constraints in children's processing of object relative clauses***
Yair Haendler, Reinhold Kliegl and Flavia Adani

357 ***Using the visual world paradigm to study retrieval interference in spoken language comprehension***
Irina A. Sekerina, Luca Campanelli and Julie A. Van Dyke

372 ***The processing of raising and nominal control: an eye-tracking study***
Patrick Sturt and Nayoung Kwon

387 ***Structure modulates similarity-based interference in sluicing: an eye tracking study***
Jesse A. Harris

405 ***Filling the silence: reactivation, not reconstruction***
Dario L. J. F. Paape

423 ***Working memory differences in long-distance dependency resolution***
Bruno Nicenboim, Shravan Vasishth, Carolina Gattei, Mariano Sigman and Reinhold Kliegl

439 ***When high-capacity readers slow down and low-capacity readers speed up: working memory and locality effects***
Bruno Nicenboim, Pavel Logačev, Carolina Gattei and Shravan Vasishth

463 ***Hyper-active gap filling***
Akira Omaki, Ellen F. Lau, Imogen Davidson White, Myles L. Dakan, Aaron Apple and Colin Phillips

481 ***Thematic orders and the comprehension of subject-extracted relative clauses in Mandarin Chinese***
Chien-Jer Charles Lin

494 ***Filling predictable and unpredictable gaps, with and without similarity-based interference: evidence for LIFG effects of dependency processing***
Kimberly Leiken, Brian McElree and Liina Pylkkänen

510 ***Task-dependency and structure-dependency in number interference effects in sentence comprehension***
Julie Franck, Saveria Colonna and Luigi Rizzi

525 ***Corrigendum: Task-dependency and structure-dependency in number interference effects in sentence comprehension***
Julie Franck, Saveria Colonna and Luigi Rizzi

526 ***Syntactic constraints and individual differences in native and non-native processing of wh-movement***
Adrienne Johnson, Robert Fiorentino and Alison Gabriele

543 ***Similarity of* wh*-phrases and acceptability variation in* wh*-islands***
Emily Atkinson, Aaron Apple, Kyle Rawlins and Akira Omaki

559 ***The D-linking effect on extraction from islands and non-islands***
Grant Goodall

570 ***Distinctiveness and encoding effects in online sentence comprehension***
Philip Hofmeister and Shravan Vasishth

583 ***Elaboration over a discourse facilitates retrieval in sentence processing***
Melissa Troyer, Philip Hofmeister and Marta Kutas

Section 3: Computing agreement and feature-based encoding

592 ***Representing number in the real-time processing of agreement: self-paced reading evidence from Arabic***
Matthew A. Tucker, Ali Idrissi and Diogo Almeida

613 ***Gender agreement attraction in russian: production and comprehension evidence***
Natalia Slioussar and Anton Malko

633 ***Minimal interference from possessor phrases in the production of subject-verb agreement***
Janet L. Nicol, Andrew Barss and Jason E. Barker

645 ***Language processing as cue integration: grounding the psychology of language in perception and neurophysiology***
Andrea E. Martin

662 ***Grammatical number processing and anticipatory eye movements are not tightly coordinated in English spoken language comprehension***
Brian Riordan, Melody Dye and Michael N. Jones

673 ***An interference account of the missing-VP effect***
Jana Häussler and Markus Bader

Section 4: Other phenomena

689 ***Listeners exploit syntactic structure on-line to restrict their lexical search to a subclass of verbs***
Perrine Brusini, Mélanie Brun, Isabelle Brunet and Anne Christophe

697 ***Dependency resolution difficulty increases with distance in persian separable complex credicates: evidence for expectation and memory-based accounts***
Molood S. Safavi, Samar Husain and Shravan Vasishth

712 ***Poor readers' retrieval mechanism: efficient access is not dependent on reading skill***
Clinton L. Johns, Kazunaga Matsuki and Julie A. Van Dyke

732 ***Processing implicit control: evidence from reading times***
Michael McCourt, Jeffrey J. Green, Ellen Lau and Alexander Williams

748 ***Dependency-dependent interference: NPI interference, agreement attraction, and global pragmatic inferences***
Ming Xiang, Julian Grove and Anastasia Giannakidou

frontiers in Psychology

EDITORIAL
published: 09 February 2017
doi: 10.3389/fpsyg.2017.00164

Editorial: Encoding and Navigating Linguistic Representations in Memory

Claudia Felser[1]*, *Colin Phillips*[2, 3] and *Matthew Wagers*[4]

[1] Potsdam Research Institute for Multilingualism, University of Potsdam, Potsdam, Germany, [2] Department of Linguistics, University of Maryland, College Park, MD, USA, [3] Language Science Center, University of Maryland, College Park, MD, USA, [4] Department of Linguistics, University of California, Santa Cruz, Santa Cruz, CA, USA

Keywords: sentence comprehension, encoding, memory retrieval, interference, anaphor resolution, agreement processing, filler-gap dependencies

Editorial on the Research Topic

Encoding and Navigating Linguistic Representations in Memory

MOTIVATIONS

We created this research topic to address two closely related needs: to support a rapidly growing area of language science, and to support the (predominantly young) scientists who are working in this area. Recent years have seen a rapid growth in the amount of psycholinguistic research being carried out in linguistics departments. This has created a venue for exploring new questions. Understanding structured mental representations and the relations within them is the bread-and-butter of much research in linguistics, but the traditional focus has been on theories at a level of analysis that assumes discrete, symbolic representations, and is agnostic about how those representations are constructed in real time, whether in comprehension, production, or acceptability judgment tasks. Now there is a community of researchers who are working to understand these phenomena in more fine-grained terms.

In the area of syntax, the growth in research at the intersection of linguistics and psychology has been fueled by a number of parallel developments.

First, by connecting linguistic representations with psychological theories of memory encoding and access. The literature on memory encoding provides only limited inspiration for theories of structured linguistic representations, because most memory research is based on unstructured lists. But the literature on memory access has served as a strong inspiration for theories of linguistic dependency formation. In particular, models of content-addressable memory (CAM) have been influential in psycholinguistics. In CAM, items in memory are accessed (or their activation-level is boosted) based on their match to a set of content-based retrieval cues, rather than based on their memory address, as in classical computational architectures. A hallmark of memory access in CAM is similarity-based interference effects. These effects have been widely documented in language processing (Gordon et al., 2001; Van Dyke, 2007) and they feature prominently in many of the papers in the current collection. A second hallmark of memory access in CAM is non-effects of structural or linear distance in retrieval times (McElree et al., 2003), and these effects are the focus of one article in this collection (Dillon et al.). The influence of CAM on psycholinguistics has been aided by an implemented CAM-based parsing model (ACT-R: Lewis and Vasishth, 2005). This model makes specific, testable predictions, and provides a useful framework for thinking about memory access in language processing.

OPEN ACCESS

Edited and reviewed by:
Manuel Carreiras, Basque Center on Cognition, Brain and Language, Spain

***Correspondence:**
Claudia Felser
felser@uni-potsdam.de

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: *18 January 2017*
Accepted: *25 January 2017*
Published: *09 February 2017*

Citation:
Felser C, Phillips C and Wagers M (2017) Editorial: Encoding and Navigating Linguistic Representations in Memory. Front. Psychol. 8:164. doi: 10.3389/fpsyg.2017.00164

Second, research on the time course of linguistic processes is now more accessible, due to portable and affordable technical resources. Many studies can be carried out on laptop computers or even via the internet. Free statistical software packages are widely used, aided by a supportive user community. And one should not underestimate the value of role models that show that linguists can do this kind of research.

Together, the theoretical and practical developments have opened up a playground of languages and linguistic phenomena that can be used to develop and test models of real-time linguistic processes. The scale and diversity of the research in this collection was not possible 10–15 years ago.

However, publication venues have not kept pace with the growth in research at the intersection of linguistics and psychology. Researchers in this area still need to choose whether to submit their work to journals with a traditional linguistics focus or journals with a traditional psychology focus. For example, the *Journal of Memory and Language* is one of the most highly regarded journals in psycholinguistics. As of the start of 2017 its editorial staff and editorial board include more than 50 individuals, and there is almost no representation from linguistics. The associate editorial board of *Linguistic Inquiry*, an influential linguistics journal, has 70 members and just a couple of psycholinguists (To its credit, the new open access linguistics journal *Glossa* has a more diverse editorial team). This polarization means that there are few hospitable outlets for research that is unapologetic in its use of both linguistic and psychological models and analyses. Psychology journals routinely tell authors that their work "will not be understandable to our readers," even for relatively basic linguistic notions, especially if not from English. Notice the quaint idea that people read journals rather than articles. Linguistics journals are more likely to ask authors, "But how does this bear on theory?," where "theory" is assumed to mean claims at the traditional level of linguistic analysis, as if psycholinguists don't build theories. Of course, a number of papers have made their way into prominent journals in either field (e.g., Sturt, 2003; Phillips, 2006), but there is limited appetite for ongoing debates that delve into the details of real-time grammatical processes.

An additional benefit of editing this research topic is that we have been pleasantly surprised by the effectiveness of the Frontiers editorial process, which departs from tradition in a number of respects.

(i) As research topic editors, we could request 1-page abstracts before inviting a full paper submission. This allowed efficient triage of unsuitable submissions, and allowed us to make suggestions to authors at an early stage in the process.
(ii) The focus in the review process on soundness rather than impact was liberating for authors and reviewers alike, as they could engage more candidly with one another about what the results did and did not show. There has been much recent hand-wringing about transparency and replicability in psychological science (Open Science Collaboration, 2015). Most discussions have focused on how to avoid problems by legislating pre-registration, data sharing, specific analyses, etc. But it's just as important to remove the unrealistic expectations that create pressure on authors to distort their claims (Nosek et al., 2012; Maner, 2014). Our experience with this research topic is that authors are refreshingly open when the system does not penalize them for that.
(iii) The interactive review process helped to keep reviewers and authors on target and on time. The discussion-like format also made the process more collegial.

The success of the research topic is also evident in various metrics. As of January 2017, the 48 articles in this collection are the largest Frontiers research topic under the heading Psychology (out of 482 research topics) and Language Sciences (out of 53 research topics). One hundred and fifty-five authors contributed, as did 74 reviewers. This leaves little doubt that there is a market for publication venues at the intersection of linguistics and psychology. Also, the publication cycles were dramatically faster than other journals in linguistics and psycholinguistics. The median time from manuscript submission through two rounds of review (independent + interactive) to online publication was just over 4 months (128 days). For traditional journals the norm is 1–2 years or more. Of course, rapid publication is particularly valuable for junior researchers whose careers are hurt by extremely slow publication cycles. The speed of publication had the unexpected consequence that articles in this collection were cited at rates that are competitive with leading journals in linguistics, although this was not part of our original goal. We refer interested readers to a fuller discussion of our editorial experience elsewhere (Phillips, 2016).

Although we were encouraged by the number and quality of the articles and by the success of the review process, there were a couple of areas where we were disappointed. We hoped to see more articles on populations other than young adult native speakers, and we hoped to see more computational contributions. In both of these areas the lower demand may reflect the availability of other accessible and fast outlets. We also hoped to see more articles focusing on theory and synthesis. If this new field of research is to be sustainable, then it will need more than a large body of findings about specific linguistic phenomena in diverse languages. Without theoretical debates that serve to organize and guide research, the field will quickly run out of steam.

TOPICAL REVIEW

To review the individual contributions in our research topic, we have divided our collection roughly by linguistic phenomena, and consider the (overlapping) subsets of articles that deal with *anaphor resolution, filler-gap dependencies*, and *agreement*. These subsets, particularly the first two, address questions of how structured, compositional information is used in online sentence comprehension, and what kinds of features or similarity relations can aid or hinder comprehension. These themes—particularly the latter two—also consider how forward-looking the parser is, both in terms of expectations it explicitly commits to, but also in terms of how it encodes present information for future retrieval.

ANAPHOR RESOLUTION

Nineteen of the articles in the collection focus on anaphor resolution (And that does not include the articles on different forms of ellipsis). Encountering an anaphoric expression is thought to trigger a memory search for a suitable antecedent. Of the linguistic phenomena represented in this collection, anaphor resolution seems the most obviously suitable one for testing theoretical assumptions about memory access and retrieval, and how different types of cue interact in guiding the search process. Linguistic constraints on anaphor resolution such as Conditions A and B of the binding theory (Chomsky, 1981; Sportiche, 2013) seem to be at odds with some of the assumptions of well-motivated retrieval models. Most of the studies on anaphor resolution in this collection focus on interference; that is, on the question of whether anaphor resolution is affected by the presence of feature-matching distractors, and whether this is the case even for distractors that are structurally illicit antecedents. Interference can serve as a probe for memory access mechanisms. Whilst earlier studies investigating the role and timing of binding constraints during anaphor resolution mostly focused on English or English-type reflexives and pronouns (e.g., Nicol and Swinney, 1989; Badecker and Straub, 2002; Sturt, 2003), the articles in the current collection considerably expand the empirical research base by examining other languages and types of anaphora, including bound variable and long-distance anaphora.

A number of articles investigate the processing of reflexives or reciprocals. Using the visual-world paradigm, Clackson and Heyer find evidence for similarity-based interference during the processing of English reflexives, with listeners being distracted by a discourse-prominent but syntactically inaccessible antecedent. Patil et al. observe interference effects during the processing of English reflexives using eye-movement monitoring during reading, and Jäger et al. report interference effects during the processing of reflexives in Mandarin. No interference effects were observed by Kush and Phillips in the processing of pre-verbal anaphors (reciprocals) in Hindi, however. Jäger et al. report a series of studies on German and Swedish reflexives, where their findings suggest that interference affects retrieval but not encoding.

Dillon et al. provide experimental and modeling evidence for a locality bias for Mandarin long-distance reflexives, using a speed-accuracy tradeoff (SAT) paradigm, and Dillon et al. show this bias to be reduced for morphologically complex reflexives. Also examining Mandarin reflexives, He and Kaiser provide reading-time evidence showing that person features can block long-distance referential dependencies. Frazier et al.'s eye-movement results show that syntactic gaps (*wh*-traces) interact with reflexive resolution in English.

Several other contributions focus on non-reflexive pronouns. Looking at pronoun resolution across sentence boundaries, Autry and Levine demonstrate that multiple distractors give rise to cumulative ("fan") effects. Schumacher et al. examine the inter-sentential resolution of German pronouns and demonstratives using ERPs. Their results suggest that both semantic and positional cues contribute to a potential antecedent's referential prominence.

Investigating the role of binding constraints during pronoun resolution, Chow et al. present reading-time evidence which indicates that the antecedent search is constrained by Condition B. This conclusion is further supported by the findings reported by Patterson et al., whose participants also included non-native speakers of English. Unlike the native group, the non-native speakers showed a bias toward matrix subject antecedents, regardless of whether or not local coreference was allowed. The eye-movement results reported by Cunnings et al. show that c-command constrains pronoun binding by a quantificational antecedent, but not coreference between a pronoun and a non-quantificational antecedent. These findings are in line with what Sportiche (2013: 196) has dubbed "Condition D" of the binding theory. Pablos et al. investigate the processing of Dutch cataphoric (rather than anaphoric) pronouns using ERPs. Their results indicate that binding Condition C constrains the search for a suitable referent. The contribution by Parker et al. provides evidence for similarity-based interference during the computation of adjunct control dependencies, showing that an overt pronoun is not necessary.

Finally Koornneef and Reuland's "hypothesis and theory" article draws largely on findings from anaphor resolution studies. The authors argue that "deep" or grammatically driven processing is not necessarily computationally more costly than "shallow" processing using extra-grammatical information sources.

The studies on anaphora in this collection ultimately present a mixed picture: On the one hand, there is clear evidence of structure-based constraints guiding the antecedent search, while on the other hand referential dependency formation has been shown to be vulnerable to similarity-based interference under specific conditions. These seemingly contradictory findings can possibly be accounted within CAM-based processing models capable of implementing structure-sensitive constraints, and the specific way of capturing them is a focus of current debate.

FILLER-GAP DEPENDENCIES

A large number of articles in this collection examine the processing of filler-gap dependencies, with most of them focusing on *wh*-movement or relative clauses. Filler-gap dependencies are mediated by hierarchical phrase structure representations, and successfully completing them involves both memory storage and retrieval (e.g., Gibson, 1998). Encountering a filler such as *which student* in a *wh*-interrogative sentence like *Which student did you say you met at the concert last night?* is thought to trigger the prediction of a corresponding gap to which the filler must be linked before it can be fully integrated into the emerging sentence representation. Piñango et al. present brain imaging evidence showing that gap search and gap completion processes can be distinguished at the neurocognitive level.

As a filler needs to be kept in memory until a suitable gap can be identified, processing filler-gap dependencies can incur measurable storage costs. The difficulty of retrieving a filler (or antecedent) at a gap site may be affected by the nature of the sentence material that intervenes between antecedent and

gap, giving rise to interference effects. Stepanov and Stateva report cross-linguistic reading-time evidence for memory storage effects during the processing of *wh*-adjunct dependencies in both English and Slovenian, and Santi et al. present brain-imaging results which show that *wh*-dependency formation is affected by the syntactic type of the intervening sentence material. Using the visual-world paradigm, Haendler et al. show that German-speaking children's ability to comprehend object relative clauses is affected by referential properties of the intervening subject. In their Methods article, Sekerina et al. demonstrate that items held in short-term memory can interfere with filler retrieval during the auditory processing of object clefts, replicating earlier findings from reading-based tasks in a different modality. Using eye-movement monitoring during reading, Sturt and Kwon show that the processing of both subject raising and nominal control dependencies is subject to facilitatory interference effects. Also taking into account Parker et al.'s findings of interference effects during the computation of adjunct control dependencies, these studies indicate that antecedent retrieval at gap sites is generally vulnerable to interference.

This conclusion is further corroborated by the two studies in this collection that have investigated sluicing, a special type of clausal ellipsis that involves fronting of a remnant *wh*-expression (as in *He lost his keys but didn't know where*). The eye-movement data presented by Harris provide evidence that antecedent retrieval during the processing of sluiced sentences is subject to similarity-based interference modulated by structural properties of the antecedent. The contribution by Paape examines how the presence of a temporary subject/object ambiguity in the antecedent affects the processing of sluiced sentences in German.

Two further studies have examined effects of individual differences in working memory (WM) capacity on the processing of filler-gap dependencies. Nicenboim et al. present reading-time evidence for such effects from Spanish, and Nicenboim et al. report further evidence showing that locality effects in Spanish and German are modulated by WM capacity.

Several contributions focus on the nature of the gap search and how this search is constrained, or on the question of how dependency formation interacts with other grammatical computations. Omaki et al.'s findings show that the gap search process is highly predictive, with direct object gaps being postulated independently of verb transitivity even in a verb-medial language like English. The contribution by Lin focuses on the role of expectation during the processing of different types of subject relative clauses in Chinese, showing that canonical thematic ordering facilitates processing. Leiken et al. investigate the role of gap predictability in the processing of English object relatives (in comparison to verb-phrase ellipsis and right-node raising structures) using magnetoencephalography. Their results suggest that the left-anterior frontal gyrus (LIFG), a brain region previously found to be involved in dependency formation, subserves memory retrieval at gap sites regardless of whether or not the gap was predictable. Franck et al. use the phenomenon of agreement attraction to demonstrate that computing filler-gap dependencies involves the creation of abstract hierarchical phrase-structure representations, and Frazier et al. demonstrate that *wh*-gaps interact with the processing of reflexives. Engaging in an active gap search does not mean that gaps are postulated freely, however. The reading-time results reported by Johnson et al. suggest that neither native English speakers nor native Korean-speaking learners of English postulate gaps in so-called "island" environments.

Other studies examine how properties of the filler affect the processing of filler-gap dependencies. Atkinson et al. report a series of acceptability judgment experiments showing that morphosyntactic and semantic features interact in ameliorating *wh*-island violations, and Goodall shows that one of the factors known to ameliorate island violations ("d-linking") also improves the acceptability of non-island sentences. Hofmeister and Vasishth's reading-time results indicate that more complex fillers are easier to retrieve at gap sites than less complex ones, and the findings reported by Troyer et al. show that elaboration also facilitates filler retrieval across short pieces of discourse.

Taken together, the above studies provide strong evidence that retrieval at gap sites is vulnerable to similarity-based interference, that more complex or elaborate fillers are easier to retrieve than less complex ones, and that both gap postulation and filler retrieval are sensitive to information encoded in hierarchical phrase-structure representations.

COMPUTING AGREEMENT AND FEATURE-BASED ENCODING

Several articles in our collection address the phenomenon of agreement attraction. Agreement attraction is a robust perceptual illusion that mirrors a speech error in production (Bock and Miller, 1991). For example, speakers are prone to produce sentences like *The **dogs** [that the shelter **rescue** in the winter] eventually got adopted*. In our example, the agreement on the verb *rescue* should be controlled by its singular, grammatical subject, *the shelter* but instead it is attracted to agree with another nearby phrase *the dogs*—the attractor. Not only are such examples easily found in natural speech and elicited in the lab, but they are also routinely missed by language perceivers: speakers experience an illusion of grammaticality.

The illusion turns out to be very sensitive to the relationship between the features on the grammatical subject, the attractor noun, and the verb. For this reason, it has served as a productive system for probing issues around linguistic encoding: what are the features with which comprehenders represent partial linguistic information in their working memory? A common, well-motivated view about how nouns are encoded relies on feature markedness. Along any feature scale, certain values are more marked than others, like plural is more marked than singular: for example, plural nouns are less common than singular nouns and are usually signaled by more complex morphological forms (*dog-s* vs. *dog-Ø*). And it appears that these marked values are more visible in the comprehension system, in terms of how they are encoded during language processing (Eberhard, 1997). But several papers in our collection demonstrate that a more nuanced view is necessary, and they do so by looking at feature systems in under-investigated languages with grammars that have more complicated syntax/morphology

mappings, ones that are more amenable to investigating this question.

Tucker et al. show that in Modern Standard Arabic (MSA) the morphological exponence of a marked feature also matters. MSA expresses the plural feature in two ways, the so-called suffixed plural and the broken (or ablaut) plural. The suffixed plural causes more attraction and with a different time-course than the ablaut plurals in MSA. Slioussar and Malko examined a more complex feature system—gender—in Russian. The Russian gender system has three values, called masculine, feminine, and neuter. This three-way distinction makes it possible to demonstrate that, in determining whether attraction will occur, the visibility of the attractor alone is insufficient, and the visibility of the head is crucial. Nicol et al. also reinforce the conclusion that the head-attractor relationship matters, but from the perspective of hierarchical relatedness. Attractors that are closer to the head noun in the sentence's phrase structure induced more attraction effects. Moreover, how saliently a noun was marked as non-nominative mattered (e.g., *women's* generated less attraction than *dogs*'). Franck et al.'s paper on filler-gap dependencies likewise demonstrates that the relative syntactic prominence of the attractor is encoded.

Research on how plurals are encoded in real-time not only feeds-back into how plurality is represented in the grammar, but it also leads the way to broader questions of the relationship between *how* dependent elements are initially encoded and how retrieval cues are identified and integrated across time. These issues are taken up by Tucker et al. in their discussion of "feature-cue" algorithms in MSA. But Martin's Hypothesis article raises higher-level questions about how a theory of cue integration might relate different kinds of linguistic representation, such as how information signaled by distinct morphemes may be integrated into the percept of a phrase. Riordan et al. specifically consider the cue validity of Number in a broad variety of syntactic contexts; they demonstrate that number cues often generate quite weak predictions about numerosity, based on anticipatory looking in the visual world paradigm.

Hofmeister and Vasishth address the encoding/retrieval relation from a different angle: investigating the processing of relative clauses, they show that only syntactic and semantic elaboration of an left dependent (the RC head) affects retrieval at the right dependent (the RC verb)—but other differences experienced at encoding, like different text colors, do not. Troyer et al. show that such elaborative effects can happen as referents are processed over the span of a discourse, and not merely locally, i.e., not only when the elaboration occurs at the targeted retrieval position itself.

When there are evidently effects of similarity-based interference at a retrieval site, it is important to assess whether such effects derive from the process of retrieval itself or whether they might have arisen during the process of encoding. Along these lines, Häussler and Bader argue that interference at retrieval can explain the "missing VP" effect observed in the processing of center self-embedded relative clauses, even in languages like German which may benefit from highly predictive encoding mechanisms. Likewise, Jäger et al.'s paper on reflexives also takes up a phenomenon to argue that the culprit in any online fallibility is explicitly not encoding interference.

The remaining articles in this collection address issues relating to prediction, memory retrieval, or the role of linguistic structure in dependency formation by examining other linguistic phenomena. Brusini et al.'s contribution investigates verb prediction in French, providing evidence that syntactic cues constrain lexical access. In a series of reading-time experiments, Safavi et al. examine the processing of complex predicates in Persian, showing that dependency resolution difficulty is affected both by predictability and distance. Using an auditory speed-accuracy tradeoff paradigm, Johns et al. show that individual differences in reading skill do not affect memory retrieval during listening. McCourt et al. present reading-time evidence which challenges previous claims to the effect that the phenomenon of implicit control involves a silent syntactic argument. Xiang et al. show that susceptibility to interference during the processing of negative polarity items (NPIs) correlates with pragmatic reasoning as measured by an autism scale, whereas susceptibility to agreement attraction does not correlate. This indicates that NPI illusions have a different source than agreement attraction.

AUTHOR CONTRIBUTIONS

All authors listed have made substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

Our work on this Editorial has been supported by an Alexander-von-Humboldt Professorship to Harald Clahsen, by NSF grant DGE-1449815 to CP, and by NSF grant BCS-1251429 to MW.

REFERENCES

Badecker, W., and Straub, K. (2002). The processing role of structural constraints on the interpretation of pronouns and anaphors. *J. Exp. Psychol. Learn. Mem. Cogn.* 28, 748–769. doi: 10.1037/0278-7393.28.4.748

Bock, K., and Miller, C. (1991). Broken agreement. *Cogn. Psychol.* 23, 45–93. doi: 10.1016/0010-0285(91)90003-7

Chomsky, N. (1981). *Lectures on Government and Binding*. Dordrecht: Foris.

Eberhard, K. M. (1997). The marked effect of number on subject–verb agreement. *J. Mem. Lang.* 36, 147–164. doi: 10.1006/jmla.1996.2484

Gibson, E. (1998). Syntactic complexity: locality of syntactic dependencies. *Cognition* 68, 1–75. doi: 10.1016/S0010-0277(98)00034-1

Gordon, P. C., Hendrick, R., and Johnson, M. (2001). Memory interference during language processing. *J. Exp. Psychol. Learn. Mem. Cogn.* 27, 1411–1423. doi: 10.1037/0278-7393.27.6.1411

Lewis, R. L., and Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cogn. Sci.* 29, 375–419. doi: 10.1207/s15516709cog0000_25

Maner, J. K. (2014). Let's put our money where our mouth is: if authors are to change their ways, reviewers (and editors) must change with

them. *Perspect. Psychol. Sci.* 9, 343–351. doi: 10.1177/1745691614528215

McElree, B., Foraker, S., and Dyer, L. (2003). Memory structures that subserve sentence comprehension. *J. Mem. Lang.* 48, 67–91. doi: 10.1016/S0749-596X(02)00515-6

Nicol, J., and Swinney, D. (1989). The role of structure in coreference assignment during sentence comprehension. *J. Psycholinguist. Res.* 18, 5–20. doi: 10.1007/BF01069043

Nosek, B. A., Spies, J. R., and Motyl, M. (2012). Scientific Utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspect. Psychol. Sci.* 7, 615–631. doi: 10.1177/1745691612459058

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science* 349:aac4716. doi: 10.1126/science.aac4716

Phillips, C. (2006). The real-time status of island phenomena. *Language (Baltim.)* 82, 795–823. doi: 10.1353/lan.2006.0217

Phillips, C. (2016). *How to Create a Top Journal by Accepting (Almost) Everything.* Available online at: http://www.colinphillips.net/?p=3470

Sportiche, D. (2013). Binding theory – structure sensitivity of referential dependencies. *Lingua* 130, 187–208. doi: 10.1016/j.lingua.2012.11.005

Sturt, P. (2003). The time-course of the application of binding constraints in reference resolution. *J. Mem. Lang.* 48, 542–562. doi: 10.1016/S0749-596X(02)00536-3

Van Dyke, J. A. (2007). Interference effects from grammatically unavailable constituents during sentence processing. *J. Exp. Psychol. Learn. Mem. Cogn.* 33, 407–430. doi: 10.1037/0278-7393.33.2.407

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2017 Felser, Phillips and Wagers. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in
PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 19 August 2014
doi: 10.3389/fpsyg.2014.00904

Reflexive anaphor resolution in spoken language comprehension: structural constraints and beyond

Kaili Clackson [1] **** and Vera Heyer*** [2]

[1] *Department of Language and Linguistics, University of Essex, Colchester, UK*
[2] *Potsdam Research Institute for Multilingualism, University of Potsdam, Potsdam, Germany*

Edited by:
Colin Phillips, University of Maryland, USA

Reviewed by:
Kepa Erdocia, University of the Basque Country, Spain
Brian Dillon, University of Massachusetts, USA

****Correspondence:***
Kaili Clackson, Department of Language and Linguistics, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, UK
e-mail: hkscla@essex.ac.uk

We report results from an eye-tracking during listening study examining English-speaking adults' online processing of reflexive pronouns, and specifically whether the search for an antecedent is restricted to syntactically appropriate positions. Participants listened to a short story where the recipient of an object was introduced with a reflexive, and were asked to identify the object recipient as quickly as possible. This allowed for the recording of participants' offline interpretation of the reflexive, response times, and eye movements on hearing the reflexive. Whilst our offline results show that the ultimate interpretation for reflexives was constrained by binding principles, the response time, and eye-movement data revealed that during processing participants were temporarily distracted by a structurally inappropriate competitor antecedent when this was prominent in the discourse. These results indicate that in addition to binding principles, online referential decisions are also affected by discourse-level information.

Keywords: binding principle A, reflexive resolution, discourse prominence, sentence processing, eye-tracking

INTRODUCTION

According to most theoretical accounts, the interpretation of a reflexive is determined solely by a structural constraint which identifies a unique referent (Chomsky, 1981, 1986; Levinson, 1987; Pollard and Sag, 1992; Reinhart and Reuland, 1993; Reinhart, 2000, Reuland, 2001; Burkhardt, 2005 among others). For example, Principle A requires that an English argument reflexive is bound by a local antecedent that falls within its governing category, so that the anaphor and its antecedent are co-indexed (i.e., have compatible number, gender and person features), and the anaphor is c-commanded by its antecedent. In (1) *Susan* is structurally accessible as an antecedent as *Susan* binds (i.e., c-commands and is co-indexed with) *herself* and falls within the governing category of *herself* (shown by square brackets). *Jane* falls outside the governing category of *herself* and so is not structurally accessible as an antecedent.

(1) Jane_1 says that [Susan_2 hurt $\text{herself}_{*1/2}$].

In recent years there has been considerable discussion about the role that such structural constraints play in online sentence processing. Of particular interest is whether the parser's search for a referent is guided principally by structural considerations, where each potential antecedent is assessed based on its structural position; or whether a more cue-based search is implemented, where a structurally illicit referent that is strongly supported by other cues (such as being of appropriate gender and number, and in a prominent position) might be briefly considered and so lead to interference effects [for further discussion see Van Dyke (2007), Phillips et al. (2010), and Dillon et al. (2013) among others]. As the referent for a reflexive can be identified on the basis of structural information alone (in contrast to pronouns where structural information rules out certain referents, but does not necessarily identify a single referent), reflexive resolution is often seen as a good test case in this debate. In the present study we ask whether a noun phrase in a position where co-reference with the reflexive would violate a constraint, henceforth termed "inaccessible," [such as *Jane* in (1)] is ever considered by the parser as a potential referent. Results from previous research have pointed to somewhat differing conclusions, leaving this question unresolved.

For example, early cross-modal priming studies (Nicol, 1988; Nicol and Swinney, 1989) suggested that during reflexive resolution, the structural constraint acts as an early filter so that the adult parser only considers structurally accessible antecedents but not structurally inaccessible ones[1]. Evidence to support this has also come from studies using more time-sensitive measures such as ERPs and eye-tracking during listening (Xiang et al., 2009; Clackson et al., 2011) where no effects of the inaccessible antecedent were found[2]. In contrast, using a self-paced reading task Badecker and Straub (2002) found that reading times on the second word following the reflexive were significantly longer when the gender of the inaccessible antecedent matched that of the reflexive compared to when it did not, suggesting that the parser briefly considered the inaccessible antecedent as a potential antecedent. Furthermore, although results from eye-tracking during reading experiments are somewhat mixed, a number of studies have found tentative evidence that the inaccessible antecedent is not fully ruled out by Principle A. For example, Cunnings and Felser (2013) found that the gender of the inaccessible antecedent affected

[1] It should be noted that priming effects were only tested for at the point of the reflexive, not shortly after where effects have subsequently been found.

[2] In both experiments numerical trends suggested an effect, but these were non-significant in the statistical analysis.

reading times both at the reflexive region and text downstream of the reflexive, while Sturt (2003) found an effect in second-pass reading times on the reflexive and later regions[3]. While a number of studies have not found evidence of interference effects (e.g., Felser et al., 2009; Dillon et al., 2013) it is possible that such null results are due to particular properties of the materials used (see Discussion section), or stem from a lack of power to detect a relatively small effect [see Chen et al. (2012) for further discussion on power].

One difficulty in interpreting previous results is that it is not certain whether participants interpreted the reflexive correctly. If previous studies included comprehension questions, they were usually not aimed at the interpretation of the critical reflexive in order to avoid drawing participants' attention to the purpose of the experiment. Therefore, in most experimental paradigms there is no offline measure of the interpretation of the reflexive, making it impossible to know whether the observed results reflect successful processing of the reflexive or not. Indeed, one offline study showed that participants incorrectly interpreted a reflexive as referring to a gender matching but structurally inaccessible antecedent in 17% of cases (Sturt, 2003). Furthermore, a number of the studies above rely on gender stereotype nouns (such as *surgeon* being assumed to be male) to create "gender match" and "gender mismatch" conditions, and again it is impossible to know if participants interpreted such nouns in the manner intended.

The present eye-tracking during listening study avoids such difficulties by only using proper names for potential antecedents and by using a "goal-directed" design. The advantage of such a design is that the participant is required to identify the referent for the reflexive for each trial, thus allowing for separate analysis of eye movements and response times for trials where participants did, and did not, interpret the reflexive correctly. Trueswell (2008) supports such designs, arguing that eye movements reflect "goal-directed behavior" and that it is only possible to infer referential decisions from eye movements when these decisions are necessary to achieve the task at hand. The "goal-directed" design was chosen because a naturalistic design, with participants simply looking at pictures while listening to auditory stimuli, can lead to less data relevant to the research question due to participants not paying attention to the pictures at critical points. For instance, Clackson et al. (2011) investigated reflexive resolution using eye-tracking during listening by asking participants to listen to stimuli and answer general comprehension questions which did not probe the referent of the reflexive. One effect of this naturalistic task was that participants' attention was in no way drawn to the non-salient reflexive. As a result, in approximately half the trials participants did not look at any potential antecedent on hearing the reflexive, considerably reducing the quantity of relevant eye movement data collected. Therefore, it is possible that the observed numerical trend showing an effect of the inaccessible antecedent soon after hearing the reflexive (i.e., fewer looks to the accessible antecedent and more looks to the inaccessible antecedent when the inaccessible antecedent matched in gender with the reflexive) did not turn out to be statistically reliable due to the limited data collected.

In the present study the participants' task was presented as a "Who is it for?" activity where participants were asked to identify as quickly as possible which character in a story received a particular object. In experimental trials the recipient was identified by a reflexive. Gaze direction across a scene which included the participants in the story was monitored, so that three responses were recorded: accuracy of identifying the recipient character, response time, and gaze direction at the point of the crucial reflexive. If manipulation of the gender of the inaccessible antecedent (matching or mismatching the gender of the reflexive) affects responses, this interference effect would suggest that the inaccessible antecedent was briefly considered as a potential antecedent in the early stages of processing.

MATERIALS AND METHODS

PARTICIPANTS

Forty-two native speakers of English (mean age: 23, range: 18–48, 16 males) were recruited at the University of Essex and were paid for their participation. All participants had normal or corrected-to-normal vision.

DESIGN AND MATERIALS

The auditory materials were taken from the reflexive conditions used by Clackson et al. (2011) consisting of spoken pairs of sentences, each involving two characters from the set of Susan, Peter, Mr. Jones, and Mrs. White. The first sentence introduced the first character and established a suitable context for the second sentence, which included the second character, an inanimate object, and the critical reflexive. In each trial, the object was for, or was given to, the second character (the recipient), referred to by a reflexive. The auditory stimulus set comprised 24 experimental items, each appearing in two conditions. In the Double-Match condition the gender of both characters matched that of the reflexive, and in the Single-Match condition only the gender of the accessible antecedent matched that of the reflexive, as illustrated in (2).

(2) *Double-Match*
Peter was waiting outside the corner shop. He watched as Mr. Jones bought a huge box of popcorn for himself over the counter.

Single-Match
Susan was waiting outside the corner shop. She watched as Mr. Jones bought a huge box of popcorn for himself over the counter.

The inaccessible antecedent [*Peter* or *Susan* in (2)] is in a discourse prominent position as it is the first-mentioned character and the subject of both main clauses (repeated as a pronoun in the second one). The accessible antecedent (here: *Mr. Jones*), in contrast, is less salient as the subject of the subordinate clause.

Auditory stimuli were recorded using splicing to ensure that each version of an item was identical except for the name and pronoun changes necessary for the experimental manipulation.

[3]A further study reporting significant interference from an inaccessible antecedent in the processing of reflexives used eye-tracking during listening to investigate the interpretation of picture noun phrases (Runner et al., 2003). However, the authors concluded that reflexives in such contexts are in fact "logophors" and thus exempt from Binding Theory [see also Runner et al. (2006)].

Experimental items from a separate pronoun experiment were presented together with those from the present reflexive study, so that in addition to the reflexive experimental trials, each participant heard 24 pronoun items which mirrored the structure of the reflexive items, and 48 filler trials comprising a range of different grammatical constructions and featuring some additional characters (Doctor, Nurse, King, and Queen). Filler trials were similar to the experimental items in that the recipient of an object was introduced by a preposition (*for*, *to*, *on*, or *at*), but other properties were manipulated to provide variety of structure: the number of characters introduced before the preposition varied from one to three and, in contrast to the experimental items, the majority of filler items identified the recipient by name. This meant that contexts in which the recipient was only introduced after the preposition could be created, thus preventing participants from assuming that the recipient would always be mentioned early in the sentence. Furthermore, the point at which it became obvious which character received the object was varied in the filler items so that participants did not know when to expect the information which provided the answer to the task. For example, the recipient of the object is mentioned quite early in (3) but fairly late in (4) (object is underlined and recipient is shown in bold).

(3) At the hospital the nurse got a glass of water for **the doctor** because he had bad hiccoughs and needed to see a patient.

(4) After the accident in the royal carriage the King and the Queen were very upset. The doctor visited them and put a plaster on **the Queen's** nose where she had cut it.

Each auditory trial was accompanied by two visual displays as shown in **Figure 1**. A picture of the inanimate object was shown in the centre of the screen prior to the start of the auditory stimulus, and this was followed by the main visual display comprising four pictures: the inanimate object and three animate characters, which was viewed while the auditory stimulus was heard. For experimental trials, two of these characters were mentioned in the auditory stimulus and one (mismatching the gender of the reflexive) served as a distracter.

The four pictures were positioned in the corners of the screen, with a small cross in the center, and the positioning of the pictures of the characters and the inanimate object was counterbalanced across items. All pictures were black-and-white line drawings, of approximately the same size, and were not noticeably different in terms of visual saliency. All pictures were selected from a set of 520 pictures from the International Picture Naming Project (http://crl.ucsd.edu/~aszekely/ipnp/) for which various normed measures are available[4]. Experimental trials were arranged in four lists according to a Latin Square design (due to the similarity between the two reflexive conditions and two pronoun conditions from a separate experiment) so that each participant saw each trial in only one condition (Double-Match or Single-Match). The same set of filler trials was used with each list, and trials were presented in a pseudo-randomized order such that no more than two experimental trials occurred consecutively. To counteract any effects of fatigue, the four lists were then reversed to create eight lists in total so that items heard early in the experiment by one participant were heard late in the experiment by another. The study received ethical approval from the University of Essex ethics committee.

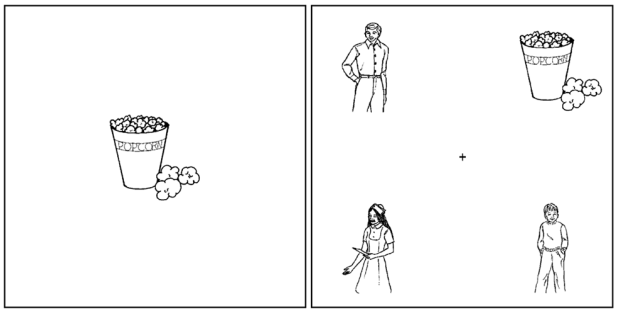


FIGURE 1 | Example visual displays for auditory stimuli shown in (2) (Double-Match condition).

PROCEDURE

Participants sat two meters away from a projection screen where the visual display measured 170 × 120 cm, and during the experiment their eye movements were recorded by a digital camcorder recording 25 frames per second (i.e., one frame every 40 ms) which was placed below the projection screen and trained on the participant's face. This set-up ensured that when the video was played back, participants' eye movements between pictures were distinct enough to be clearly interpreted. The presentation of visual and auditory stimuli was programmed using DMDX (Forster and Forster, 2003), and the sound output from the computer was split, going directly to both the headphones worn by the participant, and to the video camera so that the sound recorded by the video camera was exactly synchronized with what the participant heard. Participants were provided with full details of the procedure and gave written consent before the testing session started.

At the start of each trial, a cross appeared on screen for 1 second, followed by a picture of the object mentioned in the story, which remained in the centre of the screen for 3 seconds. The participant's task was to play a game of "Who is it for?," identifying the recipient of this object while listening to the story which followed. Following the picture of the object, the main visual display for that item was shown on screen for 1 second before the auditory stimulus began, and remained on screen until the next trial began. Participants were asked to listen carefully to the story and respond as quickly as possible once they knew who the object was for, by pressing the button on the gamepad which corresponded with the position of the selected character on the screen. For example, if the recipient was identified as being the character in the top left quadrant of the screen, the participant would press the top left button. If participants answered incorrectly the word "OOPS!" was displayed on the screen to encourage participants to pay closer attention and to discourage hasty responses before the recipient had been identified in the story. There was no feedback for correct responses. The next trial was initiated automatically, independent of the participant's response. Participants were introduced to all the characters and their pictures at the start of the session, and in order to get

[4]The selected picture stimuli could be easily recognised, as shown by their mean "visual recognisability" score of 97% (SD: 6%, range: 80–100%).

used to the pictures and the process of selecting the recipient of the object on the gamepad, the experiment was preceded by six practice trials. For these trials the stories were presented over loudspeakers to allow for immediate questions by the participant as well as to enable the experimenter to check that participants responded shortly after the key word and did not wait until the end of the story. If a participant was not completely confident with the procedure after this, the practice session was repeated. During the main experiment, participants listened to stimuli through headphones and were offered three breaks, one after every 18 items. The entire session took approximately 35 minutes.

Three dependent measures were taken and analyzed: response accuracy (the accuracy with which participants correctly interpreted the reflexive to identify the recipient of the object), response times, and eye movements. For statistical analyses, response accuracy was recorded as either correct or incorrect. Reaction times were calculated as the delay between the onset of the reflexive and when the response button was pressed. Video footage of participants' eye movements was analyzed using ELAN annotation software (Brugman and Russel, 2004), and gaze direction was recorded every frame for 2000 ms (50 frames in total) from the onset of the critical reflexive. The still image for each frame (every 40 ms), was inspected to determine the direction of gaze (toward one of the four pictures, the center of the screen or off-screen), and a target was counted as "fixated" for every frame where eyes were directed toward that picture[5]. Off-screen looks (which accounted for 2.2% of the total dataset) were treated as missing data.

RESULTS

All analyses were carried out on raw data using mixed-effects regression modeling in "R," version 3.0.1 (Baayen et al., 2008; R Development Core Team, 2010). Models included participant and item random effects, and to account for the fact that gaze direction in consecutive frames is not independent (gaze direction in any particular frame is heavily influenced by gaze direction in the previous frame), random effects of Trial were also included for analyses of eye movement data. Maximal random effects structure was used so that as well as random intercepts, all fixed effects and interaction terms had corresponding random slopes by participant, item, and trial as appropriate (Barr et al., 2013). Best fitting models were identified by adding predictors incrementally to an empty model, with those that resulted in a significant improvement of the fit of the model being retained. In the analysis of eye movements, the fixed factor of Time was added to the model in order to test for differences between conditions over time (i.e., proportions of looks increasing or decreasing differently across the two conditions). Due to the non-linear relationship between looks and Time, second and third order polynomials of Time were also tested as predictors. The response accuracy and eye movement data were analyzed using logistic regression due to the categorical nature of the data. For eye movement data the binary dependent variable encoded whether the picture of a particular antecedent was, or was not, fixated for each of the 40 ms frames. Tables/graphs show grand mean results as participant and item differences are accounted for in the mixed-effects analysis.

As the offline measure allows for the identification of trials in which the final interpretation of the reflexive was incorrect, and as response times and eye movements in trials where the inaccessible antecedent (or another incorrect answer) was selected do not reflect successful processing, incorrectly answered trials (comprising 3.6% of the total data set) were not included in the analysis of response times or eye movements.

RESPONSE ACCURACY

As shown in **Table 1**, response accuracy was high (above 95%) in both conditions. In the Double-Match condition the majority of errors were due to the selection of the inaccessible antecedent.

Table 1 | Offline button press responses.

	Correct responses	Incorrect responses	
	% Accessible antecedent	% Inaccessible antecedent	% Other erroneous responses
Double-Match	95.2	4.4	0.4
Single-Match	97.6	0.4	2.0

Analysis of accuracy scores (with each response coded as correct or incorrect) showed no effect of Condition (adding Condition as a fixed factor did not improve the fit of the model over an empty model).

RESPONSE TIMES

Table 2 shows the mean response times for correctly identified recipients. Participants took more time to identify the referent when both antecedents matched the reflexive in gender.

Table 2 | Mean response times (and standard deviation) for correctly answered trials.

	Response time
Double-Match	1155 (688)
Single-Match	1043 (687)

Statistical analyses confirmed that response times were significantly longer in the Double-Match condition [Condition (Double-Match): $\beta = 101.28$, SE $= 44.83$, $t = 2.259$].

EYE MOVEMENTS

Figure 2 shows fixations of the two potential antecedents in the two experimental conditions (Double-Match/Single-Match)

[5]To avoid gaze direction coding being influenced by coders' expectations, coding was initially done "blind," so that gaze direction was coded as being toward the top left, top right, bottom left, bottom right, center, or off-screen (i.e., participant blinking or not looking at screen), without the coder knowing the arrangement of the pictures in the visual display the participant was viewing. Gaze directions were then re-coded with reference to the visual display to show whether the participant was looking at the accessible antecedent, the inaccessible antecedent, the object, the distracter character, the center, or off-screen.

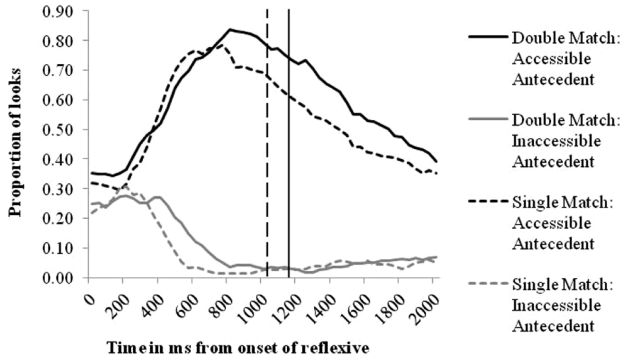


FIGURE 2 | Proportions of looks to potential antecedents.

during the 2 seconds following the onset of the critical reflexive. The *x*-axis displays the time in milliseconds from the onset of the reflexive, and the *y*-axis depicts the proportions of looks to the two potential antecedents, i.e., the number of trials in which a participant fixated on a particular picture for each 40 ms video frame as a proportion of the total number of trials in which they were looking at the screen. As it takes approximately 200 ms to program an eye movement (Rayner et al., 1983), only changes in proportions of looks after 200 ms can be attributed to participants hearing the reflexive. Note that while the graph shows grand mean data plotted on a proportional scale for ease of interpretation, the statistical analysis uses a logistic scale (as analysing data on a proportional scale can lead to inaccurate estimation of effects) and takes into account the clustering of data for each participant, item, and trial.

From 200 ms after hearing the reflexive, the proportion of looks to the accessible antecedent (black lines) increases sharply in both conditions, and looks to the inaccessible antecedent (gray lines) fall. The vertical lines in **Figure 2** indicate the mean response time for each condition (solid line = Double-Match, broken line = Single-Match). Proportions of looks to the other areas of the screen not shown in the graph (object picture, distracter picture and center of the screen) were low throughout the time window (typically between 0 and 0.15), with looks to the object gradually increasing to 0.30 after 1200 ms. The proportion of looks to each of these screen areas was similar across conditions, but slightly higher in the Single-Match condition than the Double-Match condition.

In order to investigate the time course of effects, in the statistical analysis models were fit to 400 ms time windows (200–600 ms, 600–1000 ms, 1000–1400 ms, and 1400–1800 ms). These time windows were selected following visual inspection of the data.

It is important to note that differences between conditions may be seen in two different ways: it may be that in any particular time window the average proportion of looks to an antecedent is higher in one condition than another, or it may be that the rate of increase/decrease in looks (shown by the slope or curve) differs. To investigate the first possibility, models were fit to test for an interaction between Antecedent (Inaccessible/Accessible) and Condition (Single-Match/Double-Match). To explore the second possibility, models also tested for an interaction between Antecedent, Condition, and Time. Thus findings of an Antecedent × Condition interaction, or an Antecedent × Condition × Time interaction each signify (in slightly different ways) that participant performed differently across the two conditions. In later discussion of results, the general term *effect of the inaccessible antecedent* will be used to cover both types of effect.

As shown in **Table 3**, statistical analyses revealed significant interactions between Antecedent, Condition, and Time, in the 200–600 ms and 600–1000 ms time windows. These results show that gaze direction was affected by the gender of the inaccessible antecedent until at least 1 second after the onset of the reflexive.

In order to further investigate the source of the interactions, looks to each antecedent were analyzed separately for the 200–600 ms and 600–1000 ms time windows, as shown in **Table 4**.

From 200 to 600 ms looks to the accessible antecedent increased more slowly in the Double-Match condition than in the Single-Match (shown by the negative slope for the Time × Condition interaction), while, in contrast, from 600 to 1000 ms there was a greater increase in looks to the accessible antecedent in the Double-Match condition (shown by the positive slope for the Time × Condition interaction). While the lack of significant effects in the looks to the inaccessible antecedent shows that there is not

Table 3 | Antecedent × Condition and Antecedent × Condition × Time interactions from best fitting models (full results are shown in Appendix A, found in the Supplementary Material).

Timewindow (ms)	Fixed effects	β	SE	*z* value	*p* value
200–600	Ant (Inacc.) × Condition (Double-Match)	6.694	5.152	1.299	0.194
	Time × Ant (Inacc.) × Condition (Double-Match)	31.792	10.455	3.041	0.002*
600–1000	Ant (Inacc.) × Condition (Double-Match)	−3.256	4.124	−0.790	0.430
	Time × Ant (Inacc.) × Condition (Double-Match)	−39.849	18.122	−2.199	0.028*
1000–1400	Ant (Inacc.) × Condition (Double-Match)	−5.412	8.575	−0.631	0.528
	Time × Ant (Inacc.) × Condition (Double-Match)	−39.850	22.241	−1.792	0.073
1400–1800	Ant (Inacc.) × Condition (Double-Match)	−24.920	34.553	−0.721	0.471
	Time × Ant (Inacc.) × Condition (Double-Match)	84.467	60.952	1.386	0.166

*$p < 0.05$.

Table 4 | Main effect of Condition and Time × Condition interactions from best fitting models fit to looks to each antecedent.

Time window (ms)	Fixed effects	β	SE	*z* value	*p* value
200–600	**Looks to accessible antecedent**				
	Time	59.051	8.840	6.680	<.001*
	Condition (Double-Match)	0.321	2.072	0.155	0.877
	Time × Condition (Double-Match)	−19.099	6.948	−2.749	0.006*
	Looks to inaccessible antecedent				
	Time	−302.20	124.53	−2.430	0.015*
	Condition (Double-Match)	18.460	29.21	0.632	0.527
	Time × Condition (Double-Match)	83.520	143.15	0.583	0.560
600–1000	**Looks to accessible antecedent**				
	Time	−1.120	13.919	−0.080	0.936
	condition (Double-Match)	4.959	4.689	1.058	0.290
	Time × Condition (Double-Match)	40.503	18.646	2.172	0.030*
	Looks to inaccessible antecedent				
	Time	−3.211	20.726	−0.155	0.877
	Condition (Double-Match)	−4.496	5.598	−0.803	0.422
	Time × Condition (Double-Match)	−34.00	28.025	−1.213	0.225

*$p < 0.05$.

a direct relationship between looks to the two antecedents (i.e., a lower proportion of looks to the accessible antecedent does not directly correspond with an increase in looks to the inaccessible antecedent – recall that gaze was distributed over five screen regions), it is nevertheless the case that the presence of a gender matching inaccessible antecedent leads to slower initial identification of the correct antecedent, and then to prolonged looking at the accessible antecedent prior to giving a response to identify the recipient.

SUMMARY OF RESULTS

While offline accuracy in determining the referent for the reflexive was not affected by the gender of the inaccessible antecedent, response times were significantly longer when the gender of the inaccessible antecedent matched that of the reflexive (Double-Match condition).

The analysis of eye movements also showed that the gender of the inaccessible antecedent significantly affected looks to the accessible antecedent over the first 1000 ms following the onset of the reflexive. When a gender matching competitor was present (i.e., in the Double-Match condition) participants were initially slower to identify the correct antecedent (200–600 ms), and then more likely to look at the correct antecedent as they prepared to respond to the task (600–1000 ms).

DISCUSSION

Results showed that adults are significantly distracted by a gender matching but structurally inaccessible competitor antecedent. Eye movement data revealed a two-phase pattern, with early interference effects leading to faster identification of the accessible antecedent in the Single-Match condition, and a later effect whereby participants looked more at the accessible antecedent in the Double-Match condition.

One advantage of eye-tracking during listening over reading-based measures is the ability to focus more precisely on the nature of the effect. While reading-based measures can tell us whether the presence of a gender matching inaccessible antecedent has an effect on the processing of the reflexive, eye-tracking during listening experiments allow us to investigate the origin of that effect more precisely. In this case, we have seen not only that the gender of the inaccessible antecedent has an effect, but specifically that it affects looks to the accessible antecedent. This leads to two possible interpretations of our findings[6]. Firstly, it may be (as is traditionally assumed by studies finding effects of the inaccessible antecedent) that the gender-matching inaccessible antecedent is briefly considered as a potential referent by the parser, before being discarded on the grounds of structural position. If this were the case, one might expect significant effects in the looks to both the inaccessible antecedent and the accessible antecedent (more looks to the inaccessible and fewer to the accessible antecedent). Alternatively, it may be that a gender matching inaccessible antecedent has the effect of slowing down identification of the accessible antecedent, but is not specifically considered as an antecedent itself. Since it is not clear why the gender of the inaccessible antecedent should affect processing of the reflexive unless the inaccessible antecedent were being considered as a competitor, and bearing in mind offline results showing that a gender matching inaccessible antecedent is frequently incorrectly interpreted as the referent for a reflexive

[6]We thank a reviewer for pointing out these two subtly different interpretations.

(Sturt, 2003), we are inclined to support the former interpretation (arguing that there is clearly a numerical, though non-significant, trend toward increased looks to the inaccessible antecedent in the Double-Match condition). However, we acknowledge that the latter interpretation is possible, and that future research probing this distinction is needed. Under either interpretation, it is clear that processing the reflexive involves accessing the inaccessible antecedent, thus arguing against theories which claim that the early application of structural constraints makes inaccessible antecedents "invisible" to the parser.

Our results differ from those reported by Clackson et al. (2011) who used the same materials as the present study but a naturalistic listening task and found no significant effects of the inaccessible antecedent. However, visual inspection of their results shows a numerical effect between 200 and 600 ms similar to the early effect observed here, with a slower increase in looks to the accessible antecedent, and increased looks to the inaccessible antecedent in the Double-Match condition. In order to make a direct comparison between the present study and Clackson et al.'s (2011), data from the latter was re-analyzed using the same analysis methods as presented here (400 ms time windows, maximal random effects structure and including random effects of Trial), however, results showed no significant effects of the inaccessible antecedent[7]. Nevertheless, since early differences between conditions were seen in both experiments (although not significant in Clackson et al., 2011), this suggests that this effect is task-independent, i.e., similar results found using naturalistic and goal-directed designs. In contrast, the later effect appears to be task-specific: in the goal-directed task where participants are aware that the right or wrong response depends on the correct interpretation of the reflexive, we see more looks to the accessible antecedent in the Double-Match condition from 600 to 1000 ms, whereas when participants are required only to listen to auditory stimuli with no emphasis put on processing the reflexive, no such later effect is seen.

The suggestion that later effects may be more affected by the participant's task is supported by evidence from ERP experiments where early and late ERP components differ with regard to their susceptibility to experimental variations. Both the early left anterior negativity (ELAN; occurring around 100–300 ms) and the P600 (occurring around 600–1000 ms) are associated with syntactic violations, but while the early effect is not affected by changes to the task, the later effect has been shown to be dependent on task manipulations such as the expected frequency of syntactic violations (Hahne and Friederici, 1999) and the specific instructions given to participants (Hahne and Friederici, 2002). Such results have led to the suggestion that the early effect reflects highly automatic processes, while the later effect reflects processes that are under the participant's strategic control. Friederici (2002) identifies the P600 component with a process of "reanalysis and repair." Since our participants were more likely to look at the picture of the accessible antecedent in the more challenging Double-Match condition immediately prior to responding, this may reflect a similar process of overcoming any earlier confusion and "checking" the answer. Logically, such a checking process would be absent when the task did not require the participant to give a response identifying the referent of the reflexive.

The cross-task differences in results observed for studies using the same auditory stimuli highlight the importance of identifying and separating task-independent and task-related effects. In eye-tracking during listening studies, the naturalistic listening method avoids participants adopting behavioral strategies to complete the task (as there is no task), but leaves questions about whether participants actually processed the linguistic element under investigation, and if so, whether their interpretation was in fact correct. In contrast, the goal-directed method forces participants to process the required language and gives a clear indication of the participant's interpretation, although the results may also reflect the conscious processes involved in attaining the goal. It is only by systematic comparison of results from experiments using the same materials but differing designs that the role of the task can be identified. More studies of this sort are needed to confirm which effects are truly task-independent, and in the case of eye-tracking during listening studies, to further explore how cross-condition differences between looks to the target and looks to the competitor might be interpreted.

It might be suggested that a potential explanation for the early effect is that in the Double-Match condition participants initially interpret the first syllable of "himself/herself" as the pronoun "him/her," leading to early eye movements toward the gender matching non-local antecedent before participants hear "... self." However, acoustic comparison of the first syllable of "himself/herself" and the pronouns "him/her" carried out by Clackson et al. (2011) showed that the unstressed syllable in the reflexive was significantly reduced in duration and intensity compared to the pronoun. While pronouns often occur in phonologically weak forms, in the materials used here any pronoun occurring in the position of the reflexive would naturally be pronounced as a strong form, making it unlikely that participants would interpret the weak first syllable of the reflexive as a pronoun.

As outlined in the introduction, results from previous experiments using different methodologies differ with regard to the existence and timing of interference effects. In particular, eye-tracking during reading studies have revealed conflicting patterns of results (even when the materials were very similar), and where interference effects are reported, these are usually in "later measures" corresponding with Sturt's (2003) "defeasible filter" theory, which proposes that although the inaccessible antecedent is initially blocked by the syntactic constraint, the parser may consider it at a later point in processing. In contrast, the results from the current study suggest that the interference caused by the gender matching inaccessible antecedent occurred relatively early in processing. While this apparent timing difference is still to be fully explained, it may be related to differences between auditory and visual processing or the fact that the two methodologies measure very different things, making it questionable whether reading times on the reflexive and following words can be directly compared with the probability of looking at a

[7]Perhaps because the low salience of the reflexive in the naturalistic design meant that in a large number of trials participants did not look at any potential antecedent on hearing the reflexive, thus reducing the number of valid data points and leading to a low-power analysis.

particular referent. Another contributing factor may be that the low salience of the reflexive affects reading designs in the same way that it can lead to participants failing to look at a potential antecedent in naturalistic listening designs. Specifically, the null effects in early reading measures could be due to high skipping rates and the resulting smaller amount of data points, i.e., a lack of power to detect small effects. For instance, Felser and Cunnings (2012) and Cunnings and Felser (2013) report skipping rates in the reflexive region of 11.2–15.6%, considerably higher than in the spill-over region (5.1–8.2%), raising the possibility that the reported null effect in early measures is due to a lack of power.

Connected to skipping rates, a further potential explanation for a lack of consistent effects in reading studies is the preview benefit in written texts. While orally presented sentences are presented one phoneme after the other, readers can visually inspect several letters at a time, both in the fovea and the parafovea. The fact that the reading span in English generally extends 14–15 letters to the right of the fixation allows readers to "look ahead" in the sentence [for reviews of research on parafoveal processing see Rayner (1998) and Schotter et al. (2012)]. Therefore, it is likely that in reading studies participants processed the reflexive parafoveally before actually fixating on it. With spaces and length information being very salient, the distinction between English reflexives (6–10 letters) and pronouns (2–4 letters) can easily be made on the basis of this formal information available in the parafovea. This might provide participants with a "head-start," reducing potential surprise effects which lead to longer reading times when a reflexive does not refer to the gender matching and discourse prominent, but structurally inaccessible, antecedent.

Even across methodological boundaries, it is clear that the discourse prominence of the inaccessible antecedent plays a role in determining the extent to which it can interfere with processing of the reflexive. In the present study and previous research reporting interference effects, the materials used were constructed such that the inaccessible antecedent was promoted in the discourse by being both in first-mentioned position and the matrix subject (Badecker and Straub, 2002; Sturt, 2003; Cunnings and Felser, 2013). In contrast, studies using materials where the inaccessible antecedent was not in first mentioned or matrix subject position (Xiang et al., 2009; Dillon et al., 2013), or where the prominence of the inaccessible antecedent relative to that of the accessible antecedent was reduced (Felser et al., 2009) have found no reliable effect of the inaccessible antecedent. This is consistent with recent findings showing that while sentences presented in isolation provide evidence for a syntax-based account of sentence processing, structural parsing mechanisms are influenced by discourse factors when sentences are placed in a more natural context (Yang et al., 2013).

In conclusion, our findings support a multiple constraint or cue-based retrieval approach to reflexive resolution whereby each potential antecedent is promoted by a variety of factors (both structural and discourse related), and while strong weighting is given to the structural constraint, non-structural cues or constraints (such as discourse prominence) can also affect online reflexive resolution. Furthermore, we suggest that behavioral measures may be influenced by the specific task participants are given and particularly that later occurring effects may reflect more conscious/controlled processes, as has also been reported in previous ERP research.

ACKNOWLEDGMENTS

This research was supported by an ESRC postgraduate studentship awarded to Kaili Clackson by the Department of Language and Linguistics at the University of Essex and a Ph.D. scholarship from the Potsdam Research Institute for Multilingualism awarded to Vera Heyer. We are grateful to Loay Balkhair for sharing participants and to Harald Clahsen and members of the Psycholinguistic Research Group for useful discussions.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fpsyg.2014.00904/abstract

REFERENCES

Baayen, R., Davidson, D., and Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *J. Mem. Lang.* 59, 390–412. doi: 10.1016/j.jml.2007.12.005

Badecker, W., and Straub, K. (2002). The processing role of structural constraints on interpretation of pronouns and anaphors. *J. Exp. Psychol. Learn. Mem. Cogn.* 28, 748–769. doi: 10.1037/0278-7393.28.4.748

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: keep it maximal. *J. Mem. Lang.* 68, 255–278. doi: 10.1016/j.jml.2012.11.001

Brugman, H., and Russel, A. (2004). Annotating multimedia/multi-modal resources with ELAN. *Paper presented at the 4th International Conference on Language Resources and Evaluation*, Lisbon.

Burkhardt, P. (2005). *The Syntax-discourse Interface: Representing and Interpreting Dependency*, Vol. 80. Amsterdam: John Benjamins Publishing Co.

Chen, Z., Jäger, L., and Vasishth, S. (2012). "How structure-sensitive is the parser? evidence from Mandarin Chinese," in *Empirical Approaches to Linguistic Theory: Studies in Meaning and Structure : Studies in Generative Grammar*, eds B. Stolterfoht and S. Featherston (Berlin: Mouton de Gruyter), 43–62.

Chomsky, N. (1981). *Lectures in Government and Binding*. Dordrecht: Foris.

Chomsky, N. (1986). *Knowledge of Language: Its Nature, Origin, and Use*. New York: Praeger.

Clackson, K., Felser, C., and Clahsen, H. (2011). Children's processing of reflexives and pronouns in English: evidence from eye-movements during listening. *J. Mem. Lang.* 65, 128–144. doi: 10.1016/j.jml.2011.04.007

Cunnings, I., and Felser, C. (2013). The role of working memory in the processing of reflexives. *Lang. Cogn. Process.* 28, 188–219. doi: 10.1080/01690965.2010.548391

Dillon, B., Mishler, A., Sloggett, S., and Phillips, C. (2013). Contrasting intrusion profiles for agreement and anaphora: experimental and modeling evidence. *J. Mem. Lang.* 69, 85–103. doi: 10.1016/j.jml.2013.04.003

Felser, C., and Cunnings, I. (2012). Processing reflexives in English as a second language: the role of structural and discourse-level constraints. *Appl. Psycholinguist.* 33, 571–603. doi: 10.1017/S0142716411000488

Felser, C., Sato, M., and Bertenshaw, N. (2009). The on-line application of binding principle A in English as a second language. *Biling. Lang. Cogn.* 12, 485–502. doi: 10.1017/S1366728909990228

Forster, K., and Forster, J. (2003). DMDX: a windows display program with millisecond accuracy. *Behav. Res. Methods Instrum. Comput.* 35, 116–124. doi: 10.3758/BF03195503

Friederici, A. D. (2002). Towards a neural basis of auditory sentence processing. *Trends Cogn. Sci.* 6, 78–84. doi: 10.1016/S1364-6613(00)01839-8

Hahne, A., and Friederici, A. D. (1999). Electrophysiological evidence for two steps in syntactic analysis: early automatic and late controlled processes. *J. Cogn. Neurosci.* 11, 194–205. doi: 10.1162/089892999563328

Hahne, A., and Friederici, A. D. (2002). Differential task effects on semantic and syntactic processes as revealed by ERPs. *Cogn. Brain Res.* 13, 339–356. doi: 10.1016/S0926-6410(01)00127-6

Levinson, S. C. (1987). Pragmatics and the grammar of anaphora: a partial pragmatic reduction of binding and control phenomena. *J. Linguist.* 23, 379–434. doi: 10.1017/S0022226700011324

Nicol, J. (1988). *Coreference Processing During Sentence Comprehension.* Ph.D. thesis, Massachusetts Department of Brain and Cognitive Sciences, Institute of Technology, Cambridge, MA.

Nicol, J., and Swinney, D. (1989). The role of structure in coreference assignment during sentence comprehension. *J. Psycholinguist. Res.* 18, 5–19. doi: 10.1007/BF01069043

Phillips, C., Wagers, M. W., and Lau, E. F. (2010). "Grammatical illusions and selective fallibility in real-Time language comprehension," in *Experiments at the Interfaces, Syntax and Semantics* : *Bingley*, Vol. 37. ed. J. Runner (UK: Emerald Publications).

Pollard, C. J., and Sag, I. A. (1992). Anaphors in English and the scope of binding theory. *Linguist. Inq.* 23, 261–303.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychol. Bull.* 124, 372–422. doi: 10.1037/0033-2909.124.3.372

Rayner, K., Slowiaczek, M. L., Clifton, C., and Bertera, J. H. (1983). Latency of sequential eye movements: implications for reading. *J. Exp. Psychol. Hum. Percept. Perform.* 9, 912–922. doi: 10.1037/0096-1523.9.6.912

R Development Core Team. (2010). *R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna.* Available at: http://www.R-project.org

Reinhart, T. (2000). *The Theta System: Syntactic Realization of Verbal Concepts.* Utrecht: OTS working papers in linguistics.

Reinhart, T., and Reuland, E. (1993). Reflexivity. *Linguist. Inq.* 24, 657–720.

Reuland, E. (2001). Primitives of binding. *Linguist. Inq.* 32, 439–492.

Runner, J. T., Sussman, R. S., and Tanenhaus, M. K. (2003). Assignment of reference to reflexives and pronouns in picture noun phrases: evidence from eye movements. *Cognition* 89, B1–B13. doi: 10.1016/S0010-0277(03)00065-9

Runner, J. T., Sussman, R. S., and Tanenhaus, M. K. (2006). Processing reflexives and pronouns in picture noun phrase. *Cogn. Sci.* 30, 193–241. doi: 10.1207/s15516709cog0000_58

Schotter, E. R., Angele, B., and Rayner, K. (2012). Parafoveal processing in reading. *Atten. Percep. Psychophys.* 74, 5–35. doi: 10.3758/s13414-011-0219-2

Sturt, P. (2003). The time-course of the application of binding constraints in reference resolution. *J. Mem. Lang.* 48, 542–562. doi: 10.1016/S0749-596X(02)00536-3

Trueswell, J. C. (2008). Using eye movements as a developmental measure within psycholinguistics. *Dev. Psycholinguist. On-line Methods Children's Lang. Process.* 73–96. doi: 10.1075/lald.44.05tru

Van Dyke, J. A. (2007). Interference effects from grammatically unavailable constituents during sentence processing. *J. Exp. Psychol. Learn. Mem. Cogn.* 33, 407–430. doi: 10.1037/0278-7393.33.2.407

Xiang, M., Dillon, B., and Phillips, C. (2009). Illusory licensing effects across dependency types: ERP evidence. *Brain Lang.* 108, 40–55. doi: 10.1016/j.bandl.2008.10.002

Yang, F., Mo, L., and Louwerse, M. (2013). Effects of local and global context on processing sentences with subject and object relative clauses. *J. Psycholinguist. Res.* 42, 227–237. doi: 10.1007/s10936-012-9215-5

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 17 May 2014; accepted: 29 July 2014; published online: 19 August 2014.
*Citation: Clackson K and Heyer V (2014) Reflexive anaphor resolution in spoken language comprehension: structural constraints and beyond. Front. Psychol. **5**:904. doi: 10.3389/fpsyg.2014.00904*
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology.
Copyright © 2014 Clackson and Heyer. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 26 May 2016
doi: 10.3389/fpsyg.2016.00329

Retrieval Interference in Syntactic Processing: The Case of Reflexive Binding in English

Umesh Patil[1,2]*, *Shravan Vasishth*[1] and *Richard L. Lewis*[3]

[1] Department of Linguistics, University of Potsdam, Potsdam, Germany, [2] Computational Linguistics, Institute of Cognitive Science, University of Osnabrück, Osnabrück, Germany, [3] Department of Psychology, University of Michigan, Ann Arbor, MI, USA

It has been proposed that in online sentence comprehension the dependency between a reflexive pronoun such as *himself/herself* and its antecedent is resolved using exclusively syntactic constraints. Under this *strictly syntactic search* account, Principle A of the binding theory—which requires that the antecedent c-command the reflexive within the same clause that the reflexive occurs in—constrains the parser's search for an antecedent. The parser thus ignores candidate antecedents that might match agreement features of the reflexive (e.g., gender) but are ineligible as potential antecedents because they are in structurally illicit positions. An alternative possibility accords no special status to structural constraints: in addition to using Principle A, the parser also uses non-structural cues such as gender to access the antecedent. According to cue-based retrieval theories of memory (e.g., Lewis and Vasishth, 2005), the use of non-structural cues should result in increased retrieval times and occasional errors when candidates partially match the cues, even if the candidates are in structurally illicit positions. In this paper, we first show how the retrieval processes that underlie the reflexive binding are naturally realized in the Lewis and Vasishth (2005) model. We present the predictions of the model under the assumption that both structural and non-structural cues are used during retrieval, and provide a critical analysis of previous empirical studies that failed to find evidence for the use of non-structural cues, suggesting that these failures may be Type II errors. We use this analysis and the results of further modeling to motivate a new empirical design that we use in an eye tracking study. The results of this study confirm the key predictions of the model concerning the use of non-structural cues, and are inconsistent with the strictly syntactic search account. These results present a challenge for theories advocating the infallibility of the human parser in the case of reflexive resolution, and provide support for the inclusion of agreement features such as gender in the set of retrieval cues.

Keywords: sentence processing, anaphor resolution, memory retrieval, interference, computational modeling, eye tracking

OPEN ACCESS

Edited by:
Colin Phillips,
University of Maryland, College Park, USA

Reviewed by:
Dan Parker,
College of William & Mary, USA
Jeffrey Thomas Runner,
University of Rochester, USA

***Correspondence:**
Umesh Patil
umesh.patil@gmail.com

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: *05 September 2015*
Accepted: *22 February 2016*
Published: *26 May 2016*

Citation:
Patil U, Vasishth S and Lewis RL (2016) Retrieval Interference in Syntactic Processing: The Case of Reflexive Binding in English. Front. Psychol. 7:329. doi: 10.3389/fpsyg.2016.00329

1. INTRODUCTION

Sentence comprehension involves, among other things, recovering hierarchical structure from an input string of words (e.g., Frazier, 1979). Such recovery requires the online application of grammatical constraints that delimit the possible relationships between various elements of the sentence. For example, to understand a sentence like (1), the pronoun *himself* has to be resolved

to a referent of an earlier noun *surgeon*; the reflexive cannot be associated with *Jonathan* due to Principle A of the binding theory (Chomsky, 1981)[1].

(1) The surgeon who treated Jonathan had pricked himself.

Establishing a relationship between two non-adjacent elements in a sentence requires maintaining some memory of the immediate past. The question we are concerned with here is: what role do grammatical and non-grammatical constraints play in the access to the immediate past? The binding of English reflexive pronouns is a particularly informative case, because the configurational and agreement constraints are relatively clear, and the structure admits manipulations of distance and distracting candidate antecedents (Sturt, 2003).

One proposal for how structural constraints are implicated in dependency resolution is motivated by the experiments reported in Nicol and Swinney (1989), Sturt (2003), and Xiang et al. (2009). Using different experimental methodologies, these studies found that a grammatically incorrect antecedent [e.g., *Jonathan* in (1)] does not interfere in the process of binding a reflexive pronoun by a grammatically correct antecedent [e.g., *surgeon* in (1)], at least in the early stages of processing a reflexive.

Nicol and Swinney (1989) presented evidence from a series of experiments that employed the cross modal lexical priming paradigm. Participants listened to sentences similar to those shown in (2) and responded to visually presented probe words that was presented immediately following the reflexive *himself*. The probe word was either semantically related or unrelated to one of the three previously occurring nouns in the sentence (*boxer*, *skier*, or *doctor*). Participants judged whether the probe word was a word or non-word. A significant priming effect was observed when probe words were related to grammatically accessible as antecedents [e.g., *doctor* in (2)], but not when they were related to grammatically inaccessible antecedents [e.g., *boxer* and *skier* in (2)]. Nicol and Swinney (1989) concluded that no priming was observed for words related to grammatically inaccessible antecedents because they had not been considered during co-reference resolution.

(2) The boxer told the skier that the doctor for the team would blame himself for the recent injury.

Sturt (2003) reported eye tracking studies using materials such as (3). He found that first fixation duration and first pass reading time on the region containing the reflexive were longer when the gender of the reflexive did not match the stereotypical gender of the grammatically accessible antecedent than when it matched (e.g., *herself* and *surgeon* vs. *himself* and *surgeon*). Early reading times were *not* affected by gender match between the reflexive and the grammatically inaccessible antecedent (*Jonathan* or *Jennifer*). Second pass reading time at the reflexives showed an interaction for gender match between the reflexive and the two antecedents suggesting that in later interpretation stages (but, crucially, not in earlier processing stages)[2] the inaccessible antecedent is part of the candidates being considered as antecedents. There was also an effect of the inaccessible antecedent in second pass reading time in the pre-final region, but this effect was observed only when the accessible antecedent matched the gender of the reflexive. However, these late effects of the inaccessible antecedent were not observed in the second experiment [with design as in (4)], but he pointed out that the absence of any effect of the inaccessible antecedent in Experiment 2 could have been due to the fact that the inaccessible antecedent did not c-command the reflexive and it was also not as prominent as in Experiment 1.

To gain further insight into this late-stage interpretation of the sentences, Sturt (2003) also ran a follow-up study, where a sentence-by-sentence self-paced reading task was followed by a question that directly probed for the antecedent of the reflexive. This study showed a significant interference effect, with more ungrammatical interpretations when the grammatically inaccessible antecedent matched the gender of the reflexive; the effect was bigger when the accessible antecedent did not match the gender of the reflexive. Sturt (2003) concluded that grammatical constraints are applied very early in processing, but interference from the grammatically inaccessible antecedent occurs during later processes that are related to recovery strategies, rather than during processes related to the initial interpretation of the reflexive.

(3) {Jonathan/Jennifer} was pretty worried at the City Hospital. {He/She} remembered that the surgeon had pricked {himself/herself} with a used syringe needle. There should be an investigation soon.

(4) {Jonathan/Jennifer} was pretty worried at the City Hospital. The surgeon who treated {Jonathan/Jennifer} had pricked {himself/herself} with a used syringe needle. There should be an investigation soon.

Dillon et al. (2013) reported two eye tracking studies with English reflexives using material with syntactic structure similar to Experiment 2 in Sturt (2003). They also did not find any effect of the inaccessible antecedent, but reported effects of the accessible antecedent. Xiang et al. (2009) reported similar results in an ERP study, where they found that a P600 is elicited by a reflexive that mismatches the stereotypical gender of the grammatically accessible antecedent, and is not attenuated by the presence of a matching antecedent in a grammatically inaccessible position.

Based on results from these studies, Phillips et al. (2011) suggests that:

> "*…argument reflexives are immune to interference from structurally inaccessible antecedents because antecedents are*

[1]Principle A specifies a structural constraint on the interpretation of reflexives in English: a reflexive must be bound by an antecedent in the local domain (the current clause). An antecedent X can bind a reflexive Y, if X and Y are coindexed, and X c-commands Y. The term c-command defines a hierarchical relationship between two constituents in a syntax tree. A constituent c-commands its sister constituent and every constituent *below* the sister constituent in the syntax tree. In (1), the reflexive *himself* is bound by *surgeon*; the noun *Jonathan* cannot bind the reflexive because it does not c-command it.

[2]In this paper, we follow the literature (see e.g., Sturt, 2003) in assuming that so-called early and late measures in eye tracking data map onto processes that occur (respectively) in early and late stages of parsing.

retrieved using only structural cues. In effect, we are suggesting that the person, gender, and number features of reflexives like himself, herself, and themselves play no role in the search for antecedents . . . "

An alternative possibility accords no special status to the structural constraints: in addition to using Principle A, the parser also uses non-structural cues such as gender to access the antecedent. For example, in (1), it is possible that the parser considers a relation between *Jonathan* and *himself*, due to a gender-feature match, and perhaps also due to the relative proximity of *Jonathan* compared to *surgeon*. Under at least one cue-based retrieval theory (Lewis and Vasishth, 2005; Lewis et al., 2006), this should result in interference from grammatically inaccessible antecedent while resolving the reflexive-antecedent dependency[3]. Some evidence for this account comes from studies reported in Badecker and Straub (2002), Choy and Thompson (2010), and Cunnings and Felser (2013) among others.

Badecker and Straub (2002) reported an interference effect from gender-matched distractors in a self-paced reading experiment using sentences as in (5). They found that reading times two words beyond a reflexive were slowed by the presence of a gender matching NP in a grammatically inaccessible position.

(5) {Jane/John} thought that Bill owed himself another opportunity to solve the problem.

More evidence for early retrieval interference in reflexive binding comes from the study reported in Cunnings and Felser (2013). In two eye tracking studies they tested how application of Principle A varies between low and high working memory span readers. In the first study they found a late effect of inaccessible antecedent, emerging only at regions following the reflexive region. However, in the second study where the inaccessible antecedent was closer in the surface string to the reflexive, the effect of inaccessible antecedent was observed in an early eye movement measure, namely first fixation duration, at the reflexive itself, although this effect was limited to low span readers. Consequently, Cunnings and Felser (2013) conclude that "lower span participants were more likely to consider both potential antecedents of the reflexive early on during processing, before converging on the structurally accessible antecedent later on."

Further evidence for the effect of interference from grammatically inaccessible antecedent comes from an eye tracking study in a visual world paradigm reported by Choy and Thompson (Thompson and Choy, 2009; Choy and Thompson, 2010). Although this study was targeted at aphasics' processing deficits with binding constructions, for present purposes we consider data only from unimpaired participants. Choy and Thompson (2010) recorded eye movements while the participants listened to a story as in (6), with the critical sentence containing a pronoun or a reflexive (e.g., *him* or *himself*). The visual stimuli consisted of pictures of two persons, one of which was grammatically accessible and the other inaccessible (e.g., *soldier* and *farmer*); a human-referring distractor; and an inanimate-referring noun mentioned in the story (e.g., *glasses*). The data for the reflexive condition from unimpaired participants showed an increase in the proportion of fixations to the inaccessible antecedent in the reflexive and post-reflexive regions compared to the pre-reflexive region. Although the proportion of fixations to the accessible antecedent was higher than the fixations to the inaccessible antecedent in most of the regions, the increase in fixations to the inaccessible antecedent from the onset of the reflexive indicates that participants considered the inaccessible antecedent as the potential antecedent of the reflexive, albeit less often than the accessible antecedent.

(6) Some soldiers and farmers were in a house. The soldier told the farmer with glasses to shave {himself/him} in the bathroom. And he did.

In summary, the effect of interference from a grammatically inaccessible antecedent is sometimes observed in early processing and sometimes in late processing, and in some studies the effect is completely absent.

In the remainder of this paper, we first apply an existing computational model of cue-based parsing (Lewis and Vasishth, 2005) to an empirical paradigm well established for testing the processing of reflexives in English. The model generates qualitative predictions and demonstrates that these predictions are robust against substantial variation in the quantitative parameters. We then use the theoretical perspective provided by the model to formulate conjectures for why some of the existing empirical work may have failed to detect evidence for the use of non-structural cues. Based on this analysis we advance a new experimental design which is intended to be more sensitive, and demonstrate that for many of the predictions the modified design yields larger effects in modeling simulations. We next present an eye tracking study based on the modified design, yielding several results that confirm the early use of non-structural cues in a manner consistent with the model. The paper concludes with discussion of the implication of these new results for some current theoretical approaches to dependency resolution.

2. MODELING REFLEXIVE BINDING IN THE CUE-BASED RETRIEVAL FRAMEWORK

The cue-based retrieval architecture provides a natural characterization of the retrieval steps triggered in the process of reflexive resolution. We begin by presenting a model of Experiment 1 and its follow-up in Sturt (2003), which will provide insight into the predicted effects and their robustness against parametric variation, and provide motivation for the modified design used in the eye tracking study reported here.

[3]We will follow the literature in referring to the correct antecedent as stipulated by Principle A as the *grammatically accessible* antecedent and the antecedent that is incorrect following Principle A as the *grammatically inaccessible* antecedent. Occasionally, we abbreviate these terms to *accessible* and *inaccessible* antecedents. It is important to keep in mind that under the model we advocate in this paper, the grammatically inaccessible antecedent is in fact "accessible" for memory retrieval; a more appropriate term would have been "incorrect antecedent," since this does not presuppose that the non-c-commanding antecedent is inaccessible.

The emphasis of the model described here is not on parsing the entire sentence, but on detailed modeling of the retrieval process carried out at the reflexive.

Experiment 1 in Sturt (2003) included an eye tracking experiment in which participants were required to read short texts consisting of three sentences. An example is given in (7), showing the four experimental conditions. A named referent (*Jonathan* or *Jennifer*) is introduced in the first sentence, and this referent is subsequently referred to using a pronoun (*he* or *she*) in the second sentence. The second sentence also introduces a second referent *the surgeon*, and includes a reflexive anaphor (*himself* or *herself*). The first named referent is not a grammatically accessible antecedent for the reflexive in terms of binding theory, while the second referent (*the surgeon*) is a grammatically accessible antecedent. Accessible and inaccessible antecedents either matched or did not match the gender of the reflexive. Note that even when the accessible antecedent doesn't match the gender of the reflexive, the sentences are still grammatical because a *surgeon* is only stereotypically masculine and hence a licit antecedent of *herself*.

In match-interference and match conditions [(a) and (b) in (7)] the accessible antecedent matches the gender requirement of the reflexive, and in mismatch-interference and mismatch conditions [(c) and (d) in (7)] it does not. Furthermore, in match-interference and mismatch-interference conditions the inaccessible antecedent matches the gender of the reflexive. Henceforth, we will refer to match-interference and match conditions simply as *match* conditions, and mismatch-interference and mismatch conditions as *mismatch* conditions, reflecting the fact that the accessible antecedent matches the gender of the reflexive for one pair and does not for the other. We will refer to match-interference and mismatch-interference conditions as the *interference* conditions because the gender of the inaccessible antecedent matches that of the reflexive—potentially causing interference.

(7) **Sentence 1:** {Jonathan/Jennifer} was pretty worried at the City Hospital.

Sentence 2:

a. *Accessible-match/inaccessible-match* (*Match-interference*)
He remembered that the surgeon had pricked **himself** with a used syringe needle.

b. *Accessible-match/inaccessible-mismatch* (*Match*)
She remembered that the surgeon had pricked **himself** with a used syringe needle.

c. *Accessible-mismatch/inaccessible-match* (*Mismatch-interference*)
She remembered that the surgeon had pricked **herself** with a used syringe needle.

d. *Accessible-mismatch/inaccessible-mismatch* (*Mismatch*)
He remembered that the surgeon had pricked **herself** with a used syringe needle.

Sentence 3: There should be an investigation soon.

This eye tracking study showed an early effect of the accessible antecedent (**Figure 1**). First fixation duration and first pass reading time were faster in the match conditions compared to the mismatch conditions. But no effect of the inaccessible antecedent was found in the early measures. The effect of inaccessible antecedent was found only in later measures—second pass reading time was shorter in match-interference condition compared to the match condition.

As mentioned above, Sturt (2003) also conducted a follow-up study to find out the participants' final interpretation of the reflexive. This was a sentence-by-sentence self-paced reading with the same sentences as in (7) but, instead of sentence 3, there was a question that explicitly probed for the antecedent of the reflexive [e.g., a question like *Who had been pricked with a used needle?* with possible answers, for example, for condition (a) as *Jonathan* or *surgeon*]. The follow-up study showed a main effect of accessible antecedent, inaccessible antecedent and also an interaction between these two factors (see (**Figure 1**). When the accessible antecedent did not match the gender of the reflexive, participants made a higher proportion of errors in selecting the correct antecedent. In addition, when the inaccessible antecedent matched the gender of the reflexive participants made more errors than when it did not. Moreover, the increase in error due to gender match with the inaccessible antecedent was larger when the accessible antecedent did not match the gender, resulting in the interaction between the two factors.

Thus, there are four major findings in Sturt's Experiment 1 and his follow-up study. First, gender mismatch with the default gender of the accessible antecedent resulted in lower question-response accuracies. Second, gender match with the inaccessible antecedent resulted in lower question-response accuracies. Third, early reading time measures (first fixation duration and first pass reading time) increased when the gender specification of the accessible antecedent mismatched that of the reflexive. Fourth, second pass reading time (re-reading time) was *shorter* when the gender of the inaccessible antecedent matched the gender of the reflexive (this occurred in the case where the accessible antecedent matched the reflexive in gender, i.e., in match conditions).

Interestingly, the first three of the four effects can be explained by simply assuming that the search for an antecedent includes a gender feature. Therefore, we begin our modeling by assuming that both grammatical knowledge about antecedent and gender matching is used when resolving antecedents in English. For simplicity, we model the grammatical constraint by assuming that the antecedent should be a noun and should be the subject of the clause containing the reflexive, albeit under some different implementation of this grammatical constraint, the predictions may turn out differently[4]. This choice of retrieval cues is motivated by the conjecture that including agreement features in general may be an adaptive feature of the parser, although attempting to establish this is not the purpose of the work reported here. As a result the set of retrieval cues for the

[4]There are more sophisticated ways to encode the c-command constraint but these implementation details are orthogonal to the present discussion. See Alcocer and Phillips (2012) which compares some alternatives of implementing the c-command constraint.

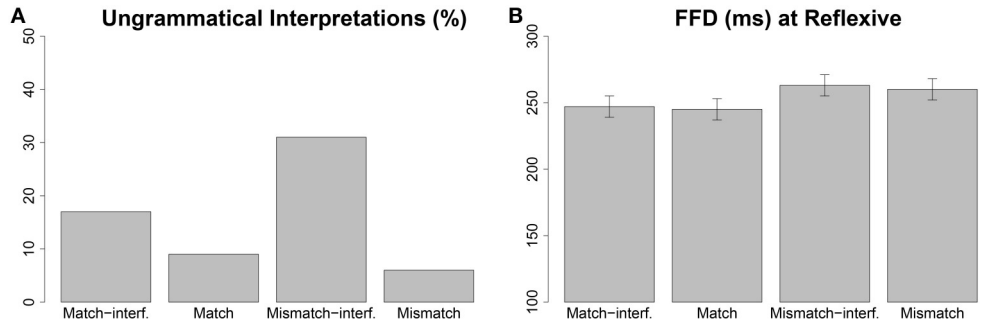


FIGURE 1 | The effects in Sturt (2003) Experiment 1: (A) proportions of ungrammatical interpretations of reflexives in the follow-up study; (B) first fixation durations in the eye tracking study. The error bars in the plot for first fixation duration show 95% confidence intervals.

TABLE 1 | The match of retrieval cues with the accessible and inaccessible antecedents for the four conditions in Sturt's experiments (cat=category).

Conditions	Accessible	Inaccessible
a (match-interference)	Gender, cat, role, clause	Gender, cat, role
b (match)	Gender, cat, role, clause	Cat, role
c (mismatch-interference)	Cat, role, clause	Gender, cat, role
d (mismatch)	Cat, role, clause	Cat, role

reflexives *himself* and *herself* are {gender = masculine/feminine, category = noun, role = subject, clause = current-clause}, differing only in the value of the *gender* feature. See **Table 1** for the list of cues matched by the two antecedents across the four conditions in Sturt's experiments.

The accessible antecedent matches all four cues in conditions (a) and (b) (the "match" conditions), but only three cues in conditions (c) and (d) (the "mismatch" conditions), since the stereotypical gender of *surgeon* is masculine, which does not match the gender retrieval cue at *herself* (gender = feminine). The inaccessible antecedent matches three cues (gender, category, role) out of a total of four cues in conditions (a) and (c), and in conditions (b) and (d) it matches two cues (category, role). As a result, interference for retrieving the antecedent will be higher in conditions (a) and (c) (the "interference" conditions) as compared to conditions (b) and (d). Note that the alternative possibility, as suggested by Phillips and colleagues, is that gender plays no role in retrieval; in that case, the cues for the match (a vs. b) and mismatch conditions (c vs. d) would be identical, leading to no interference.

The cue-based retrieval model predicts that similarity-based interference (SBI) arises at the moment of retrieval. SBI in reflexive binding is manifested in terms of delay in retrieval of the correct antecedent or an error in retrieving the correct antecedent. The delay in retrieval of the correct antecedent is a result of the fan assumption (see Equation 3 in Appendix A of Supplementary Material) that reduces the strength of association between a cue and a target as a function of the number of items associated with that cue. Reduced strength of association means reduced activation boost, which produces higher latencies. On the other hand, the error in retrieval of the correct antecedent is a combined result of activation fan and partial match. Reduction in activation boost of the accessible antecedent due to activation spreading, and partial matching between retrieval cues (the second summation component in Equation 1 in Appendix A of Supplementary Material) and any inaccessible antecedents can lead (probabilistically as a function of activation noise) to higher activation of the inaccessible antecedents. As a result, the probability of retrieving the inaccessible antecedent increases. The greater the partial match with inaccessible antecedents, the higher the percentage of errors in retrieving the accessible antecedent.

We model retrieval in sentence 2 from (7); this is the crucial sentence for generating predictions about the reflexive binding process. The predictions of the model are generated by running 1000 simulations for each condition. All model parameters are set to the values that have been used in the previous models from Lewis and Vasishth (2005), Vasishth and Lewis (2006), and Vasishth et al. (2008). A list of all the parameter values that we use is given in Table A1 in Appendix A of Supplementary Material.

The predicted retrieval error percentages accurately capture the pattern found in the Sturt (2003) follow-up study: There is a main effect of accessible antecedent, inaccessible antecedent, and an interaction between these two factors, exactly as in Sturt's follow-up study's response accuracies. First, when the accessible antecedent does not match the gender of the reflexive the model makes a higher number of errors in retrieving the correct antecedent (the mismatch effect in response accuracy). Second, when the inaccessible antecedent matches the gender of the reflexive the model makes more errors than when it does not (the interference effect in response accuracy). Third, the increase in error due to gender match with the inaccessible antecedent is greater in the mismatch conditions.

The retrieval times predicted by the model (shown in **Figure 2**) show a main effect of matching in the accessible antecedent: When the accessible antecedent does not match the gender of the reflexive, the retrieval times are higher than when it does. The model also predicts a match × interference interaction—retrieval times are predicted to be higher in the *match-interference* condition (198 ms) than in the *match* condition (194 ms); however, retrieval times are predicted to be

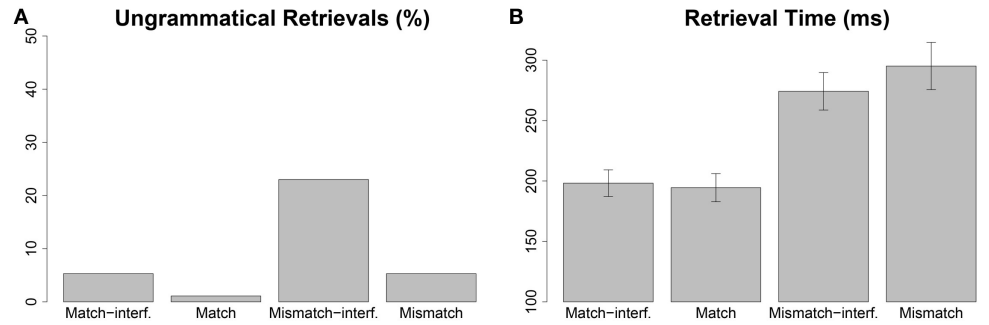


FIGURE 2 | The twofold predictions of the cue-based retrieval model for Experiment 1 in Sturt (2003), based on the parameter values listed in Table A1 in Supplementary Material. (A) Proportions of retrievals of the inaccessible antecedent; **(B)** retrieval times at the reflexive. The error bars in the plot for retrieval time show 95% confidence intervals

lower in the *mismatch-interference* condition (274 ms) compared to the *mismatch* condition (295 ms).

In order to compare the predictions to the data, we use the following terminology: the *mismatch effect* is the difference between the *match* conditions and the *mismatch* conditions; the *interference effect* is the effect between the two interference conditions and the other two conditions; the *match-interference effect* is the effect of interference in the two *match* conditions; and the *mismatch-interference effect* is the effect of interference in the two *mismatch* conditions.

The predicted ungrammatical retrievals accurately model the ungrammatical interpretations observed in the Sturt (2003) follow-up study. However, the predicted retrieval times accurately capture only the mismatch effect observed in the first fixation duration (FFD) in the eye tracking study. The interaction predicted between the *mismatch* effect and *interference* effect is not observed in the data. Thus, the model accurately captures the question-response accuracy data, but only partly characterizes the first fixation duration data.

The divergent patterns between the model's retrieval times and the first fixation durations in Sturt's study come from the differences in the patterns seen in the predicted match-interference effect and the mismatch-interference effect. The predicted match-interference effect is a consequence of spreading of activation of the gender cue which is matched by both the accessible and inaccessible antecedent. As described earlier, activation spreading reduces the strength of association between the cue and the target, causing longer retrieval latencies in the match-interference condition than the match condition. On the other hand, the mismatch-interference effect is a consequence of partial match between the cues and inaccessible antecedent: the inaccessible antecedent matches the gender cue which is not matched by the accessible antecedent (see **Table 1**), leading to higher probability of retrieving the inaccessible antecedent. This can be seen in the predicted retrieval error percentages in **Figure 2**. Moreover, in the *mismatch-interference* condition the inaccessible antecedent receives more activation from retrieval cues than in the *mismatch* condition as it matches more retrieval cues in the *mismatch-interference* condition. A substantially higher number of incorrect retrievals occur due to higher activation from the retrieval cues, and the retrieval times in the *mismatch-interference* condition are faster than the retrieval times in the *mismatch* condition, contrary to the findings in Sturt's study (see **Figure 1** vs. **Figure 2**). We return to this issue in the Section 3.6.

To summarize, the model predicts the following effects for retrieval errors (RE) and retrieval times (RT) at the reflexive:

E1. Mismatch effect (RE): the retrieval errors for the two mismatch conditions are higher than those for the two match conditions.

E2. Interference effect (RE): the retrieval errors for the match-interference and mismatch-interference conditions are higher than those for the other two conditions.

E3. Mismatch effect (RT): the retrieval times for the two mismatch conditions are longer than those for the two match conditions.

E4. Match-interference effect (RT): the retrieval times for the match-interference condition are longer than the match condition.

E5. Mismatch-interference effect (RT): the retrieval times for the mismatch-interference condition are *shorter* than the mismatch condition.

In Sturt's experiment, only the effects E1, E2 and E3 were observed. The interference effects E4 and E5 were missing in the early measures (first fixation duration and first pass reading time) of the eye tracking studies.

Here we assume that the RE translates to incorrect interpretation of the reflexive and RT translates to reading time in the experiment. We also make a simplified assumption about the lexical representation of nouns with stereotypical gender—as far as the gender feature is concerned, the representation of a stereotypically masculine or feminine noun (e.g., "soldier" or "nurse") is the same as that of an unambiguously masculine or feminine noun (e.g., "John" or "Jane"). It has been shown that the gender violation effects are stronger for definitionally masculine or feminine nouns than for stereotypically masculine or feminine nouns (Osterhout et al., 1997). This means that

our simplified assumption may lead to inflated predictions of various effects than what might be observed in an experiment. Finding out the precise difference in the representation of these two types of nouns is an important research question, but we think it lies outside the scope of this paper. We also assume that the first antecedent that is retrieved, is considered to be the correct antecedent of the reflexive in the final representation irrespective of its gender match with the reflexive; i.e., there is no reanalysis of the reflexive-antecedent dependency.

2.1. Parametric Variability in the Model

We did not estimate any parameter values for the current model. All existing parameters were set to the values that have been used in previous published versions of the cue-based retrieval model. It is possible, however, that the predictions of the model are valid only for the specific parameter values that we used here; this could be the reason behind the lack of effects E4 and E5 in the data—these effects might emerge only for a particular combination of parameter values. Conversely, the correct predictions of effects E1, E2, and E3 might depend on the specific values used by the model. To gain a better understanding of the range of possible predictions of the model, we ran the model for a range of values of three crucial ACT-R parameters: *noise*, *maximum associative strength* and *maximum difference*. The noise parameter controls the amount of instantaneous activation noise added to each chunk at retrieval; maximum associative strength is the constant "S" in Equation 3 in Supplementary Material; and the maximum difference parameter controls the penalty due to a mismatch between a retrieval cue and a feature value of a chunk. For each of these parameters, the range of values over which the predictions are generated is given in **Table 2**. The predictions are generated by running 1000 simulations for each combination of values of the three parameters. The total number of combinations of the three parameter values are 1287 (see **Table 2**). The predictions of effects E1–E5 across these sets of parameter values are plotted in **Figure 3**. Each effect is plotted against the parameter along which it varies the most. Effects E1 and E2 are influenced the most by noise, E3 and E5 are influenced the most by the maximum difference parameter, and E4 is influenced the most by the maximum associative strength parameter. Each point in the plots represents a mean over all values of the other two parameters. In short, **Figure 3** illustrates how the size of each effect varies across different parameter values.

The effect E1 varies from 0 to 23.9%, the effect E2 varies from 0 to 17.15%, the effect E3 varies from −1.55 to 228.4 ms, the effect E4 varies from −6.53 to 17.34 ms and the effect E5 varies from from −36.01 to 3.21 ms. The effects E1 and E2 are zero when the instantaneous activation noise is zero, which essentially means that the model does not make any mistake in retrieving the accessible antecedent when there is no noise added to the activations of chunks. But, in general, a non-zero value for noise parameter is necessary for modeling memory errors and response time distribution. Overall, although all effects show variation across different parameter combinations, they all remain mainly non-zero and have the same numerical sign as the predicted effects with the predefined parameter values. These results show that the model's predictions for E1–E3 are not crucially dependent on the fixed parameter values we used.

TABLE 2 | The range of parameter values used for testing the parametric variability of the cue-based retrieval models.

Parameter	Range of values
Noise	0.05–0.4, in steps of 0.05
Maximum associative strength	1–4, in steps of 0.25
Maximum difference	−1 to 0, in steps of 0.1

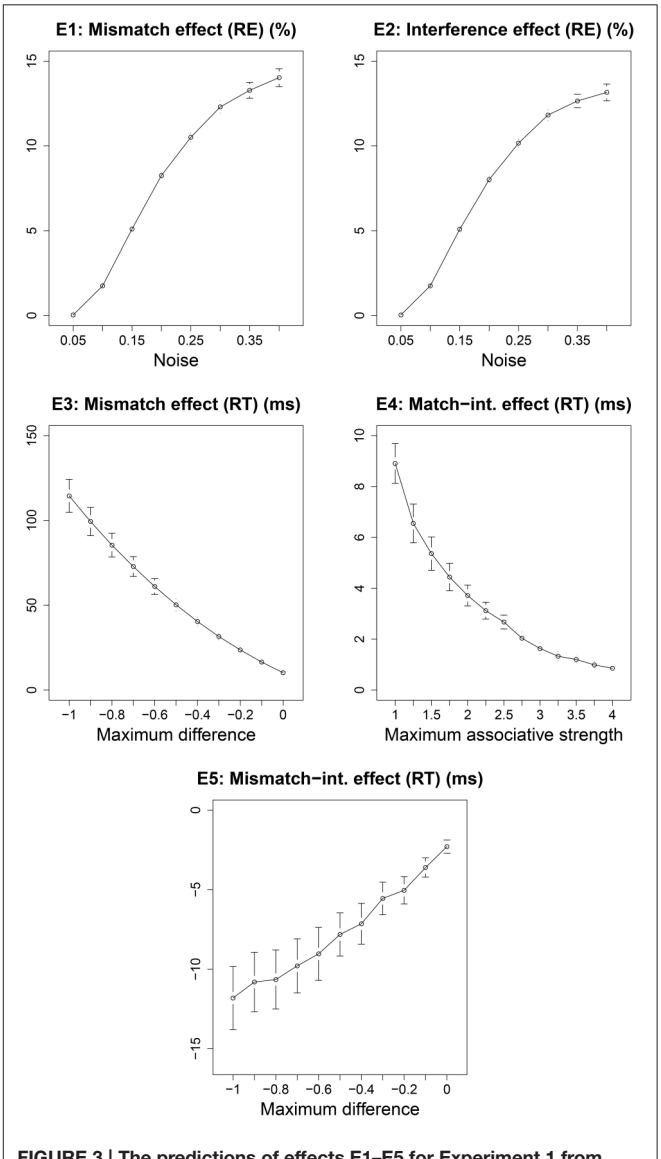


FIGURE 3 | The predictions of effects E1–E5 for Experiment 1 from Sturt (2003) across a range of parameter values. Each effect is plotted against the parameter that affects it the most.

2.2. An Alternative Explanation for the Absence of Interference Effects (E4 and E5) in Sturt (2003)

Although the lack of an interference effect in Sturt (2003) Experiment 1 could imply that non-structural cues like gender are not used in retrieval, Sturt noticed that the absence of an effect could be due to the non-local linear position of the interferer (inaccessible antecedent) with respect to the reflexive [see (7) above]. The accessible antecedent was introduced later in the string than the inaccessible antecedent, and was therefore closer to the reflexive. In his Experiment 2, Sturt (2003) modified this design by using stimuli as in (8), where the linear positions of the binding accessible and inaccessible antecedents are reversed with relation to Experiment 1, while their accessibility with respect to the binding theory is kept constant. However, this experiment also did not show any interference effect.

(8) {Jonathan/Jennifer} was pretty worried at the City Hospital. The surgeon who treated {Jonathan/Jennifer} had pricked {himself/herself} with a used syringe needle. There should be an investigation soon.

In addition to surface-string locality, we consider now another possibility for the apparent lack of interference: the degree of overlap between potential distractors and retrieval cues. We hypothesized above that reflexive binding uses grammatical category (noun, verb etc.), grammatical role (subject, object etc.) and gender as the retrieval cues to retrieve the correct antecedent. In the cue-based retrieval model, the overlap of these cues with grammatically inaccessible antecedents leads to an interference effect in both retrieval latency and retrieval errors. This formulation in the model leads to the following alternative explanation for the lack of interference effect in Sturt's Experiment 2: the interfering antecedent was the object of the relative clause [see (7) above], and hence did not match the grammatical role cue for retrieval at the reflexive. In fact, Van Dyke and McElree (2011) have recently proposed that although distractors with matching semantic cues exert interference, cues like grammatical role are weighted heavily in the retrieval process. They found that the interference effect due to the semantic match was present only when the distractors matched the grammatical cues as well. These results can also explain the lack of interference effect in Sturt's Experiment 2.

In terms of activations of memory elements, the probability of retrieving an incorrect element is higher if it has a higher activation value at the time of retrieval. The activation value of a memory element is directly dependent on its creation time, retrieval history, and its match with the retrieval cues—the more recently an element is created or retrieved, and the higher feature overlap it has with the retrieval cues, the better chances it has of being retrieved. Consequently, in Sturt's Experiment 1 the interferer has less chance of getting retrieved due to its less recent creation time with respect to the accessible antecedent, and in Experiment 2 the interferer has less chance of getting retrieved because its overlap with the retrieval cues is lower in comparison to the overlap of the accessible antecedent with the retrieval cues. In other words, in Sturt's experiments the inaccessible antecedents may not be strong enough interferers to detect their effect on the retrieval process. If this reasoning is correct, then the effect, or rather the lack of an interference effect, might be a false negative (a type II error). Concluding that an absence of an interference effect is evidence that no interference occurs has important consequences to the theory of retrieval processes in sentence comprehension. No interference in processing argument reflexives implies that the retrieval mechanism for reflexive binding is different from other retrieval mechanisms in sentence processing, e.g., subject-verb agreement, and agreement attraction. On the other hand, finding an interference effect simplifies the theory of retrieval processes considerably, since no exemption is granted to antecedent-reflexive resolution processes.

2.3. A Modified Design

In order to increase the strength of the interference effect, we can use an object relative clause [see (9)] where the inaccessible antecedent has the subject role in the clause. It is also closer to the reflexive in terms of linear distance. Under the cue-based retrieval account, the inaccessible antecedent would be more likely to interfere in the retrieval process than in the two experimental designs in Sturt (2003)—but under the structurally-constrained approach, this manipulation should not matter to the reflexive binding process. In fact, Xiang et al. (2009) used this design in their ERP study, but they did not have the crucial *match* condition. Cunnings and Felser (2013) also used this design to test the interaction of reflexive processing and memory capacity. They do find effect of inaccessible antecedent, but they did not evaluate their findings in terms any specific memory retrieval mechanism. See Section 4 for more details. Note that our design also uses the manipulation of stereotypical gender of the accessible antecedent, as in Sturt (2003), but all sentences are grammatical despite the gender mismatch between the reflexive and the stereotypical gender of the accessible antecedent.

(9)
a. *Accessible-match/inaccessible-match (match-interference)*
The tough soldier that Fred treated in the military hospital introduced himself to all the nurses.
b. *Accessible-match/inaccessible-mismatch (match)*
The tough soldier that Katie treated in the military hospital introduced himself to all the nurses.
c. *Accessible-mismatch/inaccessible-match (mismatch-interference)*
The tough soldier that Katie treated in the military hospital introduced herself to all the nurses.
d. *Accessible-mismatch/inaccessible-mismatch (mismatch)*
The tough soldier that Fred treated in the military hospital introduced herself to all the nurses.

We implemented a cue-based retrieval model for the modified design described in (9) as well as for Experiment 2 in Sturt (2003). The goal of this modeling is to compare the predictions of the cue-based retrieval theory for the five effects (E1–E5) across three designs—Experiment 1 (including the follow-up study; we

count the eye tracking study and the follow-up study as one experiment, following Sturt), Experiment 2 from Sturt (2003), and the modified design. The modeling assumptions are the same as in the model described above.

Figure 4 compares the predictions for effects E1–E5 across the three experimental designs. The predictions are generated for the range of parameter values listed in **Table 2**. The pattern of effects E1–E5 for Experiment 2 and the modified design are similar to that for Experiment 1. Across a range of parameter values, the predictions for effects E1, E2, and E5 are clearly stronger (higher numerical value) for the modified design than for Experiment 1 and 2 in Sturt (2003). Although the predictions for effect E3 are almost identical for the modified design and Experiment 2, they are nevertheless stronger than for Experiment 1. In contrast, the predictions for effect E4 are not distinguishable across the three designs. To gain better insight into the predictions for E4, we compared effect E4 across variations of the other two parameters (noise and maximum difference); see **Figure 5**. For the maximum difference parameter, effect E4 is stronger in the modified design and Experiment 2 than in Experiment 1 when the difference penalty is high (more negative), and it is weaker when the maximum difference penalty is low. For the noise parameter, effect E4 is stronger in the modified design and Experiment 2 than in Experiment 1 when the noise is low, and it is weaker when the noise is high. These patterns show that the predicted strength of effect E4 is dependent on the specific value or a range of values that are selected for these parameters. The noise parameter is a frequently modified parameter across various models (Wong et al., 2010), which is suggestive of uncertainty regarding its value across diverse cognitive tasks (cf. the decay parameter, which is usually kept fixed). The best way to estimate or, at least, restrict the noise parameter's value would be to empirically validate predictions of various models. In contrast to noise, the maximum difference penalty parameter is seldom modified, and is set to its default value of −1. For the default value of this parameter, the model clearly predicts a stronger E4 effect for the modified design and Experiment 2.

In sum, the predictions for the modified design and Experiment 2 show stronger effects than Experiment 1 across a range of parameter values. For the most part—and as expected—the effects for the modified design are much stronger than the other two designs. Next, we report an eye tracking study that we ran with the modified design (9). The goal of this study is to evaluate the predictions of our model that diverge from Sturt's findings (specifically, effects E4 and E5), as well as to replicate the effects E1–E3 that Sturt (2003) found.

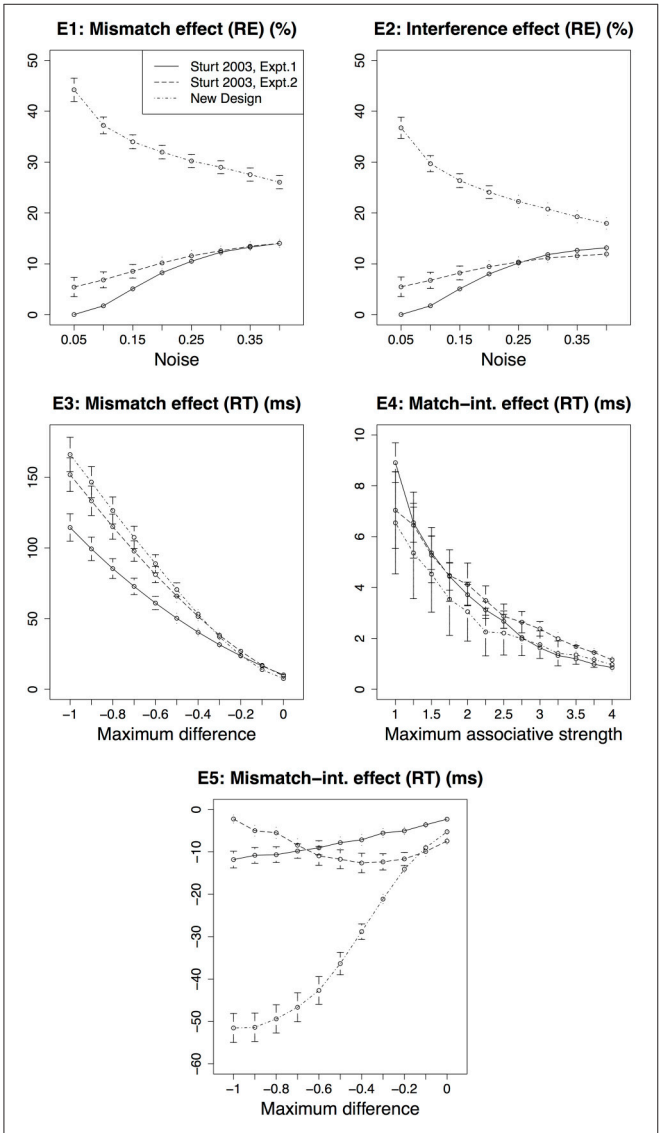


FIGURE 4 | The predictions of effects E1–E5 for three experimental designs across a range of parameter values.

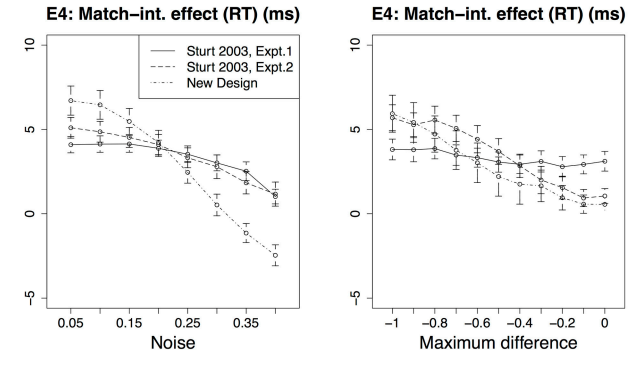


FIGURE 5 | The variations in effect E4 (match-interference (RT)) across noise and maximum difference parameters for three experimental designs.

3. EYE TRACKING EXPERIMENT

3.1. Participants

Forty English native speakers residing in Berlin, Germany participated in the eye tracking study. Data from one participant was excluded due to less than 40% accuracy on the sentence comprehension questions on all trials including experimental

and filler trials. The remaining 39 participants included 20 female participants and had a mean age of 29.5 years. The 39 native English speakers consisted of 14 British, 13 American, 8 Australian, 3 Canadian, and 1 New Zealander. All participants had normal or corrected-to-normal vision and were paid 10 Euros for their participation. The experiment had a duration of approximately 45 min, including set-up time. This study was carried out in accordance with the Helsinki declaration with written informed consent from all participants.

3.2. Design and Materials

Twenty-four stimuli were selected from the Xiang et al. (2009) study and constructed as per (9) by adding an extra *match* condition (see Appendix B in Supplementary Material for the list of stimuli). Of these 24 stimuli, 12 used stereotypically male nouns and 12 used stereotypical female nouns for the binding-accessible antecedent. There were 4 lists that comprised different item-condition combinations according to a Latin Square. Each list contained 54 filler sentences. Two-third of the target items and all fillers contained a comprehension question, and these were equally distributed across yes and no answers. In all, each participant answered 70 comprehension questions.

3.3. Procedure

Participants were seated 60 cm from an NEC Multisync 2080UX screen color monitor with 1600 × 1200 pixel resolution. They were asked to sit comfortably in front of an EyeLink 1000 eye tracker (SR Research) running at 500 Hz sampling rate (0.01° tracking resolution, and $< 0.5^{\circ}$ gaze position accuracy). Though the viewing was binocular, only the participant's right eye was tracked. The distance between the camera and the eye was 50 cm.

Participants were asked to position their head in a frame that stabilized their forehead and chin. They were asked to avoid large head movements throughout the experiment and to avoid blinking while reading the sentences. A 7-button Microsoft Sidewinder game pad was used to record button responses. The presentation of the materials and the recording of responses was controlled by two separate PCs, one running internally developed software (this is called EyeScript, and was originally developed in Richard Lewis' lab by Mason Smith, and later in Shravan Vasishth's lab by Felix Engelmann, Titus von der Malsburg, and Tobias Günther; the software is open source and available at https://github.com/tmalsburg/EyeScript) and the other running SR Research's proprietary software.

Each participant was randomly assigned one of four different stimulus lists. The list was randomized for every subject. At the start of the experiment, a standard calibration procedure was performed which involved participants looking at a grid of 13 fixation targets in random succession, in order to validate their gazes. Calibration and validation were repeated if the experimenter noticed that measurement accuracy was poor, and if participants took a break during the experiment.

Each trial consisted of the following steps: First, a fixation target in the same position as the first character of the text display was presented; two 200 ms fixations followed by one 400 ms fixation on this target triggered the presentation of the sentences (this procedure ensured that the participants always started reading in the left-most character position and helped the experimenter ensure the accuracy of calibration). Participants were instructed to read the sentence at a normal pace and to move their gaze to a dot at the bottom right of the screen after finishing the sentence. This triggered the presentation of a comprehension question in two-thirds of the trials, and in the rest it triggered the presentation of the next trial. The comprehension questions were included in order to ensure that the participants attended to the content of the sentences.

3.4. Data Analysis

All data processing and analyses were carried out in GNU-R (R Development Core Team, 2009). Fixations were detected using the algorithm described by Engbert and Kliegl (2003); an open source R package, *saccades*, developed by Titus von der Malsburg was used to carry out this step (the package is available at https://github.com/tmalsburg/saccades). Fixation and regression-based measures were extracted using another open source R package, *em2*, developed by Pavel Logačev (the package is available at: https://cran.r-project.org/src/contrib/Archive/em2/). All fixations 30 pixels above and below the sentence were included in the sentence. Fixations in the blank spaces between words were also counted; fixations in the first half of the space were included in the fixations on the preceding word and fixations in the second half were included in the fixations on the following word. All other fixations outside these regions were excluded.

Effects of accessible antecedent gender match (henceforth, *match*) and inaccessible antecedent gender match (*interference*) with the gender of the reflexive were evaluated across various eye movement measures. Data analysis was carried out using linear mixed models (Bates and Sarkar, 2007; Gelman and Hill, 2007). Linear mixed models were fit for the following eye movement measures at the reflexive. *First Fixation Duration* (FFD), the time spent during the first fixation during the first pass; *First Pass Reading time* (FPRT), the sum of all fixations during the first pass; *Re-reading Time* (RRT), the sum of all fixations in a region that occurred after first pass; *Total Reading Time* (TRT), the sum of all fixations in a region, *First Pass Regression Probability* (FPRP), the probability of regressing from a region after fixating in that region during first pass; and Re-reading Probability (RRP), the probability of reading a region during the second and subsequent passes. In the linear models, we used nested contrast coding and defined three contrasts that correspond to the three effects that we are interested in—mismatch effect, match-interference effect, and mismatch-interference effect. The interference effects were nested within the mismatch effect. The contrasts were coded such that having a positive coefficient meant that the effect was in the predicted direction. Apart from these three contrasts, trial number was used as a (centered) predictor. All linear mixed models were fit with by-participant and by-item random intercepts, and by-participant and by-item random slopes for the three contrasts. For FPRP and RRP, only random intercepts were used, since otherwise the models failed to converge. All reading times were log transformed before fitting the linear models. For FPRP and RRP, generalized linear mixed models were fit with a binomial link function.

TABLE 3 | Mean reading times at the reflexive with standard errors, percentages of first pass regressions from the reflexive, percentages of re-readings of the reflexive, and comprehension question response accuracies across four conditions.

Condition	FFD	FPRT	RRT	TRT	FPRP (%)	RRP (%)	Accuracy (%)
a. Match-interf.	258 (6)	280 (8)	108 (14)	410 (18)	13	30	84
b. Match	263 (7)	292 (10)	79 (11)	389 (17)	7	24	90
c. Mismatch-interf.	272 (8)	295 (10)	149 (20)	473 (24)	11	35	83
d. Mismatch	266 (9)	284 (9)	145 (15)	468 (20)	11	42	80

TABLE 4 | Linear mixed-effects model estimates, standard errors and *t*-values across reading time measures; the asterisk indicates statistically significant ($\alpha = 0.05$) effects.

	Effect	Estimate	Std. Error	*t*-value
FFD	Intercept	5.512	0.027	201.387
	Mismatch	0.023	0.026	0.89
	Match-interference	−0.022	0.034	−0.650
	Mismatch-interference	−0.025	0.038	−0.653
	Trial	0.001	0.001	1.751
FPRT	Intercept	5.575	0.029	194.555
	Mismatch	0.011	0.03	0.356
	Match-interference	−0.033	0.053	−0.626
	Mismatch-interference	−0.028	0.04	−0.718
	Trial	0.001	0.001	2.203*
RRT	Intercept	5.679	0.052	109.464
	Mismatch	0.043	0.066	0.65
	Match-interference	0.022	0.094	0.231
	Mismatch-interference	−0.067	0.082	−0.818
	Trial	−0.001	0.001	−0.748
TRT	Intercept	5.9	0.049	121.125
	Mismatch	0.146	0.043	3.41*
	Match-interference	0.036	0.057	0.637
	Mismatch-interference	0.016	0.057	0.273
	Trial	0	0.001	0.4

TABLE 5 | Linear model estimates, standard errors and *p*-values for FPRP and RRP; the asterisk indicates statistically significant ($\alpha = 0.05$) effects.

	Effect	Estimate	Std. Error	*p*-value
FPRP	Intercept	−2.602	0.258	$<10^{-15}$
	Mismatch	0.148	0.256	0.564
	Match-interference	0.782	0.374	0.037*
	Mismatch-interference	−0.111	0.35	0.751
	Trial	0.004	0.006	0.431
RRP	Intercept	−0.852	0.176	$<10^{-5}$
	Mismatch	0.63	0.154	$<10^{-4}$*
	Match-interference	0.359	0.225	0.11
	Mismatch-interference	0.383	0.206	0.063
	Trial	−0.006	0.003	0.067

3.5. Results

All mean reading times at the reflexive along with standard errors, FPRP from the reflexive, RRP at the reflexive and comprehension accuracy percentages are summarized in **Table 3**. The results of the statistical analysis are summarized in **Tables 4, 5**.

3.5.1. Question-Response Accuracy

Overall average accuracy for trials that included a comprehension question was 88% and average accuracy for target items was 84%. Accuracy values for comprehension questions across four conditions are listed in **Table 3**, but are not theoretically interpretable because the questions targeted different parts of the critical sentence, not just the antecedent-reflexive relation as in the Sturt follow-up study. We present these mean accuracies only for completeness.

3.5.2. Eye Tracking Dependent Measures

A statistically significant mismatch effect was observed in TRT and RRP, i.e., the conditions in which the stereotypical gender of the accessible antecedent did not match the gender of the reflexive were read more slowly and had higher probability of re-reading than the conditions where it matched. A statistically significant match-interference effect was observed in FPRP, with the high interference condition showing more regressions than the low interference condition.

3.6. Discussion

3.6.1. Early Effects

The results outlined above show an early effect of match-interference (E2) from the inaccessible antecedent in first pass regression probability, such that a gender match between the reflexive and the inaccessible antecedent leads to a higher number of first pass regressions from the reflexive (see **Figure 6**). The occurrence of a regression from a word reflects some difficulty in integrating the word when it is fixated and hence it is plausibly an early effect (Clifton et al., 2007). We are assuming that higher number of retrieval errors should reflect in higher probability of regression. First pass regressions cannot reflect the late processes triggered at the end of a sentence or the processes reflected by late measures such as second pass reading time. Assuming that first pass regressions reflect processing difficulty triggered relatively early during the first contact with the critical word, the interference effect is inconsistent with the conclusion of Sturt (2003), that the online application of Principle A is not affected by interference from the inaccessible antecedent at early stages of processing. Conclusions derived in Nicol and Swinney

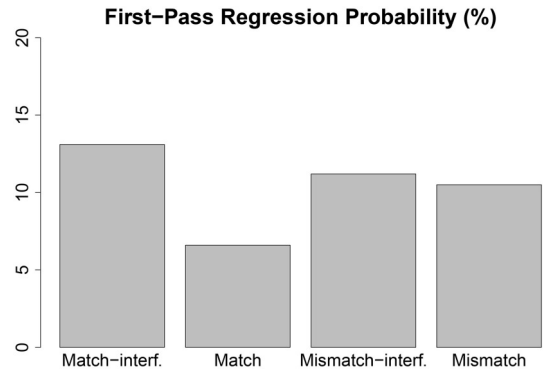


FIGURE 6 | Early effect of *interference*. Effects that are observed in "early" eye tracking measures are considered to be early effects. Usually measures like FFD, FPRT, and FPRP are considered to be early measures since they are associated with reader's first exposure to a region.

(1989) and Xiang et al. (2009) are also not compatible with these results. As a result, this study challenges the claim from Phillips et al. (2011) and Dillon et al. (2013) that an antecedent for a reflexive is retrieved using only structural cues without considering the gender feature. Our findings are consistent with those of Badecker and Straub (2002), Choy and Thompson (2010), Cunnings and Felser (2013), Thompson and Choy (2009).

3.6.2. Late Effects

The effect of accessible antecedent gender match (E1 and E3) was also observed in the RRP and TRT (see **Figure 7**) such that reading times were elevated and there were higher number of re-readings when the accessible antecedent did not match the gender of the reflexive (we are assuming that E1, the mismatch effect predicted in retrieval errors, should reflect in elevated reading times). The absence of an early effect of accessible antecedent is different from the finding of Sturt (2003), where the effect appeared at FFD. We also observed a marginal effect of mismatch-interference (E5) ($p = 0.063$) in the RRP. Although the effect doesn't reach conventional significance level, it corroborates the patterns we observe in our exploratory data analysis with cumulative progressions (see Section 3.6.4).

3.6.3. Regression Contingent Effects in FFD

As an exploratory data analysis, we analyzed FFD contingent on the first pass regressions—separate analysis for FFD followed by regressions and FFD not followed by regressions. The two patterns are plotted in **Figure 8**. FFD followed by regressions show a pattern consistent with the retrieval times predicted by the model. Although the match-interference effect (E4) ($t = 1.77$) and mismatch-interference effect (E5) ($t = 1.70$) do not quite reach conventional significance levels, they show the trend of interference effect as predicted by the model. These FFDs also show the main effect of mismatch (E1 and E3) ($t = 2.78$) which is consistent with the early mismatch effect in Sturt (2003). FFD not followed by regressions did not show this effect.

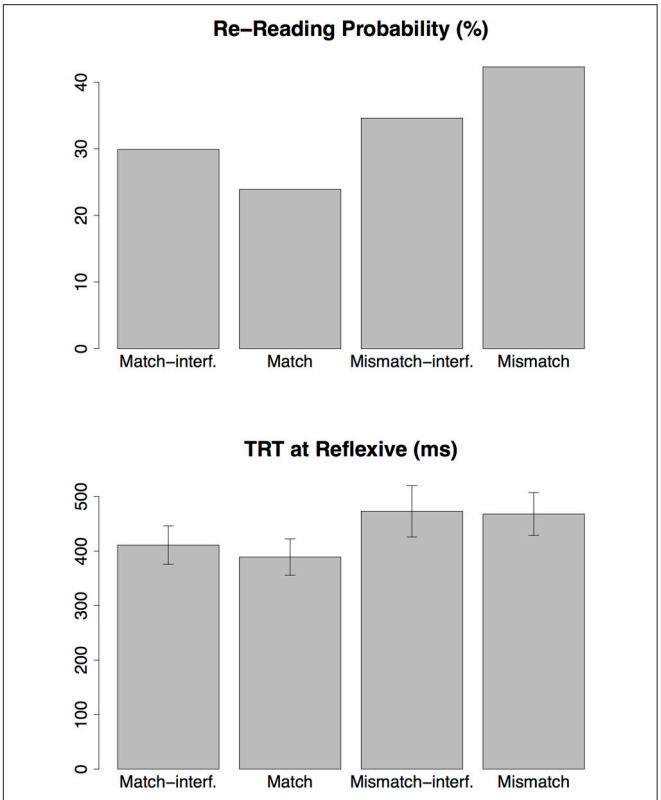


FIGURE 7 | Late effect of *match*. Effects that are observed in "late" eye tracking measures are considered to be late effects. Usually measures like RRP and TRT are considered to be late measures since they involve reader's second and possibly subsequent exposures to a region.

3.6.4. Effects Revealed in Cumulative Progressions

As another way of exploratory data analysis, we examined an eye movement measure called the cumulative progression, which has been used earlier by Kreiner et al. (2008) and Cunnings and Sturt (2014). The cumulative progression quantifies how far a reader's eyes have traveled from the region of interest. The assumption with analyzing cumulative progressions is that the further away, in the direction of reading, a reader progresses from a region (in one condition compared to another), the easier the information in that region is for processing. It is a measure of continuous eye movements, in the sense that it assigns a numeric value, the distance, at each point in time that can be recorded by an eye tracker. This makes it possible to compare the processing cost between two conditions over a continuous period of time. For example, in our case, we could examine if participants consistently progress further away from the reflexive region in the match conditions compared to the mismatch conditions after entering the reflexive region for the first time. And if they do, then, by assumption, it implies that the reflexives are easier to process in the match conditions than in the mismatch conditions. Effectively, we are assuming that faster retrievals at the reflexive will result (in faster processing, and hence) in progressions that are further away from the reflexive.

Cumulative progressions are computed by measuring the distance between the position of the first fixation in the region

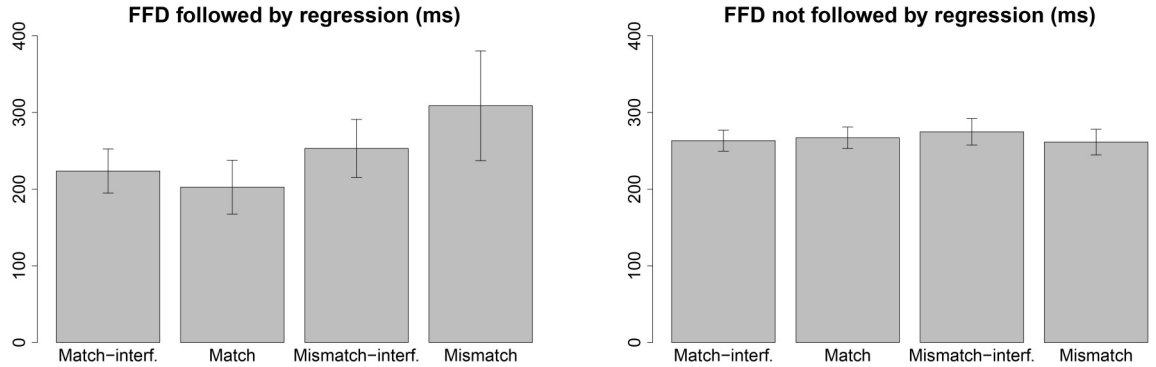


FIGURE 8 | Regression contingent FFD.

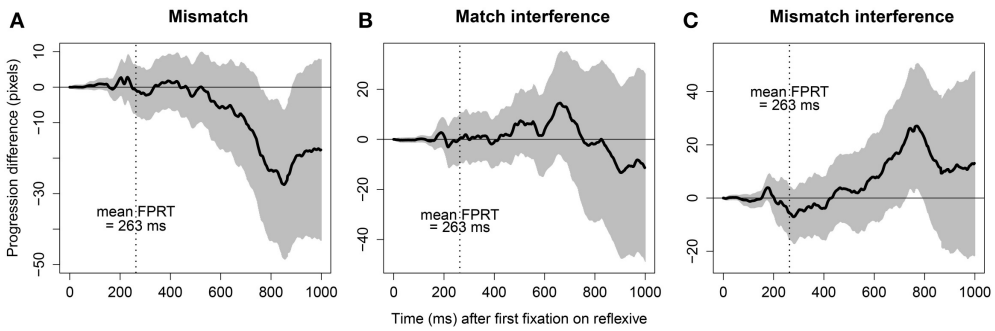


FIGURE 9 | Progression differences plotted against timestamps starting at the first fixation in the reflexive region. Each curve represents an effect and the gray area around it represents the 95% confidence interval. The dotted vertical line denotes the position of the mean FPRT for the reflexives. **(A)** The mismatch effect is obtained by subtracting the mean progressions in the two mismatch conditions from the mean progressions in the two match conditions, **(B)** the match-interference effect is obtained by subtracting the mean progressions in the match condition from the mean progressions in the match-interference condition, and **(C)** the mismatch-interference effect is obtained by subtracting the mean progressions in the mismatch condition from the mean progressions in the mismatch-interference condition.

of interest and the subsequent eye positions, ignoring word boundaries. In the earlier studies mentioned above the distance was calculated in terms of characters (the number of characters by which the current eye position is separated from the position of the first fixation in the region of interest). Only forward eye movements change the value of the measure; regressive eye movements or no eye movements, as in fixations, do not change the value of the measure. This means that the sequence of cumulative progressions for one trial is a monotonically increasing sequence—every subsequent number (representing the distance) is greater than or equal to the previous number (hence the name cumulative). Unlike in earlier studies, where the distance was calculated in terms of characters, we calculate the distance in terms of the number of screen pixels a participant has progressed, which gives a more fine-grained measure of distance. As in Cunnings and Sturt (2014), we evaluate various effects by comparing the numerical *differences* in mean cumulative progressions for different conditions.

Figure 9 plots cumulative progression differences. Each panel represents one of the three effects in reading times that we are considering here. Each point on curve is obtained by first averaging cumulative progressions across participants and items for one condition at one timestamp, and then calculating the difference between the averages across two conditions that are compared. For the mismatch effect curve, the averaging is done for two pairs of conditions and then the difference between them is calculated. The x-axis represents timestamps starting with the first fixation in the reflexive region and extends till the next 1000 ms; two consecutive timestamps are 2 ms apart since the eye tracker sampled every 2 ms (which means each curve is composed of 501 points). The y-axis represents the difference in pixels between averaged progressions of conditions that are compared. It is crucial to note here that this is an analytic approach, and it is only an exploratory data analysis. Since each data point in this figure is averaged across participants and items, it underestimates variance between participants and items. Moreover, the 95% confidence intervals are underestimates, and with more conservative approach the plots may look consistent with noise. Overall, we need a more rigorous statistical analysis to do justice to the conclusions we are drawing based on cumulative progressions.

The curve representing the mismatch-interference effect diverges at 434 ms from the x-axis on the positive side and remains on the positive side. This implies that after the

first fixation in the reflexive region, from 434 ms onwards, participants speed up in the mismatch-interference condition compared to the mismatch condition. This, in turn, is consistent with the mismatch-interference effect (E5) predicted by the model. The curve representing the mismatch effect diverges at 528 ms from the x-axis on the negative side and remains on the negative side. This implies that the two mismatch conditions are read slower than the two match conditions from 528 ms onwards, after the first fixation in the reflexive region, which is consistent with the mismatch effect (E3) predicted by the model. However, the match-interference effect curve diverges from the x-axis, initially on the positive side at 420 ms and then switches to the negative side at 754 ms and then predominantly remains on the negative side. The diversion in the positive direction is opposite to what the model predicts, but the later diversion to the negative side is consistent with the predictions of the model. Effectively, if we assume that faster cumulative progressions from the reflexive region reflect faster retrievals at the reflexive, the mismatch-interference effect (E5) and the mismatch effect (E3) are visible in the cumulative progressions. Interestingly, the mismatch-interference effect also starts earlier than the mismatch effect.

3.6.5. Timing of Mismatch and Interference Effects

It is important to note that the predictions of the model are not specific to early or late measures, but we expect that both mismatch and interference effects should occur during the same time frame in an experiment because, in the model, both the effects take place during the same (sub)process. Moreover, the absence of early mismatch effect in our experiment (and also in Cunnings and Felser, 2013 and Cunnings and Sturt, 2014) does not support the argument in Sturt (2003) that binding accessible antecedents influence early stages of processing, albeit it need not necessarily speak *against* Sturt's argument either, because early effects don't always show up in early measures (Vasishth et al., 2013).

In sum, the eye tracking study, through various measures, supported the predictions of the cue-based retrieval model of reflexive binding that assumes gender of the reflexive as one of the retrieval cues. The mismatch effects E1 and E3 were observed in total reading time, re-reading probability and in cumulative progressions from the reflexive. The interference effect E2 was observed in first pass regression probability. The interference effect E4 was observed in first pass regression probability and a trend of this effect was observed in the regression contingent first fixation duration. The interference effect E5 was observed in re-reading probability, cumulative progressions from the reflexive, and there was also a trend of this effect in first fixation durations that were followed by regressions. Along with replicating the mismatch effects, observed in Sturt's experiments, the presence of interference effects (E4 and E5), which were absent in Sturt's experiments, makes the results consistent with the model's predictions.

4. GENERAL DISCUSSION

In this paper, we investigated the question: what kinds of cues are used initially by the parser when resolving antecedent-reflexive relations? The two positions on this question are: early use of only structural cues (Nicol and Swinney, 1989; Sturt, 2003; Xiang et al., 2009; Phillips et al., 2011; Dillon et al., 2013), or early use of structural as well as other cues such as gender marking (Badecker and Straub, 2002). We framed the theoretical question within a computational model of sentence processing, the cue-based retrieval model proposed in Lewis and Vasishth (2005) and Lewis et al. (2006), and showed that if we assume that cue-based retrieval involves structural as well as non-structural cues, the model makes five predictions, repeated below:

E1. Mismatch effect (RE): the retrieval errors for the two mismatch conditions are higher than those for the two match conditions.

E2. Interference effect (RE): the retrieval errors for the match-interference and mismatch-interference conditions are higher than those for the other two conditions.

E3. Mismatch effect (RT): the retrieval times for the two mismatch conditions are longer than those for the two match conditions.

E4. Match-interference effect (RT): the retrieval times for the match-interference condition are longer than the match condition.

E5. Mismatch-interference effect (RT): the retrieval times for the mismatch-interference condition are *shorter* than the mismatch condition.

Effects E1–E3 are attested in Sturt's studies; but effects E4 and E5 are not. We hypothesized that Sturt failed to find effect E4 because the inaccessible antecedent had a different grammatical role (object) than the accessible antecedent (subject); i.e., it was distinct enough from the accessible antecedent to be rejected successfully during search. We predicted that if both the accessible and inaccessible antecedents had the subject role, then a match-interference (E4) effect would occur. Moreover if grammatical cues are weighted heavily in the retrieval process (Van Dyke and McElree, 2011), a subject distractor will induce a higher interference effect. We then conducted an eye tracking study in which both the accessible and inaccessible antecedents had the subject role, thereby increasing their similarity. We showed that in first pass regression probability a match-interference effect is indeed seen, as predicted by the model. In addition, as an exploratory data analysis, when we separately analyzed the first fixation duration contingent on regressions, the first fixation durations that were followed by regressions showed marginal effects consistent with the two interference effects E4 and E5. These first fixation durations also confirmed the effect E3. The effect E3 was observed in re-reading time and total reading time as well. This result is consistent with the model's predicted mismatch effect in retrieval times, and the predicted mismatch effect in retrieval errors. Further, in another exploratory data analysis with an eye tracking measure called cumulative progressions, which has been claimed to capture processing difficulty on a continuous time scale, we found that the interference effect E5 and the mismatch effect E3 are realized; with E5, in fact, occurring earlier than E3. Though the analysis with cumulative progressions involved only visual inspection, the

visual patterns are consistent with these two effects predicted by the model.

In sum, the eye tracking study provided empirical evidence for all the effects predicted by the model, including the interference effects that were not observed in the earlier studies such as Sturt (2003) and Xiang et al. (2009). There was clear support for the mismatch and match-interference effect predicted by the model. Although the support for the mismatch-interference effect was not equally clear—it was only marginally significant in two of the eye tracking measures, and there was some evidence in the exploratory data analysis with cumulative progressions—the two interference effects have important theoretical implications for the generality of the retrieval mechanisms in sentence processing, and so should not be ignored.

The interference effects and the mismatch effect have also been observed in some other studies[5]. The mismatch effect E3 has been found in the reading studies (eye tracking and/or self-paced reading) such as Cunnings and Felser (2013) (Experiment 1 and 2), Cunnings and Sturt (2014) (Experiment 1), Dillon et al. (2013) (Experiment 1 and 2), King et al. (2012), Parker and Phillips (2014) (Experiment 1, 2, and 3), and Sturt and Kwon (2013) (Experiment 3 and 4) with a design comparable to Sturt's experiment 1. The interference effect E4 has been found in the reading studies such as Badecker and Straub (2002) (Experiment 3 and 4) and Mansbridge and Witzel (2012), and the interference effect E5 has been found in the reading studies such as Cunnings and Felser (2013) (Experiment 2 in high working memory span readers), King et al. (2012), Parker and Phillips (2014) (Experiment 2 and 3) Sturt and Kwon (2013) (Experiment 3 and 4). In the visual world paradigm, an effect equivalent to the interference effect E4 has been reported in Choy and Thompson (2010), Clackson et al. (2011), Runner and Head (2014), and Thompson and Choy (2009). Overall the pattern appears to be that the mismatch effect has been observed robustly, although there are at least a handful of studies reporting the two interference effects as well.

4.1. Why Were the Interference Effects Found Less Often in Earlier Studies?

Apart from the reasons mentioned in the motivation for the design of the experiment reported here, namely the proximity of the inaccessible antecedent to the reflexive and it being the subject of the clause, there could be other reasons for the absence of the interference effect. The absence of the effect could just be a failure to find an effect that in fact exists, which may happen due to low power of the experiment. For example, the effect could be masked by other confounding variables. Indeed, Cunnings and Felser (2013, p. 23) found that participants with high working memory spans show (in first fixation duration) an effect in exactly the direction predicted by the cue-based retrieval model (though they didn't interpret the effect as an interference or intrusion effect). It is participants with low working memory span who show longer first fixation durations in the interference condition in the mismatch cases. If one were to ignore the working memory span in the Cunnings and Felser data, the two differently-signed effects by span would cancel out, showing no difference between the interference and no-interference condition in the mismatch cases, exactly as found in the literature. Thus, since our data and all previous experiments (except, of course, Cunnings and Felser's) do not take working memory capacity into account as a variable, it is quite possible that we are missing an effect that is correctly predicted by the model. Of course, this raises the question that the ACT-R model as currently implemented does not explicitly model high working memory capacity participants. In future work, we intend to explore the role of working memory capacity in triggering the mismatch interference effect.

Another possibility could be that the interference effects are not as strong as the mismatch effect. The model, in fact, predicts numerically smaller interference effects compared to the mismatch effect (see **Figure 4**). Recently, Parker and Phillips (2014) using sentences as in (10), found that the mismatch-interference effect is visible when the reflexive mismatches two features, such as the number and gender (e.g., herself and schoolboys), with the accessible antecedent, but not when it mismatches only one feature. King et al. (2012) using sentences as in (11), found that the mismatch-interference effect is visible when the reflexive is not adjacent to the verb [condition (b) in (11)] allowing the information about the verb's argument structure, and hence the information about the accessible antecedent, to decay. These results possibly corroborate the model's prediction that the mismatch-interference effect is weaker than the mismatch effect, and hence difficult to detect.

(10) The {librarian/janitor} said that the {schoolgirl/schoolboy/schoolboys} reminded herself about the book.

(11) a. *Verb-adjacent:* The mechanic who spoke to {John/Mary} sent {himself/herself} a package.
b. *Verb-non-adjacent:* The mechanic who spoke to {John/Mary} sent a package to {himself/herself}.

Effectively, our results are not only in line with the results from Cunnings and Felser (2013), King et al. (2012), and Parker and Phillips (2014), among others, but also provide convincing evidence for the model's predictions with manipulations independent of the memory span of the participants, with a configuration involving verb-adjacent reflexives, and with lowest possible (= single) mismatch of retrieval cues with the accessible antecedent.

4.1.1. Strictly Structured Access as an Alternative

Here, we discuss the strictly structured retrieval approach proposed in Dillon et al. (2013) and Phillips et al. (2011) for resolving reflexive-antecedent dependency, and examine its claims in the light of existing experimental and modeling findings. Although Phillips and colleagues refers to the

[5]In this paper we are considering only argument reflexives and only in English. But interference effects have also been reported for Mandarin reflexives (Jäger et al., 2015b; Chen et al., 2016) and for non-argument positions such as English reflexives inside picture noun phrases (Runner et al., 2006). On the other hand, for the pronoun-antecedent dependency (which is subject to different grammatical constraints than the reflexive-antecedent dependency), Chow et al. (2014) failed to replicate the interference effect observed in the pronoun experiments reported in Badecker and Straub (2002).

mechanism as structured access, we refer to it as a *strictly* structured access to emphasize the point that the approach suggested in this paper does not ignore the structural constraints, but it includes other constraints as well.

Dillon et al. (2013) supported evidence for the strictly structured access with a set of computational and experimental studies involving English reflexives and subject-verb agreement. This experiment essentially replicated the *interference asymmetry* effect from Wagers et al. (2009) and the absence of interference effect in processing English reflexives from Sturt (2003). Based on these results, they concluded that agreement dependency and reflexive dependency employ different retrieval mechanisms for resolving the dependencies—agreement dependencies are resolved using morphological features of the target noun phrase whereas the antecedent for a reflexive is retrieved using *only* structural constraints. They further compared the predictions of a strictly structural cue based retrieval model of reflexives to a model utilizing mixed cues—structural as well as agreement. The mixed cue model predicted an interference effect in retrieval errors (similar to E2 above) and a mismatch-interference effect in retrieval times (similar to E5 above). The prediction of the match-interference effect (E4) was not reliably non-zero, in the sense that for some parameter combinations the model didn't predict any difference between the match and match-interference condition. The structural cue based model predicted no interference effect in either retrieval errors or retrieval times. The mismatch effects in retrieval errors and retrieval times (E1 and E3), as predicted by the mixed cue model, were not discussed in Dillon et al. (2013). On the basis of these predictions, Dillon et al. (2013) concluded that the strictly structured access model captures the reflexive binding data from their Experiments better than the mixed cue model.

Although Dillon et al. (2013) replicated the findings in Sturt (2003), the lack of interference effect is subject to the same alternative explanation that we suggested for Experiment 2 in Sturt (2003): we hypothesized that reflexive binding uses the grammatical role *subject* as one of the retrieval cues for retrieving the correct antecedent. The absence of the interference effect could be due to (apart from power concerns) the fact that the interfering antecedent had an *object* role in the experiments above, which does not match one of the retrieval cues, reducing the strength of interference. Badecker and Straub (2002) also reported that the interference effect is found when the interferer is in the subject position. Moreover, Van Dyke and McElree (2011) found that, in thematic binding, the interference effect due to the semantic match was present only when the distractors matched syntactic cues along with semantic cues. If the retrieval process gives higher weight to syntactic cues than semantic cues, the absence of interference effect could simply be due to the absence of matching a grammatical role in the inaccessible antecedent.

The predictions of the structured access model hold only for a limited set of experiments and a limited set of effects in those experiments. The mismatch-effects (E1 and E3) that have been replicated in various studies like Sturt (2003), Cunnings and Felser (2013) and also the one reported in this paper cannot be explained by this model. The structured access model predicts

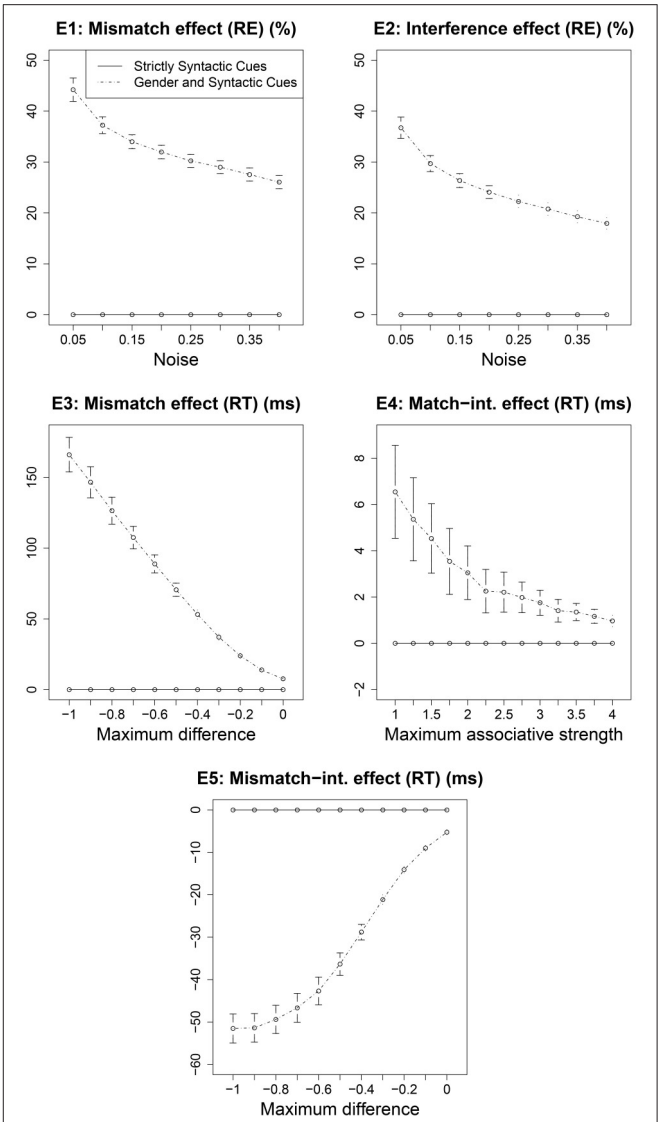


FIGURE 10 | The predictions of a strictly structural cue-based model and a model with structural and agreement feature based cues. The predictions are for the new experimental design and are generated across a set of parameter values.

no difference between match and mismatch conditions (see **Figure 10**). Furthermore, the interference-effects (E2, E4, and E5) observed in various reflexives studies like Badecker and Straub (2002) Experiment 3, Sturt (2003) follow-up study, Cunnings and Felser (2013) Experiment 2 and the one reported here cannot be explained by this model. Consequently, a model assuming structural as well as agreement features as retrieval cues predicts a broader range of data than a strictly structured access model.

Dillon et al. (2013) further claimed that the match-interference effect (E4)—higher reading times when the inaccessible noun matches gender or number of the reflexive—is not reliable evidence for interference from the grammatically inaccessible antecedent, for mainly two reasons: (1) The cue-based retrieval model doesn't predict any difference in retrieval

times between the match and match-interference conditions for certain parameter combinations, since on the one hand the cue-overlap (gender and number) between accessible and inaccessible antecedents leads to an inhibitory effect, and on the other hand the retrieval of the inaccessible antecedent leads to a facilitatory effect; (2) The match-interference effects can be explained in terms of feature-overwriting (Nairne, 1990; Gordon et al., 2001, 2004, 2006; Oberauer and Kliegl, 2006) instead of interference at the time of retrieval. Consequently, Dillon proposed that only a facilitatory effect in mismatch-interference can be considered as evidence for retrieval interference.

As far as the first argument is concerned, we, in fact, show that the cue-based retrieval model with mixed cues consistently predicts a positive match-interference effect across a set of parameter values for Sturt's two experiments (see **Figures 4**, **5**). Although the effect for the modified design is not predicted to be positive for all combinations of parameter values, for a certain set of combination of values the effect is non-zero and positive, and only for a very small set of parameter values is the effect predicted to be zero.

The second argument, in fact, applies to the mismatch-interference effect as well—the mismatch-interference effect (for this particular design) can also be explained in terms feature-overwriting or encoding interference. Encoding interference is a consequence of the feature overlap between the accessible and inaccessible antecedents. In the design discussed here, the feature overlap between the accessible and inaccessible antecedents is, in fact, higher in the mismatch condition compared to the mismatch-interference condition ("soldier" and "Fred" have the same gender in the mismatch condition whereas "soldier" and "Katie" have different genders in the mismatch-interference condition). This means between these two conditions, the interference conditions are reversed for encoding and retrieval interference. Retrieval interference predicts faster reading time for the mismatch-interference condition while encoding interference predicts *slower* reading time for the *mismatch* condition, leading to exactly the same pattern of retrieval times between the two mismatch conditions. Effectively, this configuration makes it impossible to tease apart the two types of interference theories using the experiment design considered in this paper or in similar earlier studies including Dillon et al. (2013). However, Jäger et al. (2015a) using self-paced reading and eye-tracking studies with German and Swedish reflexives, compared the predictions of the two interference theories. They could not find any evidence for encoding interference and concluded that "invoking encoding interference may not be a plausible way to reconcile interference effects with a structure-based account of reflexive processing." If we assume that the retrieval process for reflexives in German, Swedish and English are similar (especially because these are closely related languages) then we can safely conclude that, even though our design does not have the possibility of disentangling the two retrieval theories, the effects that we see in our experiment are driven by retrieval interference.

In summary, we have presented a theory and computational model of the access of antecedents for reflexive pronouns in English, and used this theory to gain insight into empirical studies that have yielded mixed results concerning the putative role of non-structural cues. We used this analysis and the results of further modeling to motivate a new empirical design that formed the basis of an eye tracking study. Many of the results of the eye tracking study are consistent with the model's assumptions concerning the early use of non-structural cues. These results present a challenge for theories advocating the infallibility of the human parser in the case of reflexive binding in English, and provide support for the inclusion of agreement features such as gender in the set of retrieval cues. In general, the results provide further support for the deployment of a rapid, parallel cue-based access mechanism in service of sentence parsing (McElree, 2000; McElree et al., 2003; Lewis and Vasishth, 2005; Lewis et al., 2006), and help to sharpen deeper explanatory questions concerning the utility and selection of cues.

AUTHOR CONTRIBUTIONS

UP wrote the model, SV and RL supervised the process. UP conceived, setup, and carried out the experiment. UP and SV analyzed the data. UP, SV, and RL wrote the manuscript.

ACKNOWLEDGMENTS

We would like to thank William Badecker, Brian Dillon, Sol Lago, Colin Phillips, Rukshin Shaher, Ming Xiang, and the two reviewers for their helpful comments, and to Felix Engelmann, Pavel Logačev and Titus von der Malsburg for their technical help at various stages of the eye tracking study. The eye tracking experiment reported here was run at the ZAS in Berlin; many thanks to Manfred Krifka for providing space to run this study. Thanks for the support of the Deutsche Forschungsgemeinschaft (Project No. BO 2142/1-1) to UP, and the National Science Foundation under Grant BCS-1152819 to RL.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpsyg.2016.00329

REFERENCES

Aho, A. V., and Ullman, J. D. (1972). *The Theory of Parsing, Translation and Compiling Volume I: Parsing*. Upper Saddle River, NJ: Prentice-Hall.

Alcocer, P., and Phillips, C. (2012). *Using Relational Syntactic Constraints in Content-Addressable Memory Architectures for Sentence Parsing*. Unpublished manuscript. Available online at: http://www.colinphillips.net/wp-content/uploads/2014/08/alcocer_phillips2012_v2.pdf (Retrieved March 4, 2016).

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., and Qin, Y. (2004). An integrated theory of the mind. *Psychol. Rev.* 111, 1036–1050. doi: 10.1037/0033-295X.111.4.1036

Badecker, W., and Straub, K. (2002). The processing role of structural constraints on the interpretation of pronouns and anaphors. *J. Exp. Psychol. Learn. Mem. Cogn.* 28, 748–769. doi: 10.1037/0278-7393.28.4.748

Bates, D., and Sarkar, D. (2007). *lme4: Linear Mixed-Effects Models Using S4 Classes (R Package Version 0.9975-11)*.

Chen, Z., Jäger, L., and Vasishth, S. (2016). "How structure sensitive is the parser? Evidence from Mandarin Chinese," in *Empirical Approaches to Linguistic Theory: Studies of Meaning and Structure, Studies in Generative Grammar*, eds B. Stolterfoht and S. Featherston (Berlin: Mouton de Gruyter), 43–62.

Chomsky, N. (1981). *Lectures on Government and Binding*. Berlin: Mouton de Gruyter.

Chomsky, N. (1986). *Barriers*. Cambridge, MA: MIT Press.

Chow, W. Y., Lewis, S., and Phillips, C. (2014). Immediate sensitivity to structural constraints in pronoun resolution. *Front. Psychol.* 5:630. doi: 10.3389/fpsyg.2014.00630

Choy, J. J., and Thompson, C. K. (2010). Binding in agrammatic aphasia: processing to comprehension. *Aphasiology* 24, 551–579. doi: 10.1080/02687030802634025

Clackson, K., Felser, C., and Clahsen, H. (2011). Children's processing of reflexives and pronouns in English: evidence from eye-movements during listening. *J. Mem. Lang.* 65, 128–144. doi: 10.1016/j.jml.2011.04.007

Clifton, C., Staub, A., and Rayner, K. (2007). "Eye movements in reading words and sentences," in *Eye Movements: A Window on Mind and Brain*, ed R. van Gompel (Amsterdam: Elsevier), 341–372.

Cunnings, I., and Felser, C. (2013). The role of working memory in the processing of reflexives. *Lang. Cogn. Process.* 28, 188–219. doi: 10.1080/01690965.2010.548391

Cunnings, I., and Sturt, P. (2014). Coargumenthood and the processing of reflexives. *J. Mem. Lang.* 75, 117–139. doi: 10.1016/j.jml.2014.05.006

Dillon, B., Mishler, A., Sloggett, S., and Phillips, C. (2013). Contrasting intrusion profiles for agreement and anaphora: experimental and modeling evidence. *J. Mem. Lang.* 69, 85–103. doi: 10.1016/j.jml.2013.04.003

Engbert, R., and Kliegl, R. (2003). Microsaccades uncover the orientation of covert attention. *Vision Res.* 43, 1035–1045. doi: 10.1016/S0042-6989(03)00084-1

Frazier, L. (1979). *On Comprehending Sentences: Syntactic Parsing Strategies*. PhD thesis, University of Massachusetts, Amherst, MA.

Gelman, A., and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. New York, NY: Cambridge University Press.

Gordon, P. C., Hendrick, R., and Johnson, M. (2001). Memory interference during language processing. *J. Exp. Psychol. Learn. Mem. Cogn.* 27, 1411–1423. doi: 10.1037/0278-7393.27.6.1411

Gordon, P. C., Hendrick, R., and Johnson, M. (2004). Effects of noun phrase type on sentence complexity. *J. Mem. Lang.* 51, 97–114. doi: 10.1016/j.jml.2004.02.003

Gordon, P., Hendrick, R., Johnson, M., and Lee, Y. (2006). Similarity-based interference during language comprehension: evidence from eye tracking during reading. *J. Exp. Psychol. Learn. Mem. Cogn.* 32, 1304–1321. doi: 10.1037/0278-7393.32.6.1304

Jäger, L. A., Benz, L., Roeser, J., Dillon, B. W., and Vasishth, S. (2015a). Teasing apart retrieval and encoding interference in the processing of anaphors. *Front. Psychol.* 6:506. doi: 10.3389/fpsyg.2015.00506

Jäger, L. A., Engelmann, F., and Vasishth, S. (2015b). Retrieval interference in reflexive processing: experimental evidence from Mandarin, and computational modeling. *Front. Psychol.* 6:617. doi: 10.3389/fpsyg.2015.00617

King, J., Andrews, C., and Wagers, M. (2012). "Do reflexives always find a grammatical antecedent for themselves?," in *The 25th Annual CUNY Conference on Human Sentence Processing*, New York, NY.

Kreiner, H., Sturt, P., and Garrod, S. (2008). Processing definitional and stereotypical gender in reference resolution: evidence from eye-movements. *J. Mem. Lang.* 58, 239–261. doi: 10.1016/j.jml.2007.09.003

Lewis, R. L., and Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cogn. Sci.* 29, 375–419. doi: 10.1207/s15516709cog0000_25

Lewis, R. L., Vasishth, S., and Van Dyke, J. A. (2006). Computational principles of working memory in sentence comprehension. *Trends Cogn. Sci.* 10, 447–454. doi: 10.1016/j.tics.2006.08.007

Mansbridge, M. P., and Witzel, J. (2012). "Binding accessibility and online anaphora processing," in *The 25th Annual CUNY Conference on Human Sentence Processing*, New York, NY.

McElree, B. (2000). Sentence comprehension is mediated by content-addressable memory structures. *J. Psycholinguist. Res.* 29, 111–123. doi: 10.1023/A:1005184709695

McElree, B., Foraker, S., and Dyer, L. (2003). Memory and language memory structures that subserve sentence comprehension. *J. Mem. Lang.* 48, 67–91. doi: 10.1016/S0749-596X(02)00515-6

Nairne, J. S. (1990). A feature model of immediate memory. *Mem. Cogn.* 18, 251–269. doi: 10.3758/BF03213879

Nicol, J., and Swinney, D. (1989). The role of structure in coreference assignment during sentence comprehension. *J. Psycholinguist. Res.* 18, 5–19. doi: 10.1007/BF01069043

Oberauer, K., and Kliegl, R. (2006). A formal model of capacity limits in working memory. *J. Mem. Lang.* 55, 601–626. doi: 10.1016/j.jml.2006.08.009

Osterhout, L., Bersick, M., and McLaughlin, J. (1997). Brain potentials reflect violations of gender stereotypes. *Mem. Cogn.* 25, 273–285. doi: 10.3758/BF03211283

Parker, D., and Phillips, C. (2014). "Selective priority for structure in memory retrieval," in *The 27th Annual CUNY Conference on Human Sentence Processing*, Columbus, OH.

Phillips, C., Wagers, M. W., and Lau, E. F. (2011). "Grammatical illusions and selective fallibility in real-time language comprehension," in *Experiments at the Interfaces (Syntax and Semantics)*, Vol. 37, ed J. T. Runner (Bingley: Emerald Group Publishing Limited), 147–180.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. ISBN: 3-900051-07-0.

Runner, J. T., and Head, K. D. L. (2014). "What can visual world eye-tracking tell us about the binding theory?," in *Empirical Issues in Syntax and Semantics 10*, ed C. Piñón (Paris: CSSP—Colloque de Syntaxe et Sémantique à Paris), 269–286.

Runner, J. T., Sussman, R. S., and Tanenhaus, M. K. (2006). Processing reflexives and pronouns in picture noun phrase. *Cogn. Sci.* 30, 193–241. doi: 10.1207/s15516709cog0000_58

Sturt, P. (2003). The time-course of the application of binding constraints in reference resolution. *J. Mem. Lang.* 48, 542–562. doi: 10.1016/S0749-596X(02)00536-3

Sturt, P., and Kwon, N. (2013). "The processing of raising and nominal control," in *The 26th Annual CUNY Conference on Human Sentence Processing*, Columbia, SC.

Thompson, C. K., and Choy, J. J. (2009). Pronominal resolution and gap filling in agrammatic aphasia: evidence from eye movements. *J. Psycholinguist. Res.* 38, 255–283. doi: 10.1007/s10936-009-9105-7

Van Dyke, J. A., and McElree, B. (2011). Cue-dependent interference in comprehension. *J. Mem. Lang.* 65, 247–263. doi: 10.1016/j.jml.2011.05.002

Vasishth, S., Brüssow, S., Lewis, R. L., and Drenhaus, H. (2008). Processing polarity: how the ungrammatical intrudes on the grammatical. *Cogn. Sci.* 32, 685–712. doi: 10.1080/03640210802066865

Vasishth, S., and Lewis, R. L. (2006). Argument-head distance and processing complexity: explaining both locality and antilocality effects. *Language* 82, 767–794. doi: 10.1353/lan.2006.0236

Vasishth, S., von der Malsburg, T., and Engelmann, F. (2013). What eye movements can tell us about sentence comprehension. *Wiley Interdisc. Rev. Cogn. Sci.* 4, 125–134. doi: 10.1002/wcs.1209

Wagers, M. W., Lau, E. F., and Phillips, C. (2009). Agreement attraction in comprehension: representations and processes. *J. Mem. Lang.* 61, 206–237. doi: 10.1016/j.jml.2009.04.002

Wong, T. J., Cokel, E. T., and Schooler, L. J. (2010). *An Online Database of ACT-R Parameters: Towards a Transparent Community-Based Approach to Model*. Philadelphia, PA: Drexel University.

Xiang, M., Dillon, B., and Phillips, C. (2009). Illusory licensing effects across dependency types: ERP evidence. *Brain Lang.* 108, 40–55. doi: 10.1016/j.bandl.2008.10.002

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2016 Patil, Vasishth and Lewis. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 27 May 2015
doi: 10.3389/fpsyg.2015.00617

Retrieval interference in reflexive processing: experimental evidence from Mandarin, and computational modeling

*Lena A. Jäger *, Felix Engelmann and Shravan Vasishth*

Department of Linguistics, University of Potsdam, Potsdam, Germany

We conducted two eye-tracking experiments investigating the processing of the Mandarin reflexive *ziji* in order to tease apart structurally constrained accounts from standard cue-based accounts of memory retrieval. In both experiments, we tested whether structurally inaccessible distractors that fulfill the animacy requirement of *ziji* influence processing times at the reflexive. In Experiment 1, we manipulated animacy of the antecedent and a structurally inaccessible distractor intervening between the antecedent and the reflexive. In conditions where the accessible antecedent mismatched the animacy cue, we found inhibitory interference whereas in antecedent-match conditions, no effect of the distractor was observed. In Experiment 2, we tested only antecedent-match configurations and manipulated locality of the reflexive-antecedent binding (Mandarin allows non-local binding). Participants were asked to hold three distractors (animate vs. inanimate nouns) in memory while reading the target sentence. We found slower reading times when animate distractors were held in memory (inhibitory interference). Moreover, we replicated the locality effect reported in previous studies. These results are incompatible with structure-based accounts. However, the cue-based ACT-R model of Lewis and Vasishth (2005) cannot explain the observed pattern either. We therefore extend the original ACT-R model and show how this model not only explains the data presented in this article, but is also able to account for previously unexplained patterns in the literature on reflexive processing.

Keywords: Chinese reflexives, ACT-R, eye-tracking, interference, cue-based retrieval, computational modeling, ziji, content-addressable memory

OPEN ACCESS

Edited by:
Colin Phillips,
University of Maryland, USA

Reviewed by:
Brian Dillon,
University of Massachusetts Amherst, USA
Patrick Sturt,
University of Edinburgh, UK

****Correspondence:***
Lena A. Jäger,
Department of Linguistics, University of Potsdam, Karl-Liebknecht-Str. 24-25, Potsdam 14476, Germany
lena.jaeger@uni-potsdam.de

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: *11 November 2014*
Accepted: *26 April 2015*
Published: *27 May 2015*

Citation:
Jäger LA, Engelmann F and Vasishth S (2015) Retrieval interference in reflexive processing: experimental evidence from Mandarin, and computational modeling. Front. Psychol. 6:617. doi: 10.3389/fpsyg.2015.00617

1. Introduction

One major task the human parser has to accomplish is to syntactically link together two or more linguistic elements that are not adjacent to each other. For example, when a reflexive is being processed, it has to be somehow linked to its antecedent even if there is intervening material. Therefore, one central question in psycholinguistics is what mechanisms the human parser uses to identify and retrieve the previously processed part of a dependency. Theoretically, there are different options how this identification and retrieval of a linguistic constituent from working memory might be accomplished: different kinds of search mechanisms on the one hand (Sternberg, 1966, 1969) and cue-based, i.e., content-addressable, retrieval on the other

hand (McElree and Dosher, 1989; Anderson and Lebiere, 1998; Anderson et al., 2004).[1] In general, a search mechanism checks certain items in memory based on their location in order to find the target. Cue-based retrieval, in contrast, assumes that retrieval targets are content-addressable and can be accessed directly by the use of certain features as retrieval cues. Over the last decade, evidence favoring a content-addressable memory underlying human sentence processing has accumulated (McElree, 2000, 2003; McElree et al., 2003; Van Dyke and McElree, 2006; Martin and McElree, 2008).

In the case of English reflexives, retrieval cues used in a content-addressable memory might be non-structural cues like gender or number along with structural cues like local c-command. Note that a reflexive's binding domain varies between languages (Büring, 2005; Reuland, 2011). Whereas in English it can be approximated by the local clause, in Chinese the reflexive *ziji* can be bound across clause boundaries (non-local binding; for a brief overview of the syntactic properties of Chinese *ziji* see below). For the sake of simplicity, we will refer to the structural feature of *c-commanding the reflexive and being contained in its binding domain* briefly as the *c-command* feature.

However, within the framework of cue-based retrieval, it is still an open question which features the parser uses as retrieval cues. On the one hand, it has been proposed that all available cues are used for retrieval with equal weights being applied to all cues (Lewis and Vasishth, 2005). We will refer to this account as the *standard cue-based retrieval account*. On the other hand, Van Dyke (2007) and Van Dyke and McElree (2011) and others argue that syntactic cues (being in a certain tree-configurational position) have some kind of priority over non-syntactic cues. In particular, it has been proposed that for the processing of reflexive-antecedent dependencies, the set of features used for retrieving a reflexive's antecedent is limited to syntactic cues such as c-command within the reflexive's binding domain (Nicol and Swinney, 1989; Sturt, 2003; Xiang et al., 2009; Phillips et al., 2011; Dillon et al., 2013; Kush and Phillips, 2014). We will refer to this proposal as *structure-based account*.

If a structure-based retrieval is applied, a noun phrase that is in a structural position that disqualifies it from being the reflexive's antecedent should not have any effect on the processing of the reflexive-antecedent dependency, no matter whether it matches non-structural features of the reflexive such as gender or number. Thus, in a sentences like (1), the gender of *Jonathan* or *Jennifer* should not affect processing times of the reflexive since they do not c-command it and hence cannot syntactically bind the reflexive.

(1) a. **Antecedent-match; distractor-match**
The *surgeon* who treated *Jonathan* had pricked *himself* …

b. **Antecedent-match; distractor-mismatch**
The *surgeon* who treated *Jennifer* had pricked *himself* …

c. **Antecedent-mismatch; distractor-match**
The *surgeon* who treated *Jennifer* had pricked *herself* …

d. **Antecedent-mismatch; distractor-mismatch**
The *surgeon* who treated *Jonathan* had pricked *herself* …

The parsing architecture developed by Lewis and Vasishth (2005), which is based on Anderson et al. (2004)'s cognitive architecture ACT-R (Adaptive Control of Thought–Rational) assumes a cue-based retrieval mechanism without syntactic constraints. This model has been used to explain interference effects in sentence processing and in reflexives in particular (e.g., Dillon et al., 2013; Parker and Phillips, 2014; Patil, Vasishth, and Lewis, "Retrieval interference in syntactic processing: The case of reflexive binding in English," unpublished manuscript). According to the ACT-R model, both latency and probability of retrieving a certain target item are determined by (i) the quality of the match between retrieval cues and target features and (ii) similarity-based mutual inhibition between the target and other matching items. Retrieval speed and probability increase with the number of cues matching the target. If, however, certain cues match the features of multiple memory items, similarity-based interference leads to a higher retrieval latency, i.e., inhibitory interference effects. The latter is the case in (1a) as compared to (1b), because in (1a) both the target *surgeon* and the distractor *Jonathan* share the feature *+masculine*. In the antecedent-mismatch conditions (1c) vs. (1d), in contrast, the target *surgeon* and the cue-matching distractor *Jennifer* in (1c) do not share the feature *+feminine*, hence, no similarity-based interference arises. Consequently, no inhibition is predicted in (1c) vs. (1d). On the contrary, because both target and distractor only partially match the retrieval cues in (1c), they are equally likely to be retrieved. Compared to (1d), this predicts a higher proportion of incorrect retrievals and a lower average retrieval latency, which is usually referred to as *facilitatory interference* or *intrusion*.

In sum, a major prediction that distinguishes standard cue-based retrieval from models assuming a limitation of the retrieval cues to structural features is that the former entails interference effects from non-target items that match (some of) the cues used for retrieval.[2]

In order to tease apart structure-based from standard cue-based retrieval, interference effects from feature-matching but syntactically illicit antecedents in the processing of reflexive-antecedent dependencies have drawn considerable attention in recent years. Several studies used a feature-match/mismatch

[1] Note that the different models of content-addressable memory differ with respect to their assumptions about the exact nature of similarity-based retrieval interference. While the model proposed by Anderson et al. (2004) predicts similarity-based retrieval interference to be observed in retrieval probabilities as well as in retrieval latencies, the model proposed by McElree (2000) predicts that similarity-based retrieval interference only affects retrieval probabilities and not retrieval latencies. In this article, we will focus on cue-based retrieval in the sense of Anderson et al. (2004).

[2] It should be noted that cueing for a c-command feature is a simplification since it actually is a tree-configurational relation between items. There is no straightforward way to attribute a feature like that in an incremental parsing mechanism in content-addressable memory. In this paper, we do not provide a detailed account of how the attribution of a c-command feature could be implemented. As an example, Patil, Vasishth, and Lewis, "Retrieval interference in syntactic processing: The case of reflexive binding in English" (unpublished manuscript) in their ACT-R model for English reflexives approximated a c-command relation by cueing for a subject in the local clause. For a discussion of possible ways to encode tree-configurational information such as c-command in content-addressable memory see Alcocer and Phillips, "Using relational syntactic constraints in content-addressable memory architectures for sentence processing" (unpublished manuscript).

design, where a non-syntactic feature (e.g., gender or number) was manipulated at the antecedent and at a structurally inaccessible distractor (see Example 1 for typical sentence material). In **Table 1**, we provide an overview of the studies examining interference effects in reflexives (including reflexives inside a prepositional phrase and possessive reflexives) and reciprocals using a feature-match/mismatch design. Studies on the processing of reflexives in so-called picture noun phrases have not been included in our review since their binding properties differ from other reflexives (Büring, 2005; Reuland, 2011). Moreover, experiments investigating specific populations such as children or L2 learners are not considered in the review. **Table 1** summarizes whether or not inhibitory (i.e., a slowdown due to the presence of a cue-matching inaccessible distractor) or facilitatory (i.e., a speed-up due to the presence of a cue-matching inaccessible distractor) interference was observed in (i) conditions with an accessible antecedent that matched the feature under examination and (ii) conditions with an accessible antecedent that mismatched the feature under examination (i.e., sentences that are either ungrammatical or at least violating the stereotypical gender of the accessible antecedent). Some studies manipulated other factors in addition to the feature-match/mismatch manipulation. In these cases, we split the respective experiments into two entries in **Table 1**, with one entry for each level of the additional factor. In particular, for Felser et al. (2009), who manipulated feature type (gender vs. c-command) as additional within-participants factor and language proficiency (native speaker vs. L2 learner) as between-participants factor, one row in **Table 1** refers to the manipulation of the c-command feature in native speakers and another row refers to the gender manipulation in native speakers. The results of the non-native group are not included in the table because this review concerns adult native speaker populations. For Chen et al. (2012), who manipulated whether the Chinese reflexive *ziji* was locally or non-locally bound, one row in **Table 1** refers to the interference effect observed in conditions with a local antecedent and a second row refers to the conditions with a non-local antecedent. Similarly, in the case of King et al. (2012), who manipulated whether the reflexive directly followed the verb or a preposition intervened, one table entry refers to the former configuration (labeled as *adjacent*) and another entry refers to the latter configuration (labeled as *non-adjacent*). In the review of Clackson et al. (2011), who primarily investigated the processing of reflexives in children, we only report the results of the adult control group. For the reviewed experiments, we report effects observed at the region containing the reflexive (labeled as *crit*) and the following regions (labeled as *crit*+*x*). Although the size of the interest areas in terms of number of words contained in one region differs between studies, which reduces the comparability of the time course of the observed effects to a certain extent, we keep the sectioning of the interest areas as in the respective publication.

In accessible antecedent-match conditions, previous studies found inhibitory interference in six cases (Badecker and Straub, 2002, Experiments 1 and 2; Felser et al., 2009, c-command manipulation in native speakers; Chen et al., 2012, non-local reflexives; Clackson and Heyer, 2014; Patil, Vasishth, and Lewis, "Retrieval interference in syntactic processing: The case of reflexive binding in English," unpublished manuscript). Statistically significant facilitatory interference in antecedent-match conditions was found in two experiments (Sturt, 2003, Experiment 1; Cunnings and Felser, 2013, Experiment 2). However, Sturt found the effect only in re-reading time two words after the reflexive and this effect could not be replicated by Cunnings and Sturt (2014), who used similar stimuli. Cunnings and Felser found the effect for readers with low working memory span (*lWM*), but not for high-span readers. In the majority of the experiments, in contrast, no interference effect was observed in antecedent-match conditions (Nicol and Swinney, 1989; Clifton et al., 1999; Badecker and Straub, 2002, Experiments 5 and 6; Sturt, 2003, Experiment 2; Felser et al., 2009, gender manipulation in native speakers; Clackson et al., 2011, adult control group of Experiment 2; Chen et al., 2012, conditions with local reflexive binding; King et al., 2012, adjacent conditions; Cunnings and Felser, 2013, Experiment 1; Dillon et al., 2013; Kush and Phillips, 2014; Cunnings and Sturt, 2014, Experiment 1; Parker and Phillips, 2014).[3]

For conditions with a feature-mismatching accessible antecedent, two studies report significant effects of facilitatory interference (King et al., 2012; Parker and Phillips, 2014) and two studies report a marginal facilitatory effect (Cunnings and Felser, 2013, Experiment 1; Patil, Vasishth, and Lewis, "Retrieval interference in syntactic processing: The case of reflexive binding in English," unpublished manuscript)—however, the latter effect was only found in a *post-hoc* analysis of regression-contingent first-fixation durations, and thus might be spurious. Marginal effects of inhibitory interference have been reported for participants with low working memory span (Cunnings and Felser, 2013, Experiment 2), in the processing of reciprocals (Kush and Phillips, 2014), and in Experiment 1 of Cunnings and Sturt (2014). The latter only report a marginal main effect of the distractor, but their reported means suggest that the effect was driven by the antecedent-mismatch conditions. This does, however, not seem very reliable because they used similar stimuli as Sturt (2003), Experiment 1, who, in contrast, had not found an effect in antecedent-mismatch conditions but a facilitation in antecedent-match conditions. A general pattern is that interference effects in antecedent-match conditions are less frequently observed than effects in antecedent-mismatch conditions.

To summarize, the literature on reflexive interference contains a mixture of results, not favoring one particular of the retrieval models in question. Studies showing a general absence of interference support structure-based accounts (Nicol and Swinney, 1989; Sturt, 2003; Xiang et al., 2009; Phillips et al., 2011; Dillon, 2011; Dillon et al., 2013; Kush and Phillips, 2014). On the other hand, observations of significant interference effects have been interpreted as evidence against purely structure-based retrieval (Badecker and Straub, 2002; Chen et al., 2012; Clackson and Heyer, 2014; Parker and Phillips, 2014). Crucially, however, taking into account the direction of the effects, there are patterns that cannot be explained by either account without

[3] King et al. (2012) report different results in their CUNY 2012 abstract and their final conference poster. We refer here to the results presented on the poster.

TABLE 1 | Overview of interference effects at reflexives categorized for feature-match/mismatch of the accessible antecedent and direction of the effect (inhibitory vs. facilitatory interference).

	Publication	Lang.	Method	Cue	Depend. type	Interf. type	Distractor position	Effect in antecedent-match		Effect in antecedent-mismatch	
								Interference	*AOI, Measure*	*Interference*	*AOI, Measure*
1	Xiang et al. '09	EN	EEG	gend	refl	retro	subj.	—		(inhib)	
2	Badecker and Straub '02 Exp5	EN	SPR	gend	refl	pro	genitive	*n.s.*		—	
3	Badecker and Straub '02 Exp6	EN	SPR	gend	refl	pro	prep.obj.	*n.s.*		—	
4	Chen et al. '12 local	CN	SPR	anim	poss.refl.	retro	subj.	*n.s.*		—	
5	Clackson et al. '11 Exp2 adults	EN	VW	gend	prep.refl.	pro	subj.(topic)	*n.s.*		—	
6	Clifton et al. '99 Exp1	EN	SPR	num	refl	retro	prep.obj.	*n.s.*		—	
7	Clifton et al. '99 Exp2	EN	SPR	gend,num	refl	retro	prep.obj.	*n.s.*		—	
8	Clifton et al. '99 Exp3	EN	SPR	gend	refl	pro	subj	*n.s.*		—	
9	Felser et al. '09 natives, gend	EN	ET	gend	refl	pro	subj.(topic)	*n.s.*		—	
10	Nicol and Swinney '89	EN	Primg	gend	refl	pro	subj.,obj.	*n.s.*		—	
11	Dillon et al. '13	EN	ET	num	refl	retro	obj.	*n.s.*		*n.s.*	
12	King et al. '12 adjacent	EN	ET	gend	refl	retro	obj.	*n.s.*		*n.s.*	
13	Sturt '03 Exp2	EN	ET	gend	refl	retro	obj.(topic)	*n.s.*		*n.s.*	
14	Cunnings and Felser '13 Exp1	EN	ET	gend	refl	pro	subj.(topic)	*n.s.*		(facil)	crit+2, FPRT
15	King et al. '12 non-adjacent	EN	ET	gend	prep.refl.	retro	prep.obj.	*n.s.*		facil	crit, FPRT
16	Parker and Phillips '14	EN	ET	nu/ge/an	refl	pro	subj.	*n.s.*		facil	crit, TFT
17	Cunnings and Sturt '14 Exp1	EN	ET	gend	refl	pro	subj.(topic)	*n.s.*		(inhib)	crit+1, FPRT
18	Kush and Phillips '14	HI	SPR	num	recipr	retro	prep.obj.	*n.s.*		(inhib)	crit+2
19	**This study Exp1**	CN	ET	anim	refl	retro	subj.	*n.s.*		inhib	crit, FFD / FPRT / RBRT / RPD
20	Cunnings and Felser '13 Exp2	EN	ET	gend	refl	retro	subj.(topic)	facil *IWM*	crit, FFD / FPRT	(inhib *IWM*)	crit, FFD
21	Sturt '03 Exp1	EN	ET	gend	refl	pro	subj.(topic)	facil	crit+2, RRT	*n.s.*	
22	**This study Exp2**	CN	ET	anim	refl	pro	memory	inhib	crit, FPRT / RBRT / RPD / TFT	—	
23	Badecker and Straub '02 Exp3	EN	SPR	gend	refl	pro	subj.	inhib	crit+1	—	
24	Badecker and Straub '02 Exp4	EN	SPR	num	recipr	pro	subj.	inhib	crit+1–crit+4	—	
25	Chen et al. '12 non-local	CN	SPR	anim	poss.refl.	retro	subj.	inhib	crit+1	—	
26	Clackson and Heyer '14	EN	VW	gend	prep.refl.	pro	subj.(topic)	inhib	gaze shift	—	
27	Felser et al. 09 natives, c-com	EN	ET	c-com	refl	pro	subj.(topic)	inhib	crit, RPD	—	
28	Patil et al. unpublished	EN	ET	gend	refl	retro	subj.	inhib	crit, FPRP / (reg.-cont. FFD)	(facil)	crit, regr.-cont. FFD

Experiments are sorted by effect pattern (effect in antecedent-match first, effect in antecedent-mismatch second). Marginal effects are in parentheses. The experiments are classified by language (English EN, Mandarin Chinese CN, Hindi HI), experimental method (self-paced reading SPR, eye-tracking while reading ET, visual world eye-tracking VW, event-related potentials ERP, cross-modal priming Primg), retrieval cues that are examined (gender, number, animacy, c-command), dependency type (reflexive, prepositional reflexive, possessive reflexive, reciprocal), interference type (proactive vs. retroactive), and by syntactic position of the distractor (subject, object, genitive attribute, sentence external memory load, and discourse topicality in parenthesis). For reading studies, the interest area (AOI) labeled crit refers to the region containing the reflexive/reciprocal. For eye-tracking experiments, Measure indicates the dependent variable in which the effect was observed (first-fixation duration FFD, first-pass reading time FPRT, total-fixation time TFT, regression-path duration RPD, right-bounded reading time RBRT, re-reading time RRT, first-pass regression probability FPRP, first-fixation duration in regression-contingent trials reg.-cont. FFD). Effects in parentheses are only marginally significant. The entry "—" means that the respective conditions were not tested in the experiment while "n.s." means that no significant effect was observed. IWM *refers to participants with low working memory capacity.*

employing additional assumptions: The cue-based retrieval account as implemented by Lewis and Vasishth (2005) and employed by Dillon (2011), Dillon et al. (2013), Kush and Phillips (2014), Parker and Phillips (2014) and Patil, Vasishth, and Lewis, "Retrieval interference in syntactic processing: The case of reflexive binding in English" (unpublished manuscript) is unable to explain facilitatory interference in antecedent-match conditions or inhibitory interference in antecedent-mismatch conditions.

The present article (i) provides further experimental evidence relating to the current debate about the use of non-structural retrieval cues and (ii) proposes two extensions to the standard cue-based retrieval architecture in order to account for the seemingly contradictory patterns of experimental results observed across studies.

We first present two eye-tracking experiments examining interference effects in the processing of the Mandarin Chinese reflexive *ziji*. There is a wide range of competing syntactic or pragmatic approaches of how to analyze *ziji* (for formal accounts see Yang, 1983; Manzini and Wexler, 1987; Pica, 1987; Kang, 1988; Tang, 1989; Huang and Tang, 1989, 1991; Cole et al., 1990, 1993; Cole and Sung, 1994; Cole and Wang, 1996; for pragmatic and non-uniform accounts see Huang et al., 1984; Yu, 1992, 1996; Xue et al., 1994; Pan, 1997; Pollard and Xue, 1998; Huang and Liu, 2001; Liu, 2010). We will restrict the following summary of the syntactic behavior of *ziji* to its properties that are relevant for the present experimental design. In contrast to English reflexives, *ziji* does not have any gender or number marking, but requires its antecedent to be animate (Tang, 1989).[4] Thus, animacy might be used as a non-structural cue to retrieve *ziji*'s antecedent. Similar to reflexives of many other languages including English, *ziji* needs to be c-commanded by its antecedent.[5] Moreover, the antecedent is required to be a subject (Huang, 1984). In contrast to English, the antecedent does not have to be contained in the local clause of the reflexive, but can also be contained in a superordinate clause (non-local binding). The processing of locally vs. non-locally bound *ziji* has been investigated by Gao et al. (2005), Liu (2009), Li and Zhou (2010), Dillon (2011), Chen et al. (2012), and Dillon et al. (2014).

The present experiments examine whether animate nouns that are in a structurally inaccessible position (i.e., not c-commanding the reflexive) induce interference effects on the processing of *ziji*. So far, the literature on interference effects in reflexives has focused on morphologically marked phi-features (gender, number). Thus, the examination of animacy in the processing of Mandarin *ziji* does not only add cross-linguistic evidence to the debate that, so far, has been centered on English, but also extends the range of investigated retrieval cues to a purely semantic feature.

Both experiments have relatively large sample sizes in order to increase statistical power. Given that the prediction of the structure-based account is that no effect should be seen (i.e., a null result), it is particularly important to conduct high power studies.

2. Experiment 1

In Experiment 1, we tested whether locally bound *ziji* is subject to interference from a structurally inaccessible distractor that fulfills the animacy requirement of *ziji*. In a 2 × 2 design we manipulated animacy of the structurally accessible antecedent (henceforth labeled as *antecedent-match* vs. *antecedent-mismatch*) and of a structurally inaccessible distractor noun that intervened between the accessible antecedent and the reflexive (henceforth labeled as *distractor-match* vs. *distractor-mismatch*). This design extends the study reported by Chen et al. (2012), who were the first to test interference effects in Mandarin *ziji*, in several respects. In contrast to Chen and colleagues, in the present experiment, *ziji* was in object position rather than being a possessive modifier and we included antecedent-mismatch conditions which Chen et al. did not test. Moreover, we used the more time-sensitive eye-tracking method rather than self-paced reading.

The ACT-R model as implemented by Lewis and Vasishth (2005) predicts an inhibitory interference effect in antecedent-match conditions and a facilitatory interference effect in antecedent-mismatch conditions at the reflexive. The structure-based account (Nicol and Swinney, 1989; Sturt, 2003; Phillips et al., 2011; Dillon, 2011; Dillon et al., 2013; Kush and Phillips, 2014), in contrast, predicts the absence of an interference effect in both antecedent-match and antecedent-mismatch conditions. Moreover, the Lewis and Vasishth ACT-R model predicts incorrect retrievals of the animate distractor (misretrievals) in both antecedent-match and antecedent-mismatch conditions, but the proportion of misretrievals is predicted to be higher in antecedent-mismatch conditions. The structure-based account predicts no misretrievals of the animate inaccessible distractor.

2.1. Materials and Method

2.1.1. Materials

We tested 48 experimental sentences which contained an either animate (antecedent-match) or inanimate (antecedent-mismatch) accessible antecedent in subject position (*yundongyuan* "athlete" vs. *pihuating* "kayak" in 2) and the reflexive as direct object. Due to the animacy requirement of *ziji*, the conditions with an inanimate accessible antecedent were ungrammatical. Between the main clause subject and the main clause verb, an adverbial clause intervened that contained an either animate (distractor-match) or inanimate (distractor-mismatch) inaccessible distractor (*lingdui* "team leader" vs. *meiti* "media" in 2). This distractor was also a subject, but did not c-command the reflexive and was therefore not a legal antecedent. The reflexive was followed by a frequency phrase or a durational phrase consisting of four characters, which was analyzed as a spillover region.

[4]There are some exceptions under which the animacy constraint can be violated, see (Tang, 1989; Pan, 1995) for a discussion. Crucially for our experimental design, in the syntactic literature, there is no example of non-emphatic, mono-morphemic *ziji* in argument position bound by a clearly inanimate NP.

[5]The c-command constraint might be violated in case of animate sub-commanding antecedents (Tang, 1989; Xue et al., 1994; Pollard and Xue, 1998), psychological verbs (Huang and Tang, 1991), passives and *ba*-constructions (Yu, 1992, but cf. Cole and Wang 1996), and in case of cataphoric binding by the subject of a matrix clause that is preceded by an adjunct clause containing *ziji* (Huang and Liu, 2001). Moreover, *ziji* can refer to the speaker of the utterance (Li, 1991), the addressee, or even a third person salient in the discourse (Pan, 2000).

(2) **Animate/Inanimate antecedent; Animate/Inanimate distractor**

运动员$_i$/*皮划艇$_i$			在	领队$_j$/媒体$_j$	施加	巨大	压力	的	情况	下
Yundongyuan$_i$/*Pihuating$_i$	[PP		zai	**lingdui$_j$/meiti$_j$**	shijia	juda	yali	de	qingkuang	xia]
athlete/kayak			*when*	*team.leader/media*	*excert*	*great*	*pressure*	MOD	*circumstance*	*under*
超越了	自己$_{i/*j}$	一共	三	次…						
chaoyue-le	**ziji$_{i/*j}$**	yigong	san	ci…						
outperform-ASP	*self*	*in total*	*three*	*times…*						

When the team leader/media excerted great pressure, the athlete/kayak outperformed himself/itself three times in total…

The experimental items were complemented with 72 filler sentences (48 grammatical, 24 ungrammatical) with varying syntactic structures including sentences containing the bare reflexive *ziji* as well as the bi-morphemic reflexive *ta-ziji* and pronouns in different syntactic positions.

Each sentence was followed by a multiple choice comprehension question that probed for the correct retrieval of the antecedent. Participants could choose between the antecedent, the distractor, an unrelated noun taken from a previous trial and the option "I am not sure." This design allowed us to examine not only whether the antecedent was retrieved correctly, but also to assess the proportion of misretrievals of the distractor. To ensure that participants also fully parsed the intervening adverbial clause containing the distractor, a second multiple-choice question targeted the adverbial clause. The same options were provided as in the first question. The questions following the filler sentences targeted various syntactic positions in the sentence.

Pretest. Since the exact binding properties of *ziji* are still subject to discussion in the syntactic literature, we conducted a paper-based questionnaire study to test our assumption that the main clause subject in the experimental items binds the reflexive. Forty native speakers of Mandarin recruited at Beijing Normal University participated in this study against payment of 25 RMB (approximately 3 EUR). None of them would participate in either of the eye-tracking experiments. Participants were presented with the antecedent-match conditions of the experimental items together with 90 filler sentences containing *ziji* in various syntactic positions and were instructed to circle the word in the given sentence *ziji* referred to or to explicitly write down the referent in case of an unbound interpretation of *ziji*.

Results. In 97.2% of all trials, participants selected the main clause subject as antecedent for the reflexive (97.0% and 97.3% when the distractor was animate or inanimate, respectively). This shows that in the experimental materials, Mandarin speakers indeed choose the main clause subject as antecedent for the reflexive.

2.1.2. Participants and Procedure

The experiment was conducted in the eye-tracking lab of the State Key Laboratory of Cognitive Neuroscience and Learning at Beijing Normal University. One hundred fifty students from different universities located in Beijing participated in the experiment against payment of 40 RMB (approximately 5 EUR). All participants were native speakers of Mandarin and had normal or corrected to normal vision.

Eye movements (right eye monocular) were recorded using an SR Research Eyelink 1000 eyetracker at a sampling rate of 1000 Hz. Participants' head was stabilized using a forehead- and chin-rest. The screen-to-eye-distance was 82 cm, the camera-to-eye-distance 75 cm. Stimuli were presented in Simplified Chinese characters (font type SimSun, black font, font size 25) on a 22 inch monitor with light gray background using SR Research Experiment Builder software. Re-calibrations were performed between trials if necessary. Each experimental session began with 6 practice trials in which feedback to the comprehension questions was provided. In the experimental trials, no feedback was given. Short breaks were given according to the participants' individual needs. The sentences were presented according to a standard Latin Square. Items were pseudo-randomized such that at least one filler sentence intervened between two experimental sentences. Each sentence was followed by two multiple choice comprehension questions as described above.

2.2. Results

All statistical analyses were carried out in R using linear mixed effects models provided by the lme4 package version 1.0-6 (Bates et al., 2014). Binary dependent variables were analyzed using a logistic link function. For both, the analysis of response accuracies and eye movements, two sets of contrasts were applied. We first ran a model testing for a main effect of antecedent (animate antecedents coded as $+0.5$; inanimate antecedents coded as -0.5), a main effect of interference (animate distractors coded as $+0.5$; inanimate distractors coded as -0.5) and the interaction between the two main effects. Second, we applied nested contrasts testing for an interference effect within antecedent-match and antecedent-mismatch conditions separately. All models were fit with a full variance-covariance matrix for participants and items (Gelman and Hill, 2007); in case the model failed to converge or the variance-covariance matrix was degenerate, random slopes for items or participants were removed.

2.2.1. Comprehension Questions

Comprehension questions targeting the reflexive-antecedent dependency were analyzed. We analyzed response accuracies and the proportion of incorrect selection of the inaccessible distractor. An overview of participants' answers is provided in **Table 2**. In the statistical analysis of response accuracies, only the main effect of antecedent reached marginal significance (estimate $= 0.34$, $SE = 0.18$, $z = 1.84$, $p = 0.07$). The antecedent (i.e., the correct option) was chosen more often in antecedent-match conditions. This effect was expected since

TABLE 2 | Experiment 1: Chosen answer to the comprehension question by condition in percentages.

Antecedent	*Distractor*	Chosen answer			
		Antecedent	Distractor	Unrelated	"Not sure"
match	*match*	82.3	5.1	0.9	11.7
	mismatch	81.6	3.6	2.4	12.4
mismatch	*match*	75.9	4.8	1.1	18.2
	mismatch	75.7	4.9	0.8	18.5

in the antecedent-mismatch conditions, no fully grammatically correct answer to the comprehension question was available (the antecedent was coded as "correct" answer, but the option "not sure" was provided as one response option in order to account for the ungrammaticality of the sentence). The analysis of the proportions of incorrect selection of the distractor revealed a main effect of antecedent: participants chose the distractor more often in antecedent-mismatch conditions than in antecedent-match conditions (estimate $= -0.45$, $SE = 0.18$, $z = -2.48$, $p < 0.05$). However, the size of this main effect was very small. We will therefore not base any conclusions on this effect. Moreover, the interaction between antecedent and distractor was significant (estimate $= 0.56$, $SE = 0.15$, $z = 3.61$, $p < 0.001$). Pairwise comparisons revealed that, within antecedent-match conditions, the distractor was chosen more often erroneously as answer to the comprehension question in case the distractor was animate (estimate $= 0.83$, $SE = 0.31$, $z = 2.70$, $p < 0.01$). But, as can be seen from **Table 2**, the animate distractor did not cause a decrease in selection probability of the antecedent but rather attracted selections from the unrelated noun. In antecedent-mismatch conditions, no interference effect was observed.

2.2.2. Eye Movements

Eye movements were analyzed at the reflexive, the pre-critical region (verb) and the spillover material consisting of the frequency/durational phrase (post-critical). In order to provide a comprehensive picture of our data, and to make our results comparable to other studies we report the whole range of eye-tracking measures common in psycholinguistic research, although some of these measures are correlated by definition. As first-pass measures, we report first-fixation duration (FFD), i.e., the duration of the first fixation in first-pass reading, and first-pass reading time (FPRT, also called gaze duration), i.e., the sum of all first-pass fixations on a word before leaving it. As regression-related measures, we report regression-path duration (RPD, also called go-past time), i.e., the sum of all fixation durations starting from the first first-pass fixation on a word including regressive fixations to previous material until a region to the right of this word is fixated, right-bounded reading time (RBRT), i.e., the sum of all fixations on a word before another region to the right of this region is fixated, and first-pass regression probability (FPRP), i.e., the proportion first-pass regressions initiated from a word. As a later-pass measure, we analyzed re-reading time (RRT), i.e., the sum of all fixations on a word that are not contained in FPRT. In addition, we analyzed total-fixation time (TFT), which is defined as the sum of FPRT and RRT. In order to achieve close to normally distributed model residuals, we log-transformed reading times (Box and Cox, 1964) and excluded all trials in which the respective continuous dependent variable was zero. First-fixation probability of the pre-critical region, the reflexive and the spillover region was 90, 62, and 87%, respectively. Re-readings occurred in 60, 33, and 45% of the trials at pre-critical region, the reflexive and the spillover region, respectively. In all models, centered log-frequencies of the antecedent and the distractor taken from the SUBLETEX-CH database (Cai and Brysbaert, 2010) were included as covariates because items had not been matched for frequencies of the antecedents and distractors. Mean raw reading times with standard errors for the pre-critical, critical and post-critical regions are provided in **Table 3**. The results of the statistical analyses of participants' eye movements are summarized in **Tables 4**, **5**.

The main effect of antecedent (longer reading times or a higher proportion of regressions in antecedent-mismatch conditions) was significant across regression-related measures (RPD, RBRT, FPRP) and late measures (TFT, RRT). In RPD and RBRT, the effect of antecedent started already at the pre-critical region and remained significant at the reflexive and the post-critical region. In FPRP, the effect was significant at the reflexive only. In TFT, the effect also started at the pre-critical region and continued to be significant at the reflexive. In RRT, the effect reached significance only at the pre-critical region.

The main effect of interference (longer reading times or higher proportion of regressions in distractor-match conditions) reached significance across first-pass, regression-related and late measures. In RPD and FPRP, the effect reached significance at the reflexive itself, in FPRT and RBRT at the post-critical region and in TFT at the pre-critical region.

The interaction between antecedent and interference reached significance at the reflexive across first-pass and regression-related measures (FFD, FPRT, RBRT). In RBRT, this interaction was already present at the pre-critical region. Pairwise comparisons revealed that the interference effect was driven by the antecedent-mismatch conditions: Within antecedent-mismatch conditions, an inhibitory interference effect was observed across first-pass, regression-related and late measures (FFD, FPRT, RBRT, RPD, TFT).[6] In FFD, FPRT, RBRT,

[6]In RPD, the effect predicted by the linear-mixed model is also an *inhibitory* one, although the opposite pattern is present in the raw means (cf. **Table 3**). This discrepancy is driven by a few very long (i.e., > 6000 ms) regression-path durations in the antecedent-mismatch/distractor-mismatch condition of one particular item. Because of the concave nature of the log-function, the log-transformation of the data reduces the impact of these extremely high values. As all of these extreme values stem from the same experimental condition, the difference in means of the log-transformed RPDs even switches the sign in antecedent-mismatch conditions (log-transformed means in antecedent-mismatch conditions: distractor-match = 5.85 log-ms; distractor-mismatch = 5.80 log-ms). This explains why the linear-mixed model estimates an inhibitory rather than a facilitatory interference effect. Removing the item which caused the extreme values yields similar results as log-transforming the data, i.e., the sign of the interference effect also switches from negative to positive (raw means in antecedent-mismatch conditions with the item causing extremely long RPDs being removed: distractor-match = 476 ms; distractor-mismatch = 469 ms).

TABLE 3 | Experiment 1: Means and standard errors of raw first-fixation duration, first-pass reading time, right-bounded reading time, regression-path duration, total fixation time, re-reading time in ms, and first-pass regression probability in percentages at the pre-critical region, the reflexive and the post-critical region.

	Pre-critical				Reflexive				Post-critical			
Antecedent	***Match***		***Mismatch***		***Match***		***Mismatch***		***Match***		***Mismatch***	
Distractor	***Match***	***Mism.***	***Match***	***Mism.***	***Match***	***Mism.***	***Match***	***Mism.***	***Match***	***Mism.***	***Match***	***Mism.***
FFD	279 (3)	277 (3)	285 (3)	279 (3)	258 (3)	259 (3)	264 (4)	251 (3)	270 (4)	274 (4)	274 (3)	268 (4)
FPRT	366 (6)	370 (6)	386 (6)	375 (6)	269 (4)	270 (4)	282 (5)	263 (4)	376 (8)	370 (8)	384 (7)	364 (7)
RBRT	397 (6)	407 (7)	425 (7)	413 (7)	286 (4)	284 (4)	302 (5)	284 (4)	436 (9)	430 (9)	448 (9)	432 (9)
RPD	484 (13)	508 (14)	537 (15)	533 (15)	430 (16)	410 (15)	484 (18)	494 (25)	688 (25)	662 (23)	759 (27)	755 (30)
FPRP	13 (1)	14 (1)	14 (1)	16 (1)	17 (1)	15 (1)	19 (1)	17 (1)	24 (1)	24 (1)	26 (1)	26 (1)
TFT	725 (14)	696 (14)	761 (15)	716 (14)	439 (10)	428 (9)	455 (10)	433 (9)	628 (14)	614 (15)	628 (15)	605 (13)
RRT	577 (17)	537 (16)	604 (17)	565 (16)	418 (15)	396 (14)	411 (14)	397 (13)	507 (18)	503 (20)	509 (21)	493 (17)

In the calculation of standard errors of continuous dependent variables, between-participants variance has been removed using the Cousineau (2005) normalization with Morey (2008)'s correction. For continuous variables, trials with a 0 as value of the respective variable have been excluded.

TABLE 4 | Experiment 1: Main effect of antecedent, main effect of interference and their interaction at the pre-critical (*ziji* − 1), critical (*ziji*), and post-critical (*ziji* + 1) regions for the dependent variables (DVs) first-fixation duration, first-pass reading time, right-bounded reading time, regression-path duration, first-pass regression probability, total fixation time, and re-reading time.

DV	Comparison	Pre-critical			Reflexive			Post-critical		
		Coef	***SE***	***t or z***	***Coef***	***SE***	***t or z***	***Coef***	***SE***	***t or z***
FFD	Antecedent	−0.02	0.01	−1.72	0.00	0.01	−0.27	0.00	0.01	−0.10
	Interference	0.01	0.01	0.84	0.02	0.01	1.42	0.01	0.01	0.45
	Ant × Int	−0.01	0.01	−0.61	**−0.02**	**0.01**	**−2.06***	−0.02	0.01	−1.77
FPRT	Antecedent	−0.03	0.01	−1.93	−0.02	0.01	−1.02	−0.02	0.02	−0.96
	Interference	0.01	0.01	0.71	0.03	0.02	1.94	**0.04**	**0.01**	**2.69***
	Ant × Int	−0.02	0.01	−1.71	**−0.04**	**0.01**	**−2.92***	−0.02	0.01	−1.24
RBRT	Antecedent	**−0.04**	**0.01**	**−2.78***	**−0.04**	**0.02**	**−2.23***	**−0.04**	**0.02**	**−2.65***
	Interference	0.00	0.01	−0.20	0.02	0.02	1.59	**0.03**	**0.01**	**2.12***
	Ant × Int	**−0.03**	**0.01**	**−2.02***	**−0.03**	**0.01**	**−2.14***	−0.01	0.01	−0.56
RPD	Antecedent	**−0.06**	**0.02**	**−2.57***	**−0.09**	**0.02**	**−3.98***	**−0.09**	**0.03**	**−3.14***
	Interference	−0.02	0.02	−1.11	**0.04**	**0.02**	**2.25***	0.03	0.02	1.11
	Ant × Int	−0.02	0.02	−1.30	−0.01	0.02	−0.73	−0.01	0.02	−0.32
FPRP	Antecedent	−0.08	0.10	−0.79	**−0.18**	**0.08**	**−2.17***	−0.12	0.09	−1.41
	Interference	−0.14	0.10	−1.46	**0.16**	**0.07**	**2.14***	0.03	0.07	0.42
	Ant × Int	0.03	0.09	0.35	0.01	0.07	0.19	0.03	0.06	0.43
TFT	Antecedent	**−0.04**	**0.02**	**−2.20***	**−0.04**	**0.02**	**−2.29***	−0.01	0.02	−0.64
	Interference	**0.04**	**0.01**	**2.76***	0.02	0.02	1.48	0.03	0.02	1.93
	Ant × Int	−0.01	0.01	−0.83	−0.02	0.02	−1.12	0.01	0.02	0.57
RRT	Antecedent	**−0.06**	**0.03**	**−2.22***	−0.05	0.03	−1.47	−0.02	0.03	−0.68
	Interference	0.04	0.02	1.74	0.02	0.03	0.87	−0.01	0.03	−0.46
	Ant × Int	0.02	0.02	0.72	−0.01	0.03	−0.42	0.03	0.03	1.06

Statistically significant ($\alpha = 0.05$) effects are marked with an asterisk and highlighted in bold.

and RPD, the effect reached significance at the reflexive itself and, in FPRT, continued to be significant at the post-critical region. In TFT, the effect reached significance at the pre-critical region only. Within antecedent-match conditions, the interference effect did not reach significance in any measure or region.

TABLE 5 | Experiment 1: Pairwise comparisons of animacy of the distractor (interference) nested within animate/inanimate antecedent (antecedent match/mismatch) at the pre-critical (*ziji* − 1), critical (*ziji*), and post-critical (*ziji* + 1) regions for the dependent variables (DVs) first-fixation duration, first-pass reading time, right-bounded reading time, regression-path duration, first-pass regression probability, total fixation time, and re-reading time.

DV	Comparison	Pre-critical			Reflexive			Post-critical		
		Coef	*SE*	*t or z*	*Coef*	*SE*	*t or z*	*Coef*	*SE*	*t or z*
FFD	Interference [ant. match]	0.00	0.02	0.26	0.00	0.02	−0.23	−0.01	0.02	−0.93
	Interference [ant. mismatch]	0.02	0.02	1.09	**0.04**	**0.02**	**2.30***	0.02	0.02	1.55
FPRT	Interference [ant. match]	−0.01	0.02	−0.71	−0.01	0.02	−0.40	0.02	0.02	1.05
	Interference [ant. mismatch]	0.03	0.02	1.68	**0.07**	**0.02**	**3.16***	**0.06**	**0.02**	**2.79***
RBRT	Interference [ant. match]	−0.03	0.02	−1.46	−0.01	0.02	−0.52	0.02	0.02	1.13
	Interference [ant. mismatch]	0.02	0.02	1.24	**0.05**	**0.02**	**2.07***	0.04	0.02	1.89
RPD	Interference [ant. match]	−0.04	0.03	−1.59	0.03	0.03	1.11	0.02	0.03	0.61
	Interference [ant. mismatch]	0.00	0.02	0.03	**0.06**	**0.03**	**2.07***	0.03	0.03	1.04
FPRP	Interference [ant. match]	−0.11	0.13	−0.84	0.17	0.10	1.64	0.03	0.09	0.33
	Interference [ant. mismatch]	−0.17	0.13	−1.36	0.14	0.10	1.40	0.00	0.09	0.00
TFT	Interference [ant. match]	0.03	0.02	1.40	0.01	0.02	0.26	0.04	0.02	1.72
	Interference [ant. mismatch]	**0.05**	**0.02**	**2.52***	0.04	0.02	1.80	0.02	0.02	1.04
RRT	Interference [ant. match]	0.06	0.03	1.84	0.01	0.04	0.33	0.02	0.04	0.41
	Interference [ant. mismatch]	0.03	0.03	0.77	0.04	0.04	0.90	−0.04	0.04	−1.06

Statistically significant ($\alpha = 0.05$) effects are marked with an asterisk and highlighted in bold.

Moreover, the models revealed that the higher frequency of the antecedent led to a significant slowdown at the reflexive in regression-based measures (RPD: estimate = 0.03, $SE = 0.01$, $t = 2.12$; RBRT: estimate = 0.02, $SE = 0.01$, $t = 2.00$) and RRT (estimate = 0.05, $SE = 0.02$, $t = 2.76$). Frequency of the distractor, in contrast, did not affect reading times at the reflexive in any measure.

One potential issue with the data analysis reported here is the so-called multiple-testing problem, that is, testing more than one dependent variable but keeping the significance threshold α unchanged at 0.05. Although in the field of psycholinguistics it is uncommon to apply an α-level correction when multiple eye-tracking measures are analyzed, we applied a Bonferroni correction to the α-level (Bonferroni, 1936; Dunn, 1959, 1961) and checked whether the effects reported above remained significant under this more conservative analysis. This is important in order to reduce the Type I error probability because, as has been noted for example by Ioannidis (2005), false positives are a serious issue in empirical science and in psychological science in particular (Simmons et al., 2011). With respect to reading studies, von der Malsburg and Angele, "The elephant in the room: False positive rates in standard analyses of eye movements in reading" (unpublished manuscript) recently showed by means of Monte Carlo simulations that testing multiple eye-tracking measures leads to a more dramatic increase of Type I errors as compared to what had been generally believed in the field. Von der Malsburg and Angele therefore recommend to apply a Bonferroni correction to the α-level. Given that we have analyzed seven dependent variables, the Bonferroni correction yields a corrected α-level of 0.007, which corresponds to an approximate t-value of $\pm$ 2.69.[7] With this adjusted α-level, the main effect of antecedent remained significant in RBRT at the pre-critical region and in RPD at the reflexive and at the post-critical region. The main effect of interference reached significance in FPRT at the post-critical region and in TFT at the pre-critical region. The interaction between antecedent and interference was significant in FPRT at the reflexive. In pairwise comparisons, the interference effect in antecedent-mismatch conditions in FPRT at the reflexive and at the post-critical region remained significant. The antecedent-frequency effect reached the Bonferroni-corrected significance threshold in RRT, but not in RPD and RBRT. In sum, although the Bonferroni correction and the considerable loss in statistical power that goes along with it makes some effects lose statistical significance, the overall pattern of results remains unchanged: An early interference effect at the reflexive present only within antecedent-mismatch conditions, an effect of antecedent in regression-related dependent variables starting already at the verb preceding the reflexive and an effect of antecedent-frequency at the reflexive.

2.3. Discussion

Comprehension questions required participants to correctly identify the reflexive's antecedent and to select it from four response options. Although participants could choose the option "not sure," they were highly likely to choose the antecedent even

[7] This t-value was approximated by using a normal distribution.

if it was inanimate and hence a semantically illicit antecedent. This shows that in their final interpretation of the reflexive they gave structural information a higher priority than semantic information. In antecedent-match conditions only, the distractor was chosen more often in case it was animate. But, crucially, this higher proportion of distractor choices was at the cost of choices of the unrelated noun, not of the antecedent. From this pattern we conclude that the observed effect reflects *offline* interference, i.e., an effect driven by meta-linguistic considerations at the moment of answering the comprehension question. If, in contrast, the effect reflected retrieval interference during the actual sentence reading, i.e., online effects, it would be expected to manifest itself in a higher proportion of misretrievals of the distractor leading to a lower proportion of choosing the *antecedent*, not the unrelated noun, because the latter is only introduced in the question.

The analyses of eye movements firstly showed that the presence of an animate distractor led to a processing slowdown (i.e., inhibitory interference) in antecedent-mismatch conditions. This slowdown was observed across first-pass, regression-related and late measures. In the more conservative analysis with Bonferroni-corrected significance threshold, this slowdown remained reliable in FPRT. In antecedent-match conditions, this interference effect did not reach significance. This pattern cannot be explained by either of the two accounts under discussion: The parser's sensitivity to the presence of an animate distractor cannot be accounted for by a structure-based retrieval mechanism. ACT-R cannot explain the results either since, in its current implementation, ACT-R predicts facilitatory rather than inhibitory interference in antecedent-mismatch conditions caused by a higher proportion of misretrievals of an animate distractor. Kush and Phillips (2014) also found inhibitory interference in antecedent-mismatch conditions in a self-paced reading experiment on Hindi reciprocals. They explain this effect in terms of interference that occurs during a later repair process of the ungrammatical sentence rather than at the moment of retrieval. Crucially, in Kush and Phillips (2014)'s experiment, the interference effect reached marginal significance only two words after the reciprocal. For the present experiment, their explanation seems implausible since the interfere effect reaches significance already in first-pass measures at the reflexive.

Second, we did *not* find any interference effects in the antecedent-match conditions. Although these results are statistically inconclusive, it is worth mentioning that this is consistent with the findings of Chen et al. (2012), who found interference effects in non-locally bound *ziji* but failed to find effects in locally-bound *ziji*.

Third, we observed a slowdown due to an inanimate antecedent in regression-related and late measures. This grammaticality effect is in line with both structure-based retrieval and the ACT-R model. In contrast to the interference effect, this effect is most pronounced at the pre-critical region. We will discuss possible explanations for this early appearance of the effect in the Discussion of Experiment 2.

Fourth, we found that lower frequency of the antecedent led to faster reading times at the reflexive. This effect might be explained by a low-frequency encoding advantage. It has been shown that the lower frequency of a word leads to a better memory encoding which results in a faster retrieval at a later point in time (Diana and Reder, 2006). Thus, low frequency antecedents might be better encoded in memory leading to a facilitated retrieval when reaching the reflexive, which shows the more prominent role of the antecedent in the retrieval process. Indeed, this facilitation due to infrequent antecedents replicates findings from English pronouns. In an eye-tracking-while-reading experiment, Van Gompel and Majid (2004) found faster FFD and FPRT at the region following the reflexive as a function of lower frequency of the antecedent.

One potential concern with the present results might be that task-related influences on interference cannot be ruled out. One of the two comprehension questions following the experimental sentences targeted the reflexive-antecedent dependency, which—in particular in the ungrammatical conditions—might have caused readers to spend some additional reading time to rule out the animate distractor. This would explain the observed inhibitory interference in the target-mismatch conditions. In the design of the experiment, we had addressed this potential issue by including ungrammatical fillers containing *ziji* with questions that did not target the reflexive-antecedent dependency. Moreover, participants had the option to answer "not sure," which allowed them not to assign any meaning to an ungrammatical sentence. If task-specifics had been an influential factor, they would most probably be reflected in repair attempts that are triggered by unexpectedly retrieving an inanimate antecedent. However, the interference effect reached significance already in FFD and FPRT. Based on a large-scale review of eye movements in reading, Clifton et al. (2007) have suggested that early measures like FFD or FPRT are unlikely to reflect repair processes since across studies, repair or reanalysis effects are typically observed in regression-related or later-pass reading measures. To the extent that Clifton et al. (2007)'s claim is correct, we can conclude that repair processes caused by the task-demands are unlikely to explain the observed results.

3. Experiment 2

This experiment extended Experiment 1 in several aspects. First, it examined proactive rather than retroactive interference; second it examined the influence of distractor items that are not a syntactic part of the sentence itself but presented as memory load; third, we tested the influence of syntactic locality on the retrieval and its interaction with interference. Previous studies report a processing slowdown in case *ziji* is non-locally bound compared to locally bound *ziji* (Gao et al., 2005; Li and Zhou, 2010; Dillon, 2011; Chen et al., 2012; Dillon et al., 2014). In the present experiment, we aimed at replicating this locality effect and investigating whether interference effects are modulated by locality of the reflexive binding.

In a dual-task paradigm, similar to Van Dyke and McElree (2006), participants were asked to remember three animate or three inanimate distractor nouns while reading a sentence

containing an either locally or non-locally bound reflexive. This resulted in a 2 × 2 design, with locality (local vs. non-local) and the distractors' animacy (animate vs. inanimate) as factors. Conditions with animate distractors are labeled as *distractors-match* and conditions with inanimate distractors as *distractors-mismatch*.

The structure-based account predicts no effect of animacy of the distractor nouns held in memory. In contrast, the standard ACT-R cue-based retrieval model predicts an inhibitory interference effect due to animacy of the distractors: retrieval times at the reflexive are predicted to be longer in distractors-match conditions. Moreover, ACT-R predicts a main effect of locality with non-local conditions being read slower. This prediction does not follow from the cue-based nature of the retrieval mechanism but rather from the ACT-R assumption of decay: The more recent, i.e., the local, antecedent has a higher level of activation than the non-local antecedent when reaching the reflexive. This difference in activation is predicted to be reflected in both, retrieval times and comprehension accuracies. Since this predicted locality effect is unrelated to the set of cues used for retrieval, the structure-based cue-based retrieval account (i.e., the ACT-R model with only structural features used as retrieval cues) makes the same prediction. Moreover, a structure-based serial search mechanism that first checks the local subject position and subsequently the non-local subject as proposed by Dillon (2011) and Dillon et al. (2014) for the processing of Mandarin *ziji* also predicts a processing slowdown in non-local conditions.

(3) a. **Local binding**

这些	数据$_i$	表明	这个	少年$_j$	耽误了	自己$_{*i/j}$	整整	三	年…
Zhe-xie	**shuju**$_i$	biaoming	[zhe-ge	**shaonian**$_j$	danwu-le	**ziji**$_{*i/j}$	zhengzheng	san	nian]…
this-CL	*data*	*demonstrate*	*this*-CL	*youngster*	*hinder*-ASP	*self*	wholly	three	years…

These data demonstrate that this youngster hindered himself three whole years…

b. **Non-local binding**

这个	少年$_i$	表明	这些	数据$_j$	耽误了	自己$_{i/*j}$	整整	三	年…
Zhe-ge	**shaonian**$_i$	biaoming	[zhe-xie	**shuju**$_j$	danwu-le	**ziji**$_{i/*j}$	zhengzheng	san	nian]…
this-CL	*youngster*	*demonstrate*	*this*-CL	*data*	*hinder*-ASP	*self*	wholly	three	years…

This youngster demonstrates that these data hindered him three whole years…

3.1. Materials and Method

3.1.1. Materials

We tested 36 experimental sentences[8] which consisted of a super-ordinate clause and an embedded clause containing the reflexive *ziji* as direct object. The locality factor of the antecedent-reflexive dependency was achieved by manipulating animacy of the local subject (i.e., the subject of the embedded clause) and the non-local subject (i.e., the subject of the superordinate clause): in the local conditions, the local subject was animate and the non-local subject was inanimate (see 3a) while in the non-local conditions, the local subject was inanimate and the non-local subject was animate (see 3b). Since *ziji* requires its antecedent to be animate, this design ensured that in the local conditions, *ziji* was bound by the local subject whereas in the non-local conditions it was bound by the subject of the superordinate clause. Similar to Experiment 1, the reflexive was followed by a spillover region consisting of four characters that formed a frequency phrase or a durational phrase. Each sentence was followed by a yes/no-comprehension question that probed for the correct binding of the reflexive. Seventy-two filler sentences containing reflexives and pronouns in varying syntactic positions were presented with memory load words of varying part-of-speech.

Pretest. In order to verify that speakers of Mandarin indeed bind the reflexive to the local subject/the superordinate subject in the local/non-local condition, respectively, we presented 40 native speakers of Mandarin recruited at Beijing Normal University with the experimental sentences in form of a paper-based questionnaire against payment of 25 RMB (approximately 3 EUR). Ninety filler sentences containing *ziji* in various syntactic positions were included. Participants were instructed to circle the word in the sentence *ziji* referred to, or, in case they found that no antecedent was available in the sentence, to write down which entity *ziji* referred to.

Results. Overall, 90.4% of all trials were answered as we had expected: In the local conditions, the animate local subject was chosen as antecedent and in the non-local conditions the animate matrix subject was selected. In the local conditions, accuracy was lower (85.1%) than in the non-local conditions (95.6%). A syntactic classification of the incorrect answers is provided in the Appendix.

3.1.2. Participants and Procedure

This experiment was conducted in the same laboratory as Experiment 1. One hundred thirty native speakers of Mandarin with normal or corrected-to-normal vision participated in the experiment against payment of 60 RMB (approximately 7 EUR). The general experimental set-up was the same as in Experiment 1. The experiment was split into two experimental sessions (40–70

[8]Originally, we had 48 items, but 12 of these were excluded based on low acceptability judgments of native speakers.

TABLE 6 | Experiment 2: Comprehension question response accuracy in percentage by experimental condition.

Locality	*Distractors*	Accuracy
local	*match*	67.1
	mismatch	68.7
non-local	*match*	71.8
	mismatch	71.9

minutes per session) conducted on two subsequent days. At the beginning of each trial, the three distractors were shown on the screen one below another for 3 seconds. When the words disappeared, the test sentence was displayed. After having finished reading the sentence, the comprehension question was presented. After having answered the comprehension question, participants were asked to serially recall the distractors: The three distractors together with three unrelated items (similarly animate or inanimate nouns) were displayed simultaneously on the screen as a numbered list in randomized order. Participants were asked to choose the distractors in their correct order from this list.

3.2. Results

For all dependent variables, we fit two sets of contrasts; the first tested for main effects of locality (local conditions coded as -0.5; non-local conditions coded as $+0.5$) and interference (animate distractors coded as $+0.5$; inanimate distractors coded as -0.5) and their interaction; in the second model pairwise comparisons of memory load nested within each level of locality were applied. In addition, experimental session (first vs. second session) was coded with sum-contrasts and its interaction with the other effects were included as predictors. All models were fit with random intercepts for items and participants, no random slopes were fit since they led to convergence failure in most of the models.

3.2.1. Comprehension Questions

Mean accuracy scores by experimental condition are shown in **Table 6**. None of the comparisons reached statistical significance.[9]

3.2.2. Memory Recall

Mean serial and non-serial recall accuracies for each of the three distractors and total serial and non-serial recall accuracy (i.e., all distractors recalled correctly) are presented in **Table 7**. In the statistical analyses of total serial recall accuracy none of the comparisons reached significance. In the analyses of total non-serial accuracies, the interaction between animacy of the distractors and locality was significant (estimate = -0.22, $SE = 0.10$, $z = -2.21$, $p < 0.05$). Pairwise comparisons revealed that this interaction was driven by a significant effect of distractors (lower recall accuracy of animate distractors) that was present only in local conditions (estimate $= -0.30$, $SE = 0.14$, $z = -2.25$, $p < 0.05$).

3.2.3. Eye Movements

The same log-transformed dependent variables as in Experiment 1 were analyzed at the reflexive, the verb preceding it (pre-critical), and the spillover material (post-critical). As in the analysis of Experiment 1, trials were excluded when the continuous variable on which the analysis was carried out was zero. First-pass fixations occurred at the pre-critical region, the reflexive, and the spillover region in 86, 50, and 85% of the trials, respectively. Re-readings were recorded in 55, 25, and 36% of the trials at pre-critical region, the reflexive, and the spillover region, respectively. Mean reading times with standard errors for each dependent variable are provided in **Table 8**.

The output of the linear-mixed models is summarized in **Tables 9** and **10**. The effect of experimental session was significant across regions and measures: Participants read faster in their second experimental session.[10] The main effect of locality reached significance across regression-based and later-pass measures (RBRT, RPD, FPRP, RRT, TFT) at the pre-critical region only. The main effect of interference was significant only in RRT at the post-critical region (longer RRTs when distractors were animate, i.e., inhibitory interference). The interaction between locality and interference was significant across first-pass, regression-based, and later-pass measures (FFD, FPRT, RBRT, RPD, TFT) at the reflexive. The pairwise comparisons revealed that the interaction was driven by a slowdown for animate distractors at the reflexive that was present only in local conditions. This inhibitory interference reached significance across first-pass, regression-based, and later-pass measures (FPRT, RBRT, RPD, TFT). For non-local conditions, a similar slowdown was observed only in RRT at the post-critical region.

As we did for Experiment 1, we checked which of the observed effects remained significant with a Bonferroni-corrected significance threshold. Given seven dependent variables, the corrected α-level is 0.007, which corresponds to an approximate t-value of $\pm$ 2.69.[11] The significance of the main effect of locality was not affected by this correction in any dependent variable, it remained significant at the pre-critical region in RBRT, RPD, FPRP, TFT, and RRT. The main effect

[9]In response accuracies the proportion of correctly answered yes-questions was strikingly higher than the proportion of correctly answered no-questions. We can exclude the possibility that this pattern can be explained by a general tendency of the participants to answer "yes" since no such difference was observed in filler sentences. We also excluded the hypothesis that this pattern might be related to the difficult nature of the dual-task paradigm by running a follow-up eye-tracking experiment ($N = 14$) with the same experimental set-up but without memory load that yielded a similar response pattern. As the pre-test on the materials had shown that native speakers indeed do the correct binding of the reflexive, we hypothesized that the response pattern was intrinsically related to the nature of the comprehension questions rather than to the experimental sentences themselves. We therefore ran another experiment ($N = 52$) in which the experimental and filler sentences appeared on the computer screen together with the respective comprehension question. Again, we observed a similar response pattern as in the online experiments. We thus conclude that the observed tendency to answer "yes" on the experimental comprehension questions reflects an offline effect, i.e., an effect which occurs at the moment when participants meta-linguistically think about how to answer the question, rather than an effect of online reflexive binding.

[10]The effect of experimental session is not of theoretical interest to our research question, therefore it is not presented in the results tables and will not be discussed further.

[11]This t-value was approximated by using a normal distribution.

TABLE 7 | Experiment 2: Mean serial and non-serial recall accuracy in percentage of the three memory load words separately and total accuracy in percentage presented by experimental condition.

	Serial accuracy				Non-serial Accuracy			
Locality	***Local***		***Non-local***		***Local***		***Non-local***	
Distractors	***Match***	***Mism.***	***Match***	***Mism.***	***Match***	***Mism.***	***Match***	***Mism.***
1st word correct	85	83	85	85	92	94	94	94
2nd word correct	79	75	81	79	93	94	94	93
3rd word correct	82	78	83	82	90	91	92	91
Total correct	68	67	71	69	77	81	82	80

TABLE 8 | Experiment 2: Means and standard errors of raw first-fixation duration, first-pass reading time, right-bounded reading time, regression-path duration, total fixation time, re-reading time in ms, and first-pass regression probability in percentages at the pre-critical region, the reflexive and the post-critical region.

	Pre-critical				Reflexive				Post-critical			
Locality	***Local***		***Non-local***		***Local***		***Non-local***		***Local***		***Non-local***	
Distractors	***Match***	***Mism.***	***Match***	***Mism.***	***Match***	***Mism.***	***Match***	***Mism.***	***Match***	***Mism.***	***Match***	***Mism.***
FFD	267 (5)	268 (5)	267 (5)	270 (5)	251 (6)	239 (5)	240 (5)	244 (6)	257 (6)	255 (5)	253 (5)	258 (6)
FPRT	342 (9)	341 (9)	351 (9)	343 (9)	260 (7)	245 (6)	250 (6)	254 (7)	325 (10)	320 (10)	322 (9)	320 (9)
RBRT	398 (11)	409 (11)	433 (12)	447 (13)	278 (7)	259 (6)	263 (7)	277 (8)	378 (11)	375 (12)	383 (12)	375 (11)
RPD	575 (25)	573 (26)	596 (24)	638 (28)	486 (28)	419 (27)	448 (29)	484 (32)	636 (35)	628 (36)	667 (41)	710 (47)
FPRP	23 (2)	20 (2)	25 (2)	28 (2)	18 (1)	14 (1)	16 (1)	16 (1)	24 (2)	24 (2)	25 (2)	25 (2)
TFT	683 (25)	666 (22)	737 (26)	763 (28)	396 (14)	354 (13)	377 (14)	379 (14)	501 (19)	491 (18)	508 (19)	479 (17)
RRT	626 (33)	592 (28)	645 (33)	704 (34)	352 (22)	365 (26)	360 (24)	345 (22)	432 (29)	414 (30)	468 (32)	413 (26)

In the calculation of standard errors of continuous dependent variables, between-participants variance has been removed using the Cousineau (2005) normalization with Morey (2008)'s correction. For continuous variables, trials with a 0 as value of the respective variable have been excluded.

of interference at the post-critical region in RRT did not reach the adjusted significance threshold. The interaction between locality and interference remained significant at the reflexive in RBRT and TFT, but did not reach significance anymore in FFD, FPRT, and RPD. In pairwise comparisons, the interference effect in local conditions at the reflexive remained significant in RBRT and TFT, but did not reach the significance threshold anymore in FPRT and RPD. The interference effect in non-local conditions that was observed at the post-critical region did not reach the adjusted significance threshold. In sum, the main effect of locality as well as the interference effect in locally bound *ziji* remained significant in various dependent variables even with an adjusted α-level. The interference effect in non-local conditions, in contrast, was not reliable under the corrected α-level.

3.3. Discussion

In the comprehension questions, no evidence for an interference effect was found. In the memory recall task, in contrast, we found that, in local conditions only, animate words were more difficult to recall than inanimate words.

First, we found evidence for a processing slowdown associated with the non-local binding of the reflexive. This locality effect replicates findings from SAT (Dillon, 2011; Dillon et al., 2014), ERP (Li and Zhou, 2010; Dillon, 2011), cross-modal priming (Liu, 2009), and self-paced reading (Chen et al., 2012), and is accounted for by the ACT-R model, no matter whether the set of retrieval cues is unconstrained or limited to structural cues. The structure-based serial search as proposed by Dillon (2011) and Dillon et al. (2014) is also in line with the observed locality effect. However, it is not fully clear why this locality effect appears at the verb preceding the reflexive rather than at the reflexive itself. One explanation would be a preview effect. Alternatively, it might be the case that the observed effect does not reflect locality of the reflexive binding but rather the verb's preference for an animate subject since the locality manipulation is achieved by having the local subject either animate or inanimate. Along the same lines, one could explain why in Experiment 1, the effect of animacy of the antecedent becomes significant already at the verb preceding the reflexive. A strong indication that the observed effect at the verb indeed reflects the verb's preference for an animate subject comes from a re-analysis of the self-paced reading data reported by Chen et al. (2012), where the locality manipulation was also achieved by varying the animacy of the local and non-local subjects, and the main clause verb also directly preceded the reflexive *ziji*. Chen et al. (2012) analyzed only the region containing the reflexive and the regions *following* the reflexive, but not the verb *preceding* the reflexive. Re-analyzing their data at the verb region revealed that the locality effect in their

TABLE 9 | Experiment 2: Main effects of locality and interference and their interaction at the pre-critical (*ziji*−1), critical (*ziji*), and post-critical (*ziji*+1) regions for the dependent variables (DVs) first-fixation duration, first-pass reading time, right-bounded reading time, regression-path duration, first-pass regression probability, total fixation time, and re-reading time.

DV	Comparison	Pre-critical			Reflexive			Post-critical		
		Coef	*SE*	*t or z*	*Coef*	*SE*	*t or z*	*Coef*	*SE*	*t or z*
FFD	Locality	0.00	0.01	−0.06	−0.01	0.01	−0.92	0.00	0.01	−0.19
	Interference	−0.01	0.01	−0.98	0.01	0.02	0.35	−0.01	0.02	−0.58
	Locality × Interference	0.00	0.02	0.00	**0.04**	**0.02**	**2.20***	0.02	0.02	1.21
FPRT	Locality	0.01	0.02	0.52	−0.01	0.02	−0.83	0.00	0.02	0.24
	Interference	0.00	0.02	0.26	0.01	0.02	0.63	0.00	0.02	−0.23
	Locality × Interference	0.00	0.02	−0.23	**0.04**	**0.02**	**2.33***	0.02	0.02	0.85
RBRT	Locality	**0.08**	**0.02**	**5.37***	−0.01	0.02	−0.34	0.01	0.02	0.36
	Interference	−0.01	0.02	−0.49	0.01	0.02	0.61	0.00	0.02	0.01
	Locality × Interference	0.00	0.02	0.03	**0.06**	**0.02**	**3.21***	0.02	0.02	0.79
RPD	Locality	**0.11**	**0.02**	**5.15***	0.03	0.03	1.10	0.02	0.03	0.78
	Interference	0.00	0.03	−0.08	0.04	0.03	1.36	−0.01	0.03	−0.30
	Locality × Interference	0.02	0.03	0.71	**0.07**	**0.03**	**2.16***	0.03	0.03	0.95
FPRP	Locality	**0.46**	**0.08**	**5.80***	0.11	0.09	1.19	0.05	0.08	0.62
	Interference	0.04	0.10	0.43	0.15	0.11	1.34	0.00	0.09	0.02
	Locality × Interference	0.14	0.10	1.41	0.09	0.11	0.83	−0.01	0.09	−0.12
TFT	Locality	**0.10**	**0.02**	**5.55***	−0.01	0.02	−0.31	0.00	0.02	−0.15
	Interference	0.01	0.02	0.57	0.05	0.02	1.92	0.02	0.02	1.01
	Locality × Interference	0.01	0.02	0.34	**0.08**	**0.02**	**3.10***	0.00	0.02	−0.16
RRT	Locality	**0.10**	**0.03**	**3.71***	−0.01	0.04	−0.23	0.06	0.04	1.73
	Interference	−0.02	0.03	−0.55	0.00	0.05	0.04	**0.09**	**0.04**	**2.10***
	Locality × Interference	0.07	0.03	1.92	−0.03	0.05	−0.64	−0.04	0.04	−0.89

Statistically significant ($\alpha = 0.05$) effects are marked with an asterisk and highlighted in bold.

data was already significant at the verb ($t = 2.5$). As preview effects are ruled out as an explanation in self-paced reading, and given the high structural similarity of our experimental materials to the ones used by Chen et al. (2012), we conclude that the effect observed at the verb in Experiment 2 is most likely due to an animacy preference of the verb. Given this—admittedly unforeseen—confounding animacy preference of the verb, we cannot draw any conclusions about the actual locality manipulation. A potential locality effect might have been masked by the stronger effect of animacy preference: when reaching the verb in the non-local conditions, readers are highly likely to re-read the previous material to overcome the difficulty associated with the verb's inanimate subject, as indicated by the highly significant effects in FPRP, RPD, and RBRT. This leads to activation of the preceding materials in the non-local conditions *directly before* reaching the reflexive, which, in turn, might have canceled out a locality effect at the reflexive. Therefore, we conclude that our data is inconclusive with respect to the locality manipulation.

Second, we found clear evidence for inhibitory interference, but the time-course of this effect was different for local and non-local conditions. In local conditions, animate distractors led to a slowdown across first-pass, regression-based, and late eye-tracking measures at the reflexive itself. Even with a Bonferroni corrected significance threshold of $\alpha = 0.007$, this effect remained significant in RBRT and TFT. In FPRT and RPD, the inhibitory interference effect did not survive Bonferroni correction. However, since these measures numerically pattern with other measures—especially with RBRT, which is closely related—it could reflect a real effect. In non-local conditions, the interference effect appeared only later in processing (in RRT at the post-critical region). However, with Bonferroni adjusted significance threshold, this effect was not reliable. In sum, the observed interference pattern extends the findings of Experiment 1 in two respects. First, Experiment 2 shows that locally bound *ziji* is subject to early interference even in case a fully cue-matching antecedent is available. The difference to Experiment 1, where the interference effect did not reach significance in antecedent-match conditions, might be explained by the different experimental paradigms: rehearsal of the distractors during reading might cause stronger interference than the sentence-internal manipulation of Experiment 1.

TABLE 10 | Experiment 2: Interference effect nested within each level of locality (local vs. non-local) at the pre-critical (*ziji*−1), critical (*ziji*), and post-critical (*ziji*+1) regions for the dependent variables (DVs) first-fixation duration, first-pass reading time, right-bounded reading time, regression-path duration, first-pass regression probability, total fixation time, and re-reading time.

DV	Comparison	Pre-critical			Reflexive			Post-critical		
		Coef	*SE*	*t or z*	*Coef*	*SE*	*t or z*	*Coef*	*SE*	*t or z*
FFD	Interference [local]	−0.01	0.02	−0.69	0.04	0.03	1.79	0.01	0.02	0.45
	Interference [non−local]	−0.01	0.02	−0.69	−0.03	0.02	−1.31	−0.03	0.02	−1.26
FPRT	Interference [local]	0.00	0.03	0.02	**0.06**	**0.03**	**2.09***	0.01	0.03	0.44
	Interference [non−local]	0.01	0.03	0.35	−0.03	0.03	−1.21	−0.02	0.03	−0.76
RBRT	Interference [local]	−0.01	0.03	−0.33	**0.07**	**0.03**	**2.70***	0.02	0.03	0.57
	Interference [non−local]	−0.01	0.03	−0.37	−0.05	0.03	−1.83	−0.02	0.03	−0.55
RPD	Interference [local]	0.02	0.04	0.44	**0.11**	**0.05**	**2.49***	0.02	0.04	0.46
	Interference [non−local]	−0.02	0.04	−0.56	−0.03	0.05	−0.57	−0.04	0.04	−0.89
FPRP	Interference [local]	0.18	0.14	1.23	0.24	0.15	1.52	−0.01	0.13	−0.07
	Interference [non−local]	−0.09	0.13	−0.74	0.06	0.15	0.36	0.01	0.13	0.10
TFT	Interference [local]	0.02	0.03	0.64	**0.12**	**0.03**	**3.56***	0.02	0.03	0.60
	Interference [non−local]	0.00	0.03	0.16	−0.03	0.03	−0.83	0.03	0.03	0.82
RRT	Interference [local]	0.05	0.05	0.94	−0.03	0.07	−0.42	0.05	0.06	0.87
	Interference [non−local]	−0.08	0.05	−1.79	0.03	0.06	0.49	**0.13**	**0.06**	**2.09***

Statistically significant ($\alpha = 0.05$) effects are marked with an asterisk and highlighted in bold.

Second, the interference profile in non-locally bound *ziji* differs from the one in locally bound *ziji* in the sense that in non-local conditions no early effect was found, but there is weak evidence for a late effect. Although the late effect in non-local conditions was not significant under Bonferroni correction, there is reason to believe in this effect when viewed against the background of previous findings by Chen et al. (2012), who found an inhibitory interference effect in non-local *ziji*.

The observed interference effects are not compatible with a structure-based retrieval mechanism since no effect of the distractors is predicted. The ACT-R model, in contrast, can account for the inhibitory interference effect. However, ACT-R is unable to explain the delayed appearance of the effect in non-local conditions.

A possible explanation for the different interference patterns in local vs. non-local conditions could be that qualitatively different mechanisms are involved in the processing of locally and non-locally bound *ziji*. In the syntactic literature, it has been proposed that only the locally bound *ziji* should be regarded as a reflexive pronoun whereas non-locally bound *ziji* should be regarded as a logophoric pronoun which is subject to pragmatic and discourse constraints rather than to purely syntactic binding principles (Huang and Liu, 2001; Huang, 2002). One prominent argument favoring this idea of two lexically different instances of *ziji* are blocking effects observed in long-distance *ziji* but not in local *ziji* (Huang, 1984, 2002; Tang, 1989; Huang and Tang, 1991; Xue et al., 1994; Pan, 2000). A qualitative distinction between locally bound *ziji* and non-local *ziji* has also been proposed in the psycholinguistic literature. Based on previous work by Gao et al. (2005), Liu (2009) conducted a cross-modal priming experiment using sentences in which both a local and a non-local animate antecedent were present (i.e., globally ambiguous sentences in terms of binding) and manipulated stimulus-onset asynchrony (0 ms, 160 ms, 370 ms). When the probe was presented directly after the offset of the reflexive (SOA = 0 ms), a semantic priming effect for probes related to the local antecedent but not for probes related to the non-local antecedent was observed. At an SOA of 160 ms, in contrast, the pattern was reversed: There was a priming effect for probes that were semantically related to the non-local antecedent, but no priming effect for probes related to the local antecedent. At an SOA of 370 ms, both the local and non-local antecedent elicited a semantic priming effect. Liu (2009) interpreted these results as evidence for *ziji* being bound by the local subject in a first stage of processing and by the non-local subject in a second stage of processing, whereas in the final stage, both bindings are possible. Along the same lines, Dillon (2011) and Dillon et al. (2014) suggested that the parser tries to first access the local subject and only at a later stage accesses non-local antecedent positions. Such a temporal delay for the triggering of the retrieval of a non-local antecedent would indeed predict the pattern observed in Experiment 2: In the local conditions, the retrieval is triggered immediately at the moment when the reflexive is first encountered. The interference effects associated with this retrieval therefore appear already in early measures at the reflexive. In non-local conditions, in contrast, the retrieval of the non-local antecedent is triggered only after a certain delay,

which causes the interference effects to occur only in RRT at the spillover region.

4. An Extended Cue-Based Retrieval Model

As has been pointed out in the experimental discussions, the interference effects observed in the experiments presented here are not compatible with structure-based accounts. The current implementation of the standard cue-based retrieval model in ACT-R (Lewis and Vasishth, 2005) cannot explain the observed patterns either. In particular, standard cue-based retrieval is unable to explain (i) why there is an effect in antecedent-match conditions in Experiment 2 but not in Experiment 1, and (ii) why there is inhibitory interference observed in antecedent-match conditions in Experiment 1. We propose an explanation of the observed patterns by adding two independently motivated assumptions to standard cue-based retrieval: that (i) similarity-based interference is modulated by *distractor prominence* and that (ii) *cue confusion* can lead to similarity-based interference between non-similar items. As discussed earlier, the difference in the interference profiles of local and non-local *ziji* might be due to a qualitative difference in processing mechanisms and was therefore not included in our modeling.

4.1. Principle 1: Prominence

In Experiment 1, we found an interference effect in antecedent-mismatch conditions but not in antecedent-match conditions. According to Wagers et al. (2009), this is an expected prediction of cue-based retrieval and, in the context of subject-verb number attraction phenomena, the authors named it "grammatical asymmetry." Their intuitively plausible explanation was that a perfectly matching antecedent (as is the case in antecedent-match conditions) must clearly outcompete a partially matching distractor, while more interference is caused when both antecedent and distractor are only partially matching candidates.

Simulations with the current ACT-R implementation (Lewis and Vasishth, 2005) revealed that the latter does not predict such asymmetry (for details, see Engelmann et al., 2015, and our forthcoming paper Engelmann, Jäger, and Vasishth, "Confusability of retrieval cues in dependency resolution: A computational model," manuscript in preparation)—at least not in a principled way: It is possible to adjust ACT-R's parameters to permanently reduce similarity-based interference. However, this would leave unexplained why in some cases effects in antecedent-match conditions do appear (see the General Discussion for details). Standardly, ACT-R predicts interference effects in match and mismatch conditions. We therefore extended the ACT-R model with a *prominence principle* that scales similarity-based interference in relation to the difference in activation between antecedent and distractor.

In standard ACT-R, a memory item i receives an amount of spreading activation S_{ji} for each retrieval cue j it matches. This activation is reduced relative to the number of distractors that match the same retrieval cue j (this number is called the fan_{ji}):

$$S_{ji} = S - \ln(fan_{ji}) \tag{1}$$

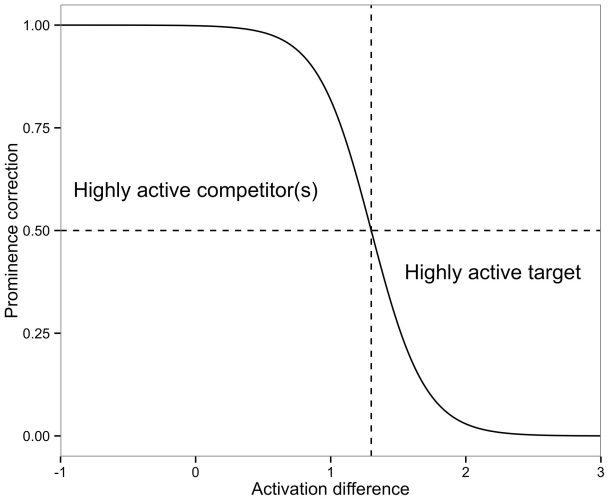


FIGURE 1 | Prominence correction by activation difference *Diff* (target − distractors) with $C = 5$ and $x_0 = 1.3$.

where S is the *maximum associative strength* parameter (*MAS*), which defaults to 1.

In our model, the fan_{ji} is transformed into fan'_{ji} by a *prominence correction*, that takes into account the distractors' relative activation:

$$fan'_{ji} = \begin{cases} \frac{1}{1+e^{-C(x_0-Diff)}} \times fan_{ji}, & \text{if } C > 0 \\ fan_{ji}, & \text{otherwise} \end{cases} \tag{2}$$

where *Diff* is the difference $A_i - \bar{A}_{Competitors}$ between the target activation A_i and the mean activation of all competitor items associated with cue j. The *prominence correction factor* C scales the steepness of the logistic *prominence correction* function and should not vary within the same model. In our simulations, we set it to 5. The function's *offset* x_0 is fixed at 1.3, which means that fan'_{ji} is $0.5 \times fan_{ji}$ at an activation difference between target and distractor of 1.3.

Figure 1 shows the change in the multiplicative term (the *prominence correction*), that determines the relation between *fan* and its transformation fan'. When the target has lower activation than the mean activation of its competitors, *Diff* is negative and the prominence correction approaches 1, which implies that the fan will correspond to the standard calculation in ACT-R, and the activation of the target will be reduced by some amount. This is the case when there are highly activated distractors present: similarity-based interference occurs in this case. *Diff* will be positive when the mean activation of the competitors is relatively low. In this case, the prominence correction will be a value less than 1, and as a consequence the second term in Equation (1) will approach 0, leading to a relatively larger amount of spreading activation to the target. In other words, there will be less interference.

This implementation of a prominence principle adds two predictions to the standard cue-based retrieval model: First, there

is generally less interference in antecedent-match conditions due to the presence of a highly activated fully matching antecedent. Second, similarity-based (inhibitory) interference in antecedent-match conditions is *increased* for distractors that are highly activated or when there are multiple distractors as in our Experiment 2.[12] Distractor base-level activation could be influenced by its grammatical role (subjects are more salient or accessible than objects, Chafe, 1976; Keenan and Comrie, 1977; Brennan, 1995; Grosz et al., 1995) and by its discourse topicality (Chafe, 1976; Givón, 1983; Du Bois, 1987, 2003; Ariel, 1990; Gundel et al., 1993; Grosz et al., 1995). Other factors contributing to the salience of the distractor and hence to its base-level activation might be first mention (Gernsbacher and Hargreaves, 1988), thematic role (Arnold, 2001), contrastive focus (Cowles et al., 2007) or animacy (Fukumura and van Gompel, 2011). In effect, the prominence principle accounts for both the absence of an effect in antecedent-match conditions of Experiment 1 and the presence of an inhibitory effect in Experiment 2. Furthermore, the prominence principle predicts greater interference effects in antecedent-match conditions for distractors in more salient positions. We will relate this prediction to the literature in the General Discussion.

4.2. Principle 2: Cue Confusion

As explained in the introduction and resulting from Equation (1), similarity-based (inhibitory) interference (or the fan effect) in ACT-R only arises when multiple memory items match the same retrieval cues. Since this is not the case in the antecedent-mismatch conditions of Experiment 1, the observed inhibitory interference is incompatible with ACT-R theory. At least this seems to be the case. We argue that this assumption of incompatibility might not be justified.

In the application of cue-based retrieval to sentence comprehension, it is generally assumed that retrieval cues perfectly distinguish matching features from non-matching ones. For instance, a *+plural* cue always activates plural items and not singular items. For our first experiment, this means that *+animate* is perfectly different from *+c-com* and no similarity-based interference is predicted in antecedent-mismatch conditions where the antecedent only matches *+c-com* and the distractor only matches *+animate*. However, the language processor might not differentiate between features categorically but rather on a continuous scale of similarity. In fact, in the general ACT-R framework, features are memory items just like the items they belong to and, therefore, could be confused with each other if they have a sufficient degree of similarity. If we assume that cue-feature associations have to be learned from language experience, it follows that these associations would somehow reflect co-occurrence statistics in the language input. Consequently, cues in a retrieval specification could, depending on the retrieval-relevant context, be associated with several features to different degrees.

A co-occurrence-based account would predict differences between English reflexives and Mandarin *ziji* in the following way: *Ziji* invariably requires its antecedent to match {*+ c-com*, *+animate*}, meaning that these two features frequently co-occur in the specific task of processing the Mandarin reflexive. English reflexives, on the other hand, have several alternative forms like *himself*, *herself*, *itself*, and *themselves*. All of these forms have the same structural requirement toward their antecedent but their non-structural retrieval cues vary in gender and number. The benefit of distinguishing features for number, gender, and structural relation in English reflexives results in a stronger one-to-one association between a cue and the corresponding feature. In the case of Mandarin *ziji*, however, there is no benefit from distinguishing *+ c-com* and *+animate* for the task of finding the appropriate antecedent. In consequence, retrieval cues might in this case be associated with both features to some degree in a kind of *crossed association*. In relation to the retrieval specification, antecedent and distractor would appear similar in this case, although they theoretically do not share any features. This confusion-induced similarity can cause similarity-based interference as of Equation (1), predicting inhibitory effects in conditions where they would not be expected in terms of standard cue-based retrieval assumptions.

We implemented cue confusion by further adjusting the measure of similarity-based interference (the *fan*) from Equation (1) to take into account all features and their strength of association with a certain cue:

$$fan_{ji} = 1 + \sum_k (1 + Q_{jk}) \tag{3}$$

where Q_{jk} is the *associative strength* between cue value j and feature value k on a scale of $[-1, 0]$, with -1 meaning no association and 0 representing maximum association. We assume that this association is dynamically adaptive to individual dependency environments. Equation (3) predicts that the stronger a cue-feature association the more this feature will contribute to similarity-based interference related to that cue. For example, if $Q_{c\text{-}com;anim}$ for *ziji* is -0.5, the resulting fan for the *+c-com* cue would be 1.5 instead of 1 as original ACT-R would predict. This increases similarity-based interference in comparison to English reflexives, where, say, $Q_{c\text{-}com;gend}$ would be standardly assumed -1, hence having a fan of 1 for each cue.

Another example of increased feature-co-occurrence are reciprocals like *each other*. In this case, the feature combination {*+ c-com*, *+plural*} is invariably required. Hence, our account predicts an increased cue-confusion level in the case of English reciprocals just like in Mandarin reflexives, possibly leading to inhibitory interference in antecedent-mismatch conditions.

With the cue confusion account, we propose that task requirements (frequent co-occurrence of certain features in similar retrieval contexts) dynamically influence how cues are treated during a retrieval request. Cue confusion therefore predicts that inhibitory interference effects in antecedent-mismatch conditions should preferably be observed

[12]Note that, for the case of multiple distractors, the original model, too, predicts increased interference. This, however, only explains the difference in effect size between Experiment 1 and 2, but neither the discrepancy between antecedent-match and antecedent-mismatch conditions in Experiment 1 nor the differences between other experiments that did not use multiple distractors.

in constructions where cues frequently co-occur. An evaluation of these predictions beyond our own experimental results will be provided in the General Discussion.

4.3. Simulation Results

We report model predictions for the full range of cue confusion values. ACT-R parameters were fixed to their defaults or to values used in previous simulations (Lewis and Vasishth, 2005): latency factor $LF = 1.5$, activation noise value $ANS = 1.5$, mismatch penalty $MP = 1.5$. We compare the model predictions with empirical FPRT on *ziji* of Experiments 1 and 2. We refer to FPRT in Experiment 2 although it was not significant under Bonferroni correction. It however patterned with an effect in RBRT, which had a similar magnitude. **Figure 2** plots the prediction space of a cue-based retrieval model that implements cue confusion and prominence (values represent the means of 2000 simulations each). For comparison, the predictions of a model without prominence are plotted in gray. The cue confusion level is plotted on a percentage scale, with 100% confusion meaning that both features, +*c-com* and +*animate*, are maximally associated with both the *c-com* and *animate* cues ($Q_{c\text{-}com;anim} = 0$ and $Q_{anim;c\text{-}com} = 0$). With *prominence correction factor* at 0 and *cue confusion level* at −1, the current model is equivalent to the original ACT-R model. The original model's predictions are therefore represented by the left-most points of the gray lines. The left panel shows the predictions for Experiment 1. With increasing cue confusion, the interference effect for the antecedent-mismatch conditions increases. At a confusion level of about 55% (indicated by the dotted vertical line), the model predicts an effect of the observed size in local conditions (19 ms in FPRT, indicated by the dashed horizontal line). In contrast to the original model, the prominence model predicts an interference effect close to zero for antecedent-match conditions in Experiment 1 for all cue confusion levels. This is in line with the absence of an effect in the data.

The right panel of **Figure 2** shows the predictions for a similar model as the left panel, but with three distractors instead of one, simulating the conditions of Experiment 2. The inhibitory effect for antecedent-match conditions increases with cue confusion in this scenario. An effect of about the observed size (15 ms in FPRT) is predicted at the same cue-confusion level as for Experiment 1.

To summarize, the extended model with cue confusion and prominence predicts the observed data of both experiments with fixed parameters at a cue-confusion level of about 55%. More specifically, the model predicts two patterns that the original ACT-R model does not predict: (i) the absence (or near absence) of an inhibitory interference effect in the antecedent-match conditions of Experiment 1 in spite of an effect present in Experiment 2 and (ii) an *inhibitory* interference effect in antecedent-mismatch conditions in Experiment 1.

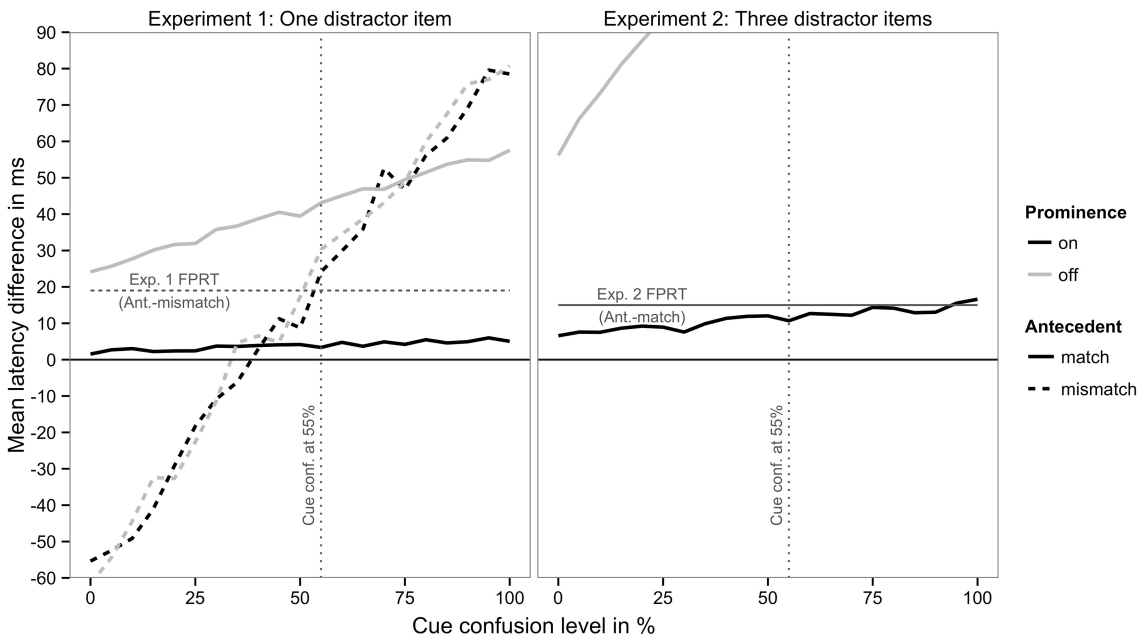


FIGURE 2 | Predicted interference effect (distractor-match − distractor-mismatch) by cue confusion level for the default model (gray lines) and the prominence model (black lines). The left panel shows the predicted interference for a single distractor (Exp. 1); the right panel for three distractors (Exp. 2). Solid lines represent the conditions where the antecedent matches the semantic cue, mismatch conditions are represented by dashed lines. The gray horizontal lines indicate the observed effect size in antecedent-mismatch conditions in Exp. 1 (left panel) and local antecedent-match conditions in Exp. 2 (right panel)—both in first pass reading time FPRT. The gray dotted vertical line intersects the x-axis at the estimated cue confusion value (55%) in both panels.

5. General Discussion

We conducted two eye-tracking experiments in which we investigated whether the reflexive *ziji* is subject to interference effects from structurally inaccessible distractor nouns that fulfill the animacy requirement of *ziji*. In Experiment 1, where only a single distractor was present in the sentence, we found inhibitory interference in antecedent-mismatch conditions but no effect in antecedent-match conditions. In Experiment 2, where three distractors were presented as memory load, we found interference effects also in antecedent-match configurations.

These results are clear evidence against a structure-based mechanism underlying memory retrieval in human sentence parsing. The interference effects observed in Experiments 1 and 2 are incompatible with a purely structure-based retrieval mechanism. However, Sturt (2003) and Kush and Phillips (2014) have proposed a potential explanation for interference effects within the structure-based account. These authors hypothesize that, in the case of retrieval failure, a later repair process might employ a retrieval with relaxed structural restrictions, giving rise to late interference effects. This late-interference account is a plausible explanation for the effect observed in the non-local conditions of Experiment 2, where the effect occurred only in RRT at the post-critical region. However, for the effects observed in locally bound *ziji* (Experiments 1 and 2), the late-interference account appears implausible given that the effects occur already in first-pass eye-tracking measures and at the critical region.[13] Also note that the effect reported in Kush and Phillips (2014) does not necessarily reflect late processes, since in self-paced reading experiments, it is very common that effects triggered at the critical region appear several words downstream.

The standard ACT-R model of cue-based retrieval (Lewis and Vasishth, 2005) does predict immediate interference effects but is not fully compatible with our results either. First, it predicts facilitatory rather than inhibitory interference in antecedent-mismatch conditions and, second, it cannot explain the absence of an effect in the antecedent-match conditions of Experiment 1. In fact, in the literature on reflexive processing, hardly any study can be found that reports the exact pattern predicted by the standard ACT-R model, namely inhibitory interference in antecedent-match conditions and facilitatory interference in antecedent-mismatch conditions.[14] An approach of extending the ACT-R model in favor of a structure-based mechanism has been taken by Parker and Phillips (2014). They have proposed that structural cues are weighted higher than semantic or morphological cues, so that interference effects occur only in case of an abnormally poor match of the accessible antecedent. This is a plausible explanation for their data and offers an account for the fact that interference is hard to find in reflexives. However, with respect to our results, it neither explains the inhibitory interference in antecedent-match conditions nor the difference in effect sizes in antecedent-match vs. antecedent-mismatch conditions.

In order to account for our results and the diverse patterns in the literature, we have introduced two concepts as an extension of the standard cue-based retrieval model. The *prominence principle* implements the idea that a perfectly matching or otherwise highly activated antecedent is only marginally affected by similarity-based interference from comparably poorly matching distractors. This explains the discrepancy between Experiments 1 and 2 (absence of an effect in antecedent-match conditions in Experiment 1 vs. an inhibitory interference effect in Experiment 2). With the concept of *cue confusion*, we proposed that the retrieval cues can be associated with several features of memory items and that the strength of these associations depends on experience with a specific linguistic context. For special cases, this can cause similarity-based interference between items that do not match the same retrieval cues. We argued that *ziji* is such a special case, which would explain the observed inhibitory interference in antecedent-mismatch conditions of Experiment 1.

In the following, we compare the predictions of the extended ACT-R model with the literature on reflexives. Prominence predicts that interference in antecedent-match conditions is generally low compared to antecedent-mismatch conditions but increases as a function of distractor activation. If we assume that distractor position (grammatical role and discourse topicality) affects its base-level activation in memory, the literature summary in **Table 1** seems to conform with these predictions: Among the studies which tested both antecedent-match and antecedent-mismatch conditions, about 75% report an interference effect (including marginal effects) in antecedent-mismatch conditions while only 50% of the studies found an effect in antecedent-match conditions. All studies that did report an effect in antecedent-match conditions had the distractor either in subject position (Badecker and Straub, 2002; Chen et al., 2012; Patil, Vasishth, and Lewis, "Retrieval interference in syntactic processing: The case of reflexive binding in English," unpublished manuscript), in topicalized subject position[15] (Felser et al., 2009; Cunnings and Felser, 2013; Clackson and Heyer, 2014), or had multiple distractors (Experiment 2 reported here). On the other hand, only half of the studies reporting no interference effect in antecedent-match conditions had the distractor in subject position. Obviously, not all studies that have the distractor in subject position report an effect, but the literature review suggests that subject position increases the probability of finding one. For the absence of an antecedent-match interference effect in our Experiment 1, there might be a specific reason: Dillon et al. (2015) have shown that items within restrictive relative clauses

[13]This is assuming that the pre-critical effects in Experiments 1 and 2 are due to difficulty with an inanimate subject, as discussed above, rather than reflecting an early application of binding during the parafoveal preview of the reflexive.

[14]It should be noted that the (marginal) facilitatory interference in antecedent-match conditions reported by three studies presented in **Table 1** (Sturt, 2003; Cunnings and Felser, 2013) is compatible with the ACT-R model although this may not be intuitively obvious. An exceptionally highly activated distractor (in all three of these experiments, the distractor is a discourse prominent subject) can lead to facilitatory interference (see Engelmann et al., 2015, and our forthcoming publication Engelmann, Jäger, and Vasishth, "Confusability of retrieval cues in dependency resolution: A computational model," manuscript in preparation).

[15]With distractors in "topicalized subject position" we here refer to distractor nouns in subject positions which appear as the current discourse topic in the test sentence because they were introduced in a preceding context sentence.

cause more interference as compared to items in appositive relative clauses. They attribute this difference to the idea that, in contrast to restrictive relative clauses, appositive relative clauses constitute a speech act separate from the one of the main utterance (Potts, 2005; Arnold, 2007). More generally, their results suggest that the embedding environment containing a distractor influences the strength of interference caused by this distractor. In terms of ACT-R, one might think of this as different base-level activations as a function of the type of embedding environment. It might be possible that the interposed adverbial structures which contain the distractor in our materials belong to those embedding environments which cause a relatively low degree of interference. This seems a plausible assumption since in our materials, the adverbial clause can simply be ignored by the parser without affecting the grammaticality or plausibility of the whole sentence.

For antecedent-mismatch conditions, cue confusion predicts stronger inhibition the higher the crossed association between cues and features is assumed to be, that is, in contexts with frequently co-occurring cue combinations. However, note that cue confusion is compatible with both facilitatory and inhibitory effects, and even with the absence of an effect, as all this is part of the effect continuum that is illustrated in **Figure 2**. This raises the concern of how to determine a sensible confusion level in each case, since a model allowing arbitrary predictions is not useful. Currently, the model prediction can only be treated as a predicted difference between two conditions in one or the other direction along the effect continuum. In other words, a prediction should be stated in terms of whether the antecedent-mismatch interference effect of one dependency tends more toward inhibition or toward facilitation in comparison to another dependency like, e.g., English reflexives. In the reasoning we apply here, we refer to English reflexives as a baseline with zero cue confusion and spot special cases where a different feature-co-occurrence rate can be assumed that would motivate a higher confusion level. We have argued that inhibitory interference was observed in antecedent-mismatch conditions in our Experiment 1 because *ziji* is a special case in the sense that the feature combination {+ *c-com*, +*animate*} is constant compared to the variable combinations in the different forms of English reflexives. The same logic with respect to {+ *c-com*, +*plural*} would apply to reciprocals. In the literature there is one study by Kush and Phillips (2014) that tested the Hindi equivalent of the reciprocal *each other* and indeed found the predicted inhibitory interference in antecedent-mismatch conditions.

Although the *post-hoc* nature of our proposals here is an important limitation that needs to be addressed with new empirical tests, theory development necessarily is data-driven, and the existing data suggest that our proposal constitutes one possible explanation. Indeed, currently it is the only computational account of the patterns of findings discussed here. In order to empirically test the predictions of cue confusion, it is necessary to experimentally manipulate feature-co-occurrence within a minimal pair. A potential experiment could use stimuli like in Example (4) to compare the interference effect in antecedent-mismatch conditions for *themselves* and *each other*. Cue confusion predicts a smaller facilitation or even an inhibition for *each other*. Furthermore, it should be possible to derive a numerical metric of cue confusion for a range of dependencies by computing co-occurrence frequencies in a treebank that contains dependency information as well as information about retrieval relevant features such as gender, number, and animacy.

(4) a. **Reflexive; distractor-match**
The *nurse* who cared for the *children* had pricked *themselves* . . .
b. **Reflexive; distractor-mismatch**
The *nurse* who cared for the *child* had pricked *themselves* . . .
c. **Reciprocal; distractor-match**
The *nurse* who cared for the *children* had pricked *each other* . . .
d. **Reciprocal; distractor-mismatch**
The *nurse* who cared for the *child* had pricked *each other* . . .

A more thorough test of the extended model's predictions will be presented in a forthcoming publication (Engelmann et al., "Confusability of retrieval cues in dependency resolution: A computational model," manuscript in preparation) that includes quantitative simulations of a range of previous studies on reflexive processing and subject-verb dependencies.

As a rather speculative point we want to add that the cue confusion level of a certain dependency might not only be influenced by feature-co-occurrence but also by task demands and individual differences. If cue-feature associations are subject to an adaptive learning process, they might also be affected by resource-preserving strategies. An example where strategic adaptation of comprehension processes has been found are relative clause attachment ambiguities. Swets et al. (2008) and Logačev and Vasishth (2015) have found that processing effort in ambiguity resolution was adapted to the type of comprehension questions. Also, effects of individual differences in working memory span have been found by Traxler (2007) and von der Malsburg and Vasishth (2012) for the processing of attachment ambiguities. If analogously to task- and resource-related underspecification in attachment ambiguities, cue-feature associations are affected by resource-preserving strategies in the sense of *good-enough processing* (Ferreira et al., 2002), we would expect that low-span readers tend to have greater cue confusion and, thus, exhibit interference effects further toward inhibition in the continuum than high-span readers. The marginal inhibitory effect for low-span readers in antecedent-mismatch conditions of Experiment 2 by Cunnings and Felser (2013) would fit with this expectation. However, more experimental data is needed in order to evaluate effects of individual differences and task-demands on cue-feature associations.

6. Conclusion

We have presented experimental evidence that is incompatible with structure-based accounts of reflexive processing and also inconsistent with the original cue-based ACT-R model of

sentence processing. In order to account for the observed pattern, we have proposed to add two new principles, prominence and cue confusion, to the ACT-R model. This extension to the ACT-R model is not only able to explain the pattern observed in the data presented in this article, but can also account for a range of previously unexplained patterns reported in the literature on reflexive processing. Naturally, this proposal needs to be evaluated with novel experimental data.

Acknowledgments

We thank Prof. Hua Shu and the State Key Laboratory of Cognitive Neuroscience and Learning at Beijing Normal University, China, for allowing us to conduct our experiments in their eye-tracking lab. We are grateful to Dr. Ming Yan, Dr. Jing'er Pan and Wei Zhou for their kind assistance and their very helpful comments on the experimental materials. We thank Prof. Reinhold Kliegl for his valuable advice on the statistical analyses and Dr. Zhong Chen and Maja Stegenwallner-Schütz for constructive discussions. We thank the audience at AMLaP 2012 and 2014 for comments. This work was partly funded by the Studienstiftung des deutschen Volkes and the Potsdam Graduate School by awarding a scholarship to the first and the second author, respectively. Publication of this article was funded by the Deutsche Forschungsgemeinschaft and the Open Access Publishing Fund of the University of Potsdam.

References

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., and Qin, Y. (2004). An integrated theory of the mind. *Psychol. Rev.* 111, 1036–1060. doi: 10.1037/0033-295X.111.4.1036

Anderson, J. R., and Lebiere, C. (1998). *Atomic Components of Thought*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Ariel, M. (1990). *Accessing Noun-Phrase Antecedents*. London: Routledge.

Arnold, D. (2007). Non-restrictive relatives are not orphans. *J. Linguist.* 43, 271–309. doi: 10.1017/S0022226707004586

Arnold, J. E. (2001). The effect of thematic roles on pronoun use and frequency of reference continuation. *Discourse Process.* 31, 137–162. doi: 10.1207/S15326950DP3102_02

Badecker, W., and Straub, K. (2002). The processing role of structural constraints on the interpretation of pronouns and anaphors. *J. Exp. Psychol.* 28, 748–769. doi: 10.1037/0278-7393.28.4.748

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2014). *lme4: Linear mixed-effects models using Eigen and S4*. R package version 1.0-6. Available online at: http://CRAN.R-project.org/package=lme4

Bonferroni, C. E. (1936). Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze* 8, 3–62.

Box, G. E. P., and Cox, D. R. (1964). An analysis of transformations. *J. R. Stat. Soc.* 26, 211–234.

Brennan, S. E. (1995). Centering attention in discourse. *Lang. Cogn. Process.* 10, 137–167.

Büring, D. (2005). *Binding Theory*. Cambridge, UK: Cambridge Textbooks in Linguistics; Cambridge University Press.

Cai, Q., and Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PLoS ONE* 5:e10729. doi: 10.1371/journal.pone.0010729

Chafe, W. L. (1976). "Givenness, contrastiveness, definiteness, subjects, topics, and point of view," in *Subject and Topic*, ed C. N. Li (New York, NY: Acedemic Press), 25–56.

Chen, Z., Jäger, L., and Vasishth, S. (2012). "How structure-sensitive is the parser? Evidence from Mandarin Chinese," in *Empirical Approaches to Linguistic Theory: Studies of Meaning and Structure (Studies in Generative Grammar)*, eds B. Stolterfoht and S. Featherston (Berlin: Mouton de Gruyter), 43–62.

Clackson, K., Felser, C., and Clahsen, H. (2011). Children's processing of reflexives and pronouns in English: evidence from eye-movements during listening. *J. Mem. Lang.* 65, 128–144. doi: 10.1016/j.jml.2011.04.007

Clackson, K., and Heyer, V. (2014). Reflexive anaphor resolution in spoken language comprehension: structural constraints and beyond. *Front. Psychol.* 5:904. doi: 10.3389/fpsyg.2014.00904

Clifton, C. J., Frazier, L., and Deevy, P. (1999). Feature manipulation in sentence comprehension. *Riv. Linguist.* 11, 11–39.

Clifton, C. J., Staub, A., and Rayner, K. (2007). "Eye movements in reading words and sentences," in *Eye Movements: A Window on Mind and Brain, Chapter 15*, eds R. P. Van Gompel, M. H. Fischer, W. S. Murray, and R. L. Hill (Oxford: Elsevier), 341–372.

Cole, P., Hermon, G., and Sung, L. (1990). Principles and parameters of long-distance reflexives. *Linguist. Inq.* 21, 1–22.

Cole, P., Hermon, G., and Sung, L.-M. (1993). Feature percolation. *J. East Asian Linguist.* 2, 91–118.

Cole, P., and Sung, L.-M. (1994). Head movement and long-distance reflexives. *Linguist. Inq.* 25, 355–406.

Cole, P., and Wang, C. (1996). Antecedents and blockers of long-distance reflexives: the case of Chinese *ziji*. *Linguist. Inq.* 27, 357–390.

Cousineau, D. (2005). Confidence intervals in within-subject designs: a simpler solution to Loftus and Masson's method. *Tutorial Quant. Methods Psychol.* 1, 42–45.

Cowles, H. W., Walenski, M., and Kluender, R. (2007). Linguistic and cognitive prominence in anaphor resolution: topic, contrastive focus and pronouns. *Topoi* 26, 3–18. doi: 10.1007/s11245-006-9004-6

Cunnings, I., and Felser, C. (2013). The role of working memory in the processing of reflexives. *Lang. Cogn. Process.* 28, 188–219. doi: 10.1080/01690965.2010.548391

Cunnings, I., and Sturt, P. (2014). Coargumenthood and the processing of reflexives. *J. Mem. Lang.* 75, 117–139. doi: 10.1016/j.jml.2014.05.006

Diana, R. A., and Reder, L. M. (2006). The low-frequency encoding disadvantage: word frequency affects processing demands. *J. Exp. Psychol.* 32, 1805–1815. doi: 10.1037/0278-7393.32.4.805

Dillon, B. (2011). *Structured Access in Sentence Comprehension*. Ph.D. thesis, University of Maryland, College Park, MD.

Dillon, B., Chow, W.-Y., Wagers, M., Guo, T., Liu, F., and Phillips, C. (2014). The structure-sensitivity of memory access: evidence from Mandarin Chinese. *Front. Psychol.* 5:1025. doi: 10.3389/fpsyg.2014.01025

Dillon, B., Clifton, C., Sloggett, S., and Frazier, L. (2015). "Not all relative clauses interfere equally in filler-gap processing," in *28th Annual CUNY Conference on Human Sentence Processing* (Los Angeles, CA: University of Southern California), 281.

Dillon, B., Mishler, A., Sloggett, S., and Phillips, C. (2013). Contrasting intrusion profiles for agreement and anaphora: experimental and modeling evidence. *J. Mem. Lang.* 69, 85–103. doi: 10.1016/j.jml.2013.04.003

Du Bois, J. W. (1987). The discourse basis of ergativity. *Language* 63, 805–855.

Du Bois, J. W. (2003). "Argument structure: grammar in use," in *Preferred Argument Structure: Grammar as Architecture for Function, Vol. 14 of Studies in Discourse and Grammar*, eds J. W. Du Bois, L. E. Kumpf, and W. J. Ashby (Amsterdam: John Benjamins), 11–60.

Dunn, J. O. (1959). Estimation of the medians for dependent variables. *Ann. Math. Stat.* 30, 192–197.

Dunn, J. O. (1961). Multiple comparisons among means. *J. Am. Stat. Assoc.* 56, 52–64.

Engelmann, F., Jäger, L., and Vasishth, S. (2015). "Cue confusion and distractor prominence can explain inconsistent interference effects," in *28th Annual CUNY Conference on Human Sentence Processing* (Los Angeles, CA: University of Southern California), 168.

Felser, C., Sato, M., and Bertenshaw, N. (2009). The on-line application of Binding Principle A in English as a second language. *Bilingualism* 12, 485–502. doi: 10.1017/S1366728909990228

Ferreira, F., Ferraro, V., and Bailey, K. G. D. (2002). Good-enough representations in language comprehension. *Curr. Dir. Psychol. Sci.* 11, 11–15. doi: 10.1111/1467-8721.00158

Fukumura, K., and van Gompel, R. P. G. (2011). The effect of animacy on the choice of referring expression. *Lang. Cogn. Process.* 26, 1472–1504. doi: 10.1080/01690965.2010.506444

Gao, L., Liu, Z., and Huang, Y. (2005). Ziji shi shei? Dui yueshu yuanze de shiyan yanjiu [Who is *ziji*? An experimental research on Binding Principles]. *Yuyan Kexue [Linguist. Sci.]* 4, 39–50.

Gelman, A., and Hill, J. (2007). *Data Analysis using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.

Gernsbacher, M. A., and Hargreaves, D. J. (1988). Accessing sentence participants: the advantage of first mention. *J. Mem. Lang.* 27, 699–717.

Givón, T. (1983). *Topic Continuity in Discourse*. Amsterdam: John Benjamins.

Grosz, B. J., Weinstein, S., and Joshi, A. K. (1995). Centering: a framework for modeling the local coherence of discourse. *Comput. Linguist.* 21, 203–225.

Gundel, J., Hedberg, N., and Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language* 69, 274–307.

Huang, C.-T. J. (2002). "Distributivity and reflexivity," in *On the Formal Way to Chinese Languages*, eds S.-W. Tang and C.-S. L. Liu (Stanford, CA: CSLI Publications), 21–44.

Huang, C.-T. J., Huang, Y.-H., Teng, T.-H., and Tiedeman, R. (1984). "Reflexives in Chinese and the teaching of Chinese," in *Proceedings of the First World Conference on Chinese Language* (Taipeih: World Chinese Language Association), 205–215.

Huang, C.-T. J., and Liu, C.-S. L. (2001). "Logophoricity, attitudes, and *ziji* at the interface," in *Long-Distance Reflexives, Vol. 33 of Syntax and Semantics*, eds P. Cole, G. Hermon, and C.-T. J. Huang (New York, NY: Academic Press), 141–195.

Huang, C.-T. J., and Tang, C.-C. J. (1989). "The local nature of the long-distance reflexive in Chinese," in *Proceedings of NELS 19* (Amherst, MA: GLSA Publications), 191–206.

Huang, C.-T. J., and Tang, C.-C. J. (1991). "The local nature of the long-distance reflexive in Chinese," in *Long-Distance Anaphora*, eds J. Koster and E. Reuland (Cambridge: Cambridge University Press), 263–282.

Huang, Y.-H. (1984). Reflexives in Chinese. *Stud. Eng. Lit. Linguist.* 10, 163–188.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Med.* 2:e124. doi: 10.1371/journal.pmed.0020124

Kang, B.-M. (1988). Unbounded reflexives. *Linguist. Philos.* 11, 415–456.

Keenan, E. L., and Comrie, B. (1977). Noun phrase accessibility and universal grammar. *Linguist. Inq.* 8, 63–99.

King, J., Andrews, C., and Wagers, M. (2012). "Do reflexives always find a grammatical antecedent for themselves?," in *25th Annual CUNY Conference on Human Sentence Processing* (New York, NY: The CUNY Graduate Center), 67.

Kush, D., and Phillips, C. (2014). Local anaphor licensing in an SOV language: implications for retrieval strategies. *Front. Psychol.* 5:1252. doi: 10.3389/fpsyg.2014.01252

Lewis, R. L., and Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cogn. Sci.* 29, 375–419. doi: 10.1207/s15516709cog0000_25

Li, N. (1991). *Perspective-Taking in Mandarin Discourse*. Ph.D. thesis, State University of New York at Buffalo, Buffalo, NY.

Li, X., and Zhou, X. (2010). Who is *ziji*? ERP responses to the Chinese reflexive pronoun during sentence comprehension. *Brain Res.* 1331, 96–104. doi: 10.1016/j.brainres.2010.03.050

Liu, L. (2010). A pragmatic account of anaphora: the cases of the bare reflexive in Chinese. *J. Lang. Teach. Res.* 1, 796–806. doi: 10.4304/jltr.1.6.796-806

Liu, Z. (2009). The cognitive process of Chinese reflexive processing. *J. Chin. Linguist.* 37, 1–27.

Logačev, P., and Vasishth, S. (2015). A multiple-channel model of task-dependent ambiguity resolution in sentence comprehension. *Cogn. Sci*. doi: 10.1111/cogs.12228. [Epub ahead of print].

Manzini, M. R., and Wexler, K. (1987). Parameters, binding theory, and learnability. *Linguist. Inq.* 18, 413–444.

Martin, A. E., and McElree, B. (2008). A content-addressable pointer mechanism underlies comprehension of verb-phrase ellipsis. *J. Mem. Lang.* 58, 879–906. doi: 10.1016/j.jml.2007.06.010

McElree, B. (2000). Sentence comprehension is mediated by content-addressable memory structures. *J. Psycholinguist. Res.* 29, 111–123. doi: 10.1023/A:1005184709695

McElree, B. (2003). Accessing recent events. *Psychol. Learn. Motiv.* 46, 155–200. doi: 10.1016/S0079-7421(06)46005-9

McElree, B., and Dosher, B. (1989). Serial position and set size in short-term memory: the time course of recognition. *J. Exp. Psychol.* 118, 346–373.

McElree, B., Foraker, S., and Dyer, L. (2003). Memory structures that subserve sentence comprehension. *J. Mem. Lang.* 48, 67–91. doi: 10.1016/S0749-596X(02)00515-6

Morey, R. D. (2008). Confidence intervals from normalized data: a correction to Cousineau (2005). *Tutorial Quant. Methods Psychol.* 4, 61–64.

Nicol, J., and Swinney, D. (1989). The role of structure in coreference assignment during sentence comprehension. *J. Psycholinguist. Res.* 18, 5–19.

Pan, H. (1995). *Locality, Self-Ascription, Discourse Prominence, and Mandarin Reflexives*. Ph.D. thesis, University of Texas at Austin, Austin, TX.

Pan, H. (1997). *Constraints on Reflexivization in Mandarin Chinese*. New York, NY: Garland Publishing.

Pan, H. (2000). "Why the blocking effect?," in *Long-Distance Reflexives, Vol. 33 of Syntax and Semantics*, eds P. Cole, G. Hermon, and C.-T. J. Huang (New York, NY: Academic Press), 279–316.

Parker, D., and Phillips, C. (2014). "Selective priority for structure in memory retrieval," in *27th Annual CUNY Conference on Human Sentence Processing* (Columbus, OH: Ohio State University), 100.

Phillips, C., Wagers, M. W., and Lau, E. F. (2011). "Grammatical illusions and selective fallibility in real-time language comprehension," in *Experiments at the Interfaces, Vol. 37 of Syntax and Semantics*, ed J. Runner (Bingley: Emerald Group Publishing Limited), 147–180.

Pica, P. (1987). "On the nature of the reflexivation circle," in *Proceedings of the 17th Annual Meeting of the North East Linguistic Society*, eds J. McDonough and B. Plunkett (Amherst, MA: GLSA Publications), 483–499.

Pollard, C., and Xue, P. (1998). Chinese reflexive *ziji*: syntactic reflexives vs. nonsyntactic reflexives. *J. East Asian Linguist.* 7, 287–318.

Potts, C. (2005). *The Logic of Conventional Implicatures*. Oxford: Oxford University Press.

Reuland, E. (2011). *Anaphora and Language Design*. Cambridge, MA: MIT Press.

Simmons, J. P., Nelson, L. D., and Simonsohn, U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* 22, 1359–1366. doi: 10.1177/0956797611417632

Sternberg, S. (1966). High-speed scanning in human memory. *Science* 153, 652–654.

Sternberg, S. (1969). Memory-scanning: mental processes revealed by reaction-time experiments. *Am. Sci.* 57, 421–457.

Sturt, P. (2003). The time-course of the application of binding constraints in reference resolution. *J. Mem. Lang.* 48, 542–562. doi: 10.1016/S0749-596X(02)00536-3

Swets, B., Desmet, T., Clifton, C., and Ferreira, F. (2008). Underspecification of syntactic ambiguities: evidence from self-paced reading. *Mem. Cogn.* 36, 201–216. doi: 10.3758/MC.36.1.201

Tang, C.-C. J. (1989). Chinese reflexives. *Nat. Lang. Linguist. Theory* 7, 93–121.

Traxler, M. J. (2007). Working memory contributions to relative clause attachment processing: a hierarchical linear modeling analysis. *Mem. Cogn.* 35, 1107–1121. doi: 10.3758/BF03193482

Van Dyke, J. (2007). Interference effects from grammatically unavailable constituents during sentence processing. *J. Exp. Psychol.* 33, 407–430. doi: 10.1037/0278-7393.33.2.407

Van Dyke, J., and McElree, B. (2006). Retrieval interference in sentence comprehension. *J. Mem. Lang.* 55, 157–166. doi: 10.1016/j.jml.2006.03.007

Van Dyke, J., and McElree, B. (2011). Cue-dependent interference in comprehension. *J. Mem. Lang.* 65, 247–263. doi: 10.1016/j.jml.2011.05.002

Van Gompel, R. P., and Majid, A. (2004). Antecedent frequency effects during the processing of pronouns. *Cognition* 90, 255–264. doi: 10.1016/S0010-0277(03)00161-6

von der Malsburg, T., and Vasishth, S. (2012). Scanpaths reveal syntactic underspecification and reanalysis strategies. *Lang. Cogn. Process.* 28, 1–34. doi: 10.1080/01690965.2012.728232

Wagers, M. W., Lau, E. F., and Phillips, C. (2009). Agreement attraction in comprehension: representations and processes. *J. Mem. Lang.* 61, 206–237. doi: 10.1016/j.jml.2009.04.002

Xiang, M., Dillon, B., and Phillips, C. (2009). Illusory licensing effects across dependency types: ERP evidence. *Brain Lang.* 108, 40–55. doi: 10.1016/j.bandl.2008.10.002

Xue, P., Pollard, C., and Sag, I. (1994). "A new perspective on Chinese *ziji*," in *Proceedings of the 13th Annual Meeting of the West Coast Conference on Formal Linguistics* (Stanford, CA: CSLI Publications), 432–447.

Yang, D.-W. (1983). The extended binding theory of anaphora. *Lang. Res.* 19, 169–192.

Yu, W. X. (1992). Challenging Chinese reflexive data. *Linguist. Rev.* 9, 285–294.

Yu, X.-F. W. (1996). *A Study of Chinese Reflexives*. Ph.D. thesis, University of London, London, UK.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2015 Jäger, Engelmann and Vasishth. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Appendix

TABLE A1 | Pretest of Experiment 2: Classification of the participants' answers in the "incorrect" trials by experimental condition.

Chosen antecedent of ziji	Condition		
	Local	**Non-local**	
Local inanimate subj.	*n.a.*	15 (1.56%)	*Sentence-internal antecedent*
Non-local inanimate subj.	40 (4.17%)	*n.a.*	
Cataphoric binding	40 (4.17%)	14 (1.46%)	
Speaker	11 (1.15%)	4 (0.42%)	*Sentence-external antecedent*
Context	27 (2.81%)	7 (0.73%)	
Author of media noun	19 (1.98%)	0	
Recipient of media noun	2 (0.21%)	0	

Percentages refer to the total number of trials including the correct trials.

frontiers in PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 05 November 2014
doi: 10.3389/fpsyg.2014.01252

Local anaphor licensing in an SOV language: implications for retrieval strategies

***Dave Kush*[1]* *and Colin Phillips*[2,3]**

[1] *Haskins Laboratories, New Haven, CT, USA*
[2] *Linguistics, University of Maryland, College Park, MD, USA*
[3] *Maryland Language Science Center, University of Maryland, College Park, MD, USA*

Edited by:
Claudia Felser, University of Potsdam, Germany

Reviewed by:
Shravan Vasishth, University of Potsdam, Germany
Patrick Sturt, University of Edinburgh, UK

****Correspondence:***
Dave Kush, Haskins Laboratories, 300 George St., Suite 900, New Haven, CT 06511, USA
e-mail: dave.w.kush@gmail.com

Because morphological and syntactic constraints govern the distribution of potential antecedents for local anaphors, local antecedent retrieval might be expected to make equal use of both syntactic and morphological cues. However, previous research (e.g., Dillon et al., 2013) has shown that local antecedent retrieval is not susceptible to the same morphological interference effects observed during the resolution of morphologically-driven grammatical dependencies, such as subject-verb agreement checking (e.g., Pearlmutter et al., 1999). Although this lack of interference has been taken as evidence that syntactic cues are given priority over morphological cues in local antecedent retrieval, the absence of interference could also be the result of a confound in the materials used: the post-verbal position of local anaphors in prior studies may obscure morphological interference that would otherwise be visible if the critical anaphor were in a different position. We investigated the licensing of local anaphors (reciprocals) in Hindi, an SOV language, in order to determine whether pre-verbal anaphors are subject to morphological interference from feature-matching distractors in a way that post-verbal anaphors are not. Computational simulations using a version of the ACT-R parser (Lewis and Vasishth, 2005) predicted that a feature-matching distractor should facilitate the processing of an unlicensed reciprocal if morphological cues are used in antecedent retrieval. In a self-paced reading study we found no evidence that distractors eased processing of an unlicensed reciprocal. However, the presence of a distractor increased difficulty of processing following the reciprocal. We discuss the significance of these results for theories of cue selection in retrieval.

Keywords: memory retrieval, anaphor resolution, Hindi, self-paced reading, computational modeling

INTRODUCTION

In order to establish grammatical dependencies between words across a distance during routine sentence processing comprehenders rely heavily on their ability to encode and retrieve items from memory. For example, processing of a local anaphor such as the reflexive *themselves* or the reciprocal *each other* in (1) requires recalling the previously seen noun phrase (NP) *the people* from memory so that it may be interpreted as the antecedent.

(1) The people talked to *themselves/each other*.

The mechanism by which previously encountered items are retrieved for subsequent processing has been the subject of recent research. A number of recent studies have motivated a processing model that exploits a cue-based access mechanism to retrieve items from content-addressable memory (e.g., McElree, 2000; McElree et al., 2003; Lewis et al., 2006; Van Dyke, 2007; Martin and McElree, 2008, 2009; Van Dyke and McElree, 2011).

A hallmark property of cue-based retrieval is that it is susceptible to interference (Nairne, 1990). Task-irrelevant items in memory whose features overlap with a probe's retrieval cues (distractors) can exert influence on the retrieval of a target item. In the context of sentence processing retrieval interference is said to occur when grammatically inappropriate distractors influence the processing of a phrase that must enter into a dependency with a previously encountered head. The influence of distractors can be *inhibitory*: a distractor may increase the difficulty of retrieving an appropriate item. Van Dyke (2007) found that distractor NPs increased the difficulty of retrieving a grammatically appropriate subject for the purposes of thematic integration with a verb (see also Van Dyke and McElree, 2006, 2011). A distractor's influence may also be *facilitatory* if its presence decreases the difficulty of processing an otherwise ungrammatical or unlicensed element. Comprehenders have repeatedly showed signs of facilitatory interference during the processing of subject-verb agreement (e.g., Pearlmutter et al., 1999; Wagers et al., 2009). Wagers and colleagues found that reading times immediately following the plural verb *were*, which mismatched the features of the singular subject *key*, were decreased when an intervening distractor [*cabinet(s)*] was plural, compared to when the distractor was singular.

(2) The key to the **cabinet(s)** unsurprisingly *were* rusty from years of disuse.

The authors argued that facilitation arose because comprehenders erroneously retrieved the plural distractor on some portion of trials when attempting to find a licensor for the plural marking on the verb. These kinds of facilitatory interference effects have also been observed in the processing of other grammatical dependencies such as negative polarity item (NPI) licensing (e.g., Drenhaus et al., 2005; Vasishth et al., 2008; Xiang et al., 2009; Parker and Phillips, submitted), and the retrieval of antecedents for null pronominal subjects (PRO) in adjunct clauses (Parker et al., 2012) and many authors have attributed these effects to misretrieval of a distractor under (partial) match with a set of retrieval cues.

Although facilitatory interference has been repeatedly observed in the processing of some dependencies, other dependencies that recruit retrieval have displayed virtual immunity to facilitation from distractors. Recent work has found that the processing of a local anaphor that lacks a grammatical antecedent is unaffected by the morphological feature-content of intervening distractors (e.g., Sturt, 2003; Dillon et al., 2013). For instance, Dillon et al. (2013) demonstrated that the processing of the unlicensed plural reflexive *themselves* in (3) is not influenced by plural-marking on the distractor *manager(s)*.

(3) The new executive who oversaw the manager(s) apparently doubted *themselves*...

The lack of facilitatory interference effects is unexpected on the assumption that the same cues as those used to find licensors for agreement dependencies (e.g., morphological features such as number) are used to identify potential antecedents of reflexives. As with agreement, reflexives must match their licensors in number and gender, so the use of morphological features as cue for retrieval of appropriate antecedents would appear to be motivated. On analogy to agreement licensing, the use of these morphological cues should in turn render antecedent retrieval subject to interference.

The results suggest that morphological features may play a different role in antecedent retrieval for local anaphors than they do in agreement licensing. One option, advocated by Dillon et al. (2013), is that antecedent retrieval forgoes the use of interference-prone morphological features, opting instead to exclusively use *positional* syntactic features to access the local subject. Another option is that antecedent retrieval preferentially weights syntactic cues over morphological cues instead of avoiding them altogether. This second account predicts a small but non-negligible interference effect that the first does not, but previous experiments may not have had sufficient power to find this effect, so they cannot distinguish between the two competing explanations.

Although the two accounts differ, they both assign priority to positional cues. This goes against the general assumption that retrieval identifies targets through the use of a maximal cue set that uniformly weights lexical, morphological, syntactic, and semantic features (see Van Dyke and McElree, 2011 for discussion).

As it stands the previous studies may not be sufficient to establish a preference for positional features. It is possible that the absence of facilitatory interference could be attributed to a confound that masks the contribution of morphological features that are weighted equally to syntactic cues. In almost all previous studies the critical anaphor immediately followed its verb, which could potentially play a role in reducing the incidence of facilitatory interference (see King et al., 2012 for a similar suggestion).

As Dillon et al. (2013) note, the post-verbal position can provide an anaphor with privileged access to the local subject by means of recent activation alone. If subjects are retrieved by their verbs for thematic integration, the local subject *the executive* in (3) should be recalled by the verb *doubted*. Retrieval of the local subject entails that it should have the highest baseline activation out of all other items in memory immediately following the verb. At the time that a verb-adjacent reflexive is encountered, this high degree of activation may be strong enough to guarantee retrieval of the local subject instead of the feature-matching distractor even if morphological cues were used.

Alternatively, it may be that previous studies on reflexives do not provide a measure of susceptibility to facilitatory interference because establishing a dependency between the local subject and a post-verbal anaphor might not require retrieval at all. Some theories assume that the most recently retrieved item is maintained in a state that the parser can access without retrieval. In some theories this state is referred to as the *focus of attention* (e.g., McElree, 2000), in others such as Lewis and Vasishth's (2005) parsing model it is the *problem buffer*. When an anaphor is encountered immediately following the verb, it is possible that it consults the contents of this buffer to find its antecedent rather than initiating a retrieval from memory.

In this study we address the extent to which the lack of facilitatory interference in anaphoric licensing depends on an anaphor's post-verbal position. If the absence of interference is a consequence of the target anaphor occupying an immediately post-verbal position, then in languages where anaphors uniformly precede their verbs, local anaphor licensing should display facilitatory effects that have not been seen in English. We tested this prediction by investigating the processing of Hindi reciprocals. Hindi is a language in which all arguments and adjuncts precede the verb in unmarked word order. In (4), for example, the subject *LaRkoN* ("boys"), the reciprocal object *ek-dusre* ("each other"), and the adjunct *kal* ("tomorrow") precede the verb *dekhaa* ("saw").

(4) LaRkoN-ne ek-dusre-ko kal dekhaa.
Boys-Erg each.other-Acc yesterday saw.
'(The) boys saw each other yesterday.'

Hindi reciprocals provide a minimal contrast to English reflexives because they are subject to nearly identical licensing conditions as English local anaphors. Their antecedent must have matching morphological features: in order to license the reciprocal in (5), the local subject must bear plural features. The reciprocal's antecedent must be contained in the same local clause as the reciprocal: the main clause subject in (6) cannot antecede the

reciprocal in the embedded clause, despite bearing correct number marking, because it is not local to the reciprocal. Finally, the reciprocal's antecedent must also c-command the reciprocal (cf. Dayal, 1994). In (7), the plural NP *boys* does not c-command the reciprocal because it is embedded inside the adjunct phrase *at the boys' party*. It is therefore ineligible to license the anaphor.

(5) *LaRk*-{*-e/oN}-ne *ek-dusre*-ko kal dekhaa.
Boy-{Sing./Pl.}-Erg each.other-Acc yesterday saw.
'(The) boy*(s) saw each other yesterday.'

(6) **LaRkoN*-ne kahaa ki Mary-ne *ek-dusre*-ko dekhaa.
Boys-Erg said that Mary-Erg each.other-Acc saw.
*'(The) boys said that Mary saw each other.'

(7) *Mary-ne [*larkoN*-ki parTi me] *ek-dusre*-ko dekhaa.
Mary-Erg boys' party in one-another-Acc saw
*'Mary saw each other at the boys' party.'

We test whether morphological number features engender facilitatory interference effects during the processing of Hindi reciprocals.

SIMULATIONS

We ran a series of computational simulations that modeled local anaphor resolution in Hindi using equally-weighted morphological and positional features as cues for retrieval. Modeling was carried out to obtain qualitative predictions about the character and direction of interference from the distractor's morphological features that could then be compared with empirical reading times in the self-paced reading experiment.

PROCEDURE

We implemented a modified version of Lewis and Vasishth's (2005) ACT-R model of sentence processing [using code originally developed by Badecker and Lewis (2007)]. ACT-R is a general cognitive architecture that has been used to model a wide range of phenomena in cognitive psychology (Anderson, 1990). In the model, items are stored as "chunks" in a content-addressable memory and are retrieved with a success proportional to their overall activation at the time of retrieval, which is in turn determined by the overlap of their features with those of a retrieval probe. Memory access is modeled as a rational procedure that employs a general retrieval mechanism that minimizes retrieval error in the limit (Anderson, 1989; Anderson and Milson, 1989; Anderson and Schooler, 1991). Although fully implemented ACT-R parsing models exist (e.g., Lewis and Vasishth's, 2005 ACT-R parser), the simulations here focus solely on modeling retrieval latencies, abstracting away from the contributions of other modules. Retrieval latencies do not exhaust the processes that must be carried out in order to advance to the next word in a parsing task (other operations include structural attachment and integration), but for current purposes we adopt the standard assumption that longer retrieval latencies entail longer RTs (Anderson and Milson, 1989).

In the model the probability of retrieving an item i is governed by its activation A_i, computed as in (8). B_i is chunk i's baseline activation. The weight assigned to the individual cue j is represented w_j. For the purposes of our simulations cues were assigned uniform weights, so this term can be effectively dropped. S_{ji} is the strength of association between cue j and chunk i. PM in the equation below is a term that penalizes partial matches. The term ε introduces stochastic noise.

$$(8)\quad A_i = B_i + \Sigma w_j S_{ji} + PM + \varepsilon_i$$

S_{ji} is calculated according to the Equation in (9), where S is a parameter that specifies the maximum strength of association allowed. The fan_j term reflects the number of items that bear cue j. The term provides a way of quantifying the distinctiveness of a particular cue. The fan serves to decrease the associative strength between item i and cue j as a function of the number of total cues in memory that bear j.

$$(9)\quad S_{ji} = S - \ln(fan_j)$$

Baseline activation is calculated according to (10), where d is the decay rate of a chunk's activation in memory at a given point since retrieval time t_m.

$$(10)\quad B_i = \ln[\Sigma_m\, t_m^{-d}]$$

The chunk with the highest activation has the shortest retrieval latency (T_i) as calculated according to the equation below, where *F* is a scaling parameter. The chunk with the shortest retrieval latency is the chunk that is retrieved in simulations.

$$(11)\quad T_i = Fe^{-A_i}$$

The model equations above contain a number of free parameters whose settings could impact the results of the simulation. We ran a series of simulations that systematically combined parameter values from across the range of those reported in previous work. Values of the *total source activation*, *activation noise*, *fan*, *decay rate*, and *match-penalty* parameters were manipulated[1]. The scaling factor (F) was held constant at 0.75 across all simulations. This resulted in the construction of 324 different models with unique parameter value combinations. As noted by Dillon et al. (2013), conducting such a sweep through the space of possible parameter values and combinations enables the identification of model predictions that are independent of idiosyncratic parameter combinations. 10,000 Monte Carlo simulations were run for each model, providing for each simulation a prediction of the most probable retrieval target and its retrieval latency.

MATERIALS

We simulated antecedent retrieval time-locked to a position corresponding to the critical reciprocal in a sentence that contained

[1]Total source activation took one of three values across our simulations: 1.0, 1.25, 1.5. Four values were possible for the activation noise parameter: 0.3, 0.4, 0.5, 0.6. Three values were used for the fan parameter: 1.0, 1.5, 2.0, centered at the default value of 1.5 (Lewis and Vasishth, 2005). Three decay rates were used: 0.5 (the default rate of decay; Lewis and Vasishth, 2005), 0.25, and.001. Finally, we used three values for the match-penalty parameter: −0.2, −0.4, and −0.6.

three preceding NPs. The first NP, the *subject*, corresponded to a structurally appropriate antecedent for the reciprocal. The second NP, introduced at a lag after the subject NP, corresponded to a structurally inappropriate distractor. A third NP (NP3) was also introduced to more directly model the materials in our self-paced reading (SPR) experiment, the design of which is discussed below. The three NPs were introduced at 300 ms, 900 ms, and 1500 ms after simulation onset. Retrieval of the critical reciprocal was scheduled at 2400 ms after simulation onset.

Each NP in the simulation was marked with three features relevant for retrieval: its *category*, *number*, and *clause index*. All NPs bore the NP category feature. Number features could be either *singular* or *plural*. The *clause index* feature was used as a proxy feature for encoding an NP's structural appropriateness for the purposes of binding the reciprocal: the local licensing requirement is assumed to be satisfied if the antecedent bears the same clause index as the reciprocal. This indexing scheme can be viewed as a feature-based implementation of the clause-mate constraint on local anaphor licensing (see Lasnik, 2002 for a review of such constraints, which can differ in formulation from the c-command constraints of Chomsky, 1981; Reinhart, 1983).

Models were run to simulate four distinct conditions, corresponding to different feature combinations on the subject and distractor. The number features on the subject and the distractor were manipulated, resulting in the 2 × 2 factorial design schematized in (12). In grammatical conditions the subject was plural-marked, in ungrammatical conditions the subject was singular. In *NoInterference* conditions the distractor was singular, while in *Interference* conditions it was plural-marked. The structurally appropriate subject NP was marked with the *main clause* feature, while both the distractor and NP3 were marked as *embedded* and were therefore ineligible to antecede the reciprocal.

(12) a. *Grammatical-NoInterference*
[Subject]+PL…[Distractor]+SG…[NP3]+SG…
[RECIPROCAL]+PL
b. *Grammatical-Interference*
[Subject]+PL…[Distractor]+PL…[NP3]+SG…
[RECIPROCAL]+PL
c. *Ungrammatical-NoInterference*
[Subject]+SG…[Distractor]+SG…[NP3]+SG…
[RECIPROCAL]+PL
b. *Ungrammatical-Interference*
[Subject]+SG…[Distractor]+PL…[NP3]+SG…
[RECIPROCAL]+PL

Antecedent retrieval at the reciprocal was modeled as specifying *NP* as a category cue and *main clause* as the clause cue. The number feature *plural* was also used in the retrieval cue set, to measure the interference effect associated with morphological features.

RESULTS

We report three measures of interest from the simulations run for each condition: (i) predicted error rate, (ii) average predicted latency by condition, and (iii) predicted interference effect.

Predicted error rate corresponds to the percentage of the runs when the distractor, rather than the appropriate subject, was retrieved as an antecedent for the reciprocal. This measure is a relevant index of facilitatory interference in the ungrammatical conditions if facilitation stems from erroneous retrieval of the distractor instead of an appropriate target NP.

Predicted latency provides a measure of how long on average the winning retrieval should take in each condition. In simulations, the chunk with the shortest retrieval latency is the chunk that is retrieved from memory. According to the fully implemented ACT-R model, reading times on a particular word or phrase are the sum of the latency of retrieval triggered at that phrase and the amount of time associated with subsequent processing required by that word or phrase. Retrieval latencies should therefore map monotonically to reading times, with longer retrieval latencies corresponding to longer overall reading times, although the mental processes that intervene between retrieval and button-press may interact or contribute additional difficulty in such a way as to distort the underlying pattern of retrieval. Despite the possibility of later processing concealing underlying retrieval patterns, previous work has found a degree of relative transparency between the qualitative pattern of retrieval latencies furnished by the model and observed effects of facilitatory interference in self-paced reading or eye-tracking measures (see e.g., Wagers et al., 2009; Dillon et al., 2013).

The interference effect is a difference measure that compares average retrieval latencies between two conditions that differ on a single feature, as a way of estimating the magnitude and direction of interference contributed by the retrieval probe matching that one feature. We report two interference effects: the difference between the two grammatical conditions, as well as the difference between the two ungrammatical conditions. These comparisons provide a quantitative prediction of the effect of distractor plural marking when the features of the appropriate subject are held constant.

PREDICTED ERROR RATES

Error rates are reported in **Table 1**. The error rates are consistent with a profile of facilitatory interference. Between the *Ungrammatical* conditions, plural marking on the distractor is predicted to increase rates of erroneous retrieval compared to when there is no NP in the sentence that matches the reciprocal in features (26.1 vs. 6.5%). On some proportion of trials, the recency of the distractor is predicted to increase the NP's baseline level of activation enough to result in it being the most highly-activated NP at retrieval. In the *Ungrammatical-NoInterference* condition, the distractor does not share any features with the reciprocal's cue set, so the main subject is still more likely to be retrieved, as it matches the retrieval probe's clause index cue. Error rate is

Table 1 | Retrieval error rates by condition for retrieval using morphological and syntactic cues calculated as the percentage of trials on which the distractor was retrieved across 10,000 runs each of 324 different models with unique parameter combinations.

	NoInterference (%)	Interference (%)
Grammatical	2.0	7.4
Ungrammatical	6.5	26.1

expected to differ slightly between the two grammatical conditions: misretrieval of the distractor is 5.4% more common when it bears plural marking and the main subject matches the retrieval cues completely.

AVERAGE PREDICTED RETRIEVAL LATENCIES

In the simulations the presence of a feature-matching subject has a facilitative effect on retrieval latencies (see **Figure 1**). Overall, retrieval times should be faster in the grammatical conditions because the grammatical subject, which matches the reciprocal's morphological and syntactic retrieval cues completely, is highly activated. Increased activation due to greater feature-match with the probe results in faster retrieval latencies in accordance with Equation (11). In the *Ungrammatical* conditions, where the main subject matches only on syntactic cues, retrieval latencies should be longer because the retrieved chunk should never match the probe completely. The appropriate subject only matches the probe's category and positional cues. The distractor matches the category cue and, in the *Ungrammatical-Interference* condition, the reciprocal's number feature. A pairwise difference is also predicted between the average retrieval latencies in the *Ungrammatical-NoInterference* and *Ungrammatical-Interference* conditions, which can be linked to the presence of morphological plural marking on the distractor. On the proportion of trials where the distractor is retrieved in the *Ungrammatical-Interference* condition, latencies are reduced relative to when the main clause subject is retrieved. This results in a reduction of average latency across retrievals.

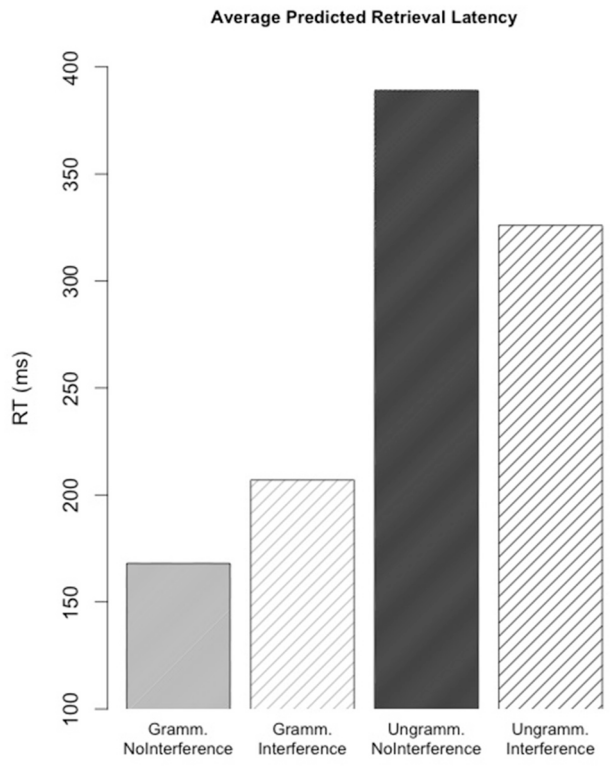


FIGURE 1 | Retrieval latencies by condition as predicted by the model in Experiment 1. Reported retrieval latencies represent the mean latency by condition across all simulations.

INTERFERENCE EFFECTS

Predicted interference effects are shown in **Table 2**. The grammatical interference effect was calculated by subtracting the average predicted latency in the *Grammatical-Interference* condition from the predicted latency in the *Grammatical-NoInterference* condition. The same difference was calculated for the two ungrammatical conditions. 95% confidence intervals represent the range of predicted interference effects across simulations.

The simulation results predict that a plural-marked distractor should cause facilitatory interference in the ungrammatical conditions. The *Ungrammatical-Interference* condition exhibits faster average retrieval latencies than the *Ungrammatical-NoInterference* condition. Though the size of the effect varies, a facilitatory effect was consistently observed across all parameter combinations.

A small effect of inhibitory interference is also predicted in the grammatical conditions. This inhibition can be attributed to the fan effect (see Equation 8). In the *Grammatical-Interference* condition, the strength of association between the appropriate subject and the plural retrieval cue is decreased relative to the *Grammatical-NoInterference* condition, due to the presence of another plural-marked NP (the distractor).

DISCUSSION

The goal of the simulations was to obtain predictions about the effect that a feature-matching but syntactically inappropriate distractor would have on the retrieval of an antecedent for a local reflexive if that retrieval used morphological features as cues that were weighted equally to syntactic cues.

The simulations show that when morphological cues are assigned the same weight as syntactic cues, the presence of a feature-matching distractor should decrease the parser's ability to retrieve a syntactically appropriate but feature-mismatching subject as an antecedent for a local anaphor. Some proportion of the time, the distractor is expected to be erroneously retrieved as a result of partial overlap with the retrieval cues. This misretrieval is predicted to have a facilitating effect on reading times in comparison to a case of retrieval when neither the distractor nor the local subject match the reflexive.

SELF-PACED READING EXPERIMENT

The modeling results predict that retrieval of a pre-verbal reciprocal's antecedent should display facilitatory interference effects from structurally inappropriate distractors, if morphological cues

Table 2 | Average interference effects across 10,000 runs each of 324 different models.

	Interference effect	Middle 95% of simulated distributions
Grammatical	+36 ms	+11, +82 ms
Ungrammatical	−63 ms	−18, −142 ms

such as number are assigned the same weight as syntactic cues in retrieval. The experiment below used the self-paced reading method to investigate whether evidence of the predicted facilitatory interference would be found.

MATERIALS

The experiment had a 2×2 factorial design that matched the simulated conditions. The design manipulated the factors GRAMMATICALITY and INTERFERENCE. The structure of the test items is schematized in (13) and an example item is given in (14). All conditions contained a critical reciprocal (*ek-dusre*) that required a plural-marked antecedent in the main clause. The reciprocal was contained in a post-positional phrase that preceded a manner adverbial (*gupt-rup-se,* "*secretly*") and the main clause verb (*baat kii,* "*chatted*" *lit.* "*chat did*").

GRAMMATICALITY was manipulated by changing whether the main clause subject was plural-marked [*Doctor(-oN),* "*doctor(s)*"]. Plural marking was unambiguously marked by the inflectional suffix *–oN*. In the grammatical conditions, the main clause subject was plural and could therefore act as a grammatical antecedent for the reciprocal. In the ungrammatical conditions, the local subject was singular and the reciprocal therefore lacked a clause-mate antecedent. The factor INTERFERENCE manipulated whether the distractor [*mariiz-(oN),* "*patient(s)*"] was plural-marked.

In previous studies on local anaphor licensing (e.g., Sturt, 2003; Dillon et al., 2013) distractors have been positioned within relative clauses (RCs) attached to the main clause subject. RC-modification of subjects is a marked construction in Hindi, so the present study embedded the distractor inside a locative phrase that preceded the critical reciprocal.

The locative phrase contained an NP denoting a location modified by an animate possessor (*nurse-ke steSan,* "the nurse's station"). The distractor was embedded as the object of a verb within a prenominal RC that was attached to this possessor. In this position the distractor was not a clause-mate of the reciprocal and was therefore ineligible to act as a potential antecedent.

Critical reciprocals were always followed by a case marking post-position, either the genitive *ke*, the objective *ko*, or the dative *se*. When followed by the genitive, reciprocals were embedded in a complex post-position that was an argument to the main verb (e.g., *ke bare-me* "about" in 14). In sentences with *ko* or *se*, adverbial material was introduced after the post-position to maintain consistent length across sentences.

(13) Subject-{sg/pl} [$_{PP}$[$_{RC}$ Distractor-{sg/pl} V] NP's Location] Reciprocal P Adv V

(14) a. *Grammatical-NoInterference*
DoctoroN-ne mariiz-ki dekhbaal karne-wali nars-ke sTeSan-me ek-dusre ke-bare-me gupt-rup-se baat kii.
Doctors-Erg patient-Gen care doing-RP nurse's station-in each-other aboutsecretly chat did.
'The doctors secretly spoke about each other in the station of the nurse taking care of (a/the) patient.'

b. *Grammatical-Interference*
DoctoroN-ne mariizoN-ki dekhbaal karne-wali nars-ke sTeSan-me ek-dusre ke-bare-me gupt-rup-se baat kii.
Doctors-Erg patients-Gen care doing-RP nurse's station-in each-other about secretly chat did.
'The doctors secretly spoke about each other in the station of the nurse taking care of (the) patients.'

c. *Ungrammatical-NoInterference*
Doctor-ne mariiz-Gen dekhbaal karne-wali nars-ke sTeSan-me ek-dusre ke-bare-me gupt-rup-se baat kii.
Doctor-Erg patient-ki care doing-RP nurse's station-in each-other about secretly chat did.
'The doctor secretly spoke about each other in the station of the nurse taking care of (a/the) patient.'

d. *Ungrammatical-Interference*
Doctor-ne mariizoN-ki dekhbaal karne-wali nars-ke sTeSan-me ek-dusre ke-bare-me gupt-rup-se baat kii.
Doctor-Erg patients-Gen care doing-RP nurse's station-in each-other about secretly chat did.
'The doctor secretly spoke about each other in the station of the nurse taking care of (the) patients.'

Inside the pre-nominal RC the distractor bore either accusative or genitive case (according to the verb's requirements). Although this increased the contrast between the nominative grammatical subject and the distractor, it is unlikely that the case difference would play a role in distinguishing appropriate from inappropriate NPs, as accusative and genitive-marked NPs can serve as antecedents for local anaphors under the right structural conditions (see, e.g., Dayal, 1994; Mohanan, 1994; Bhatt and Dayal, 2007).

A second concern with the experimental materials is that there exists the potential for temporary misanalysis of the structural position of the distractor during incremental parsing. When it initially encounters the distractor, the parser has not yet encountered any information that indicates that the distractor is contained within an embedded clause. In the absence of this information, an incremental parser is likely to analyze the distractor as a constituent of the main clause. This type of temporary misparse is common in head-final languages where embedded arguments can be encountered prior to the verb that licenses them (Inoue, 1991; Mazuka and Itoh, 1995; Miyamoto, 2003). The misanalysis would be disconfirmed at the relative pronoun *wali*, at which point the object would be correctly reanalyzed as a constituent of the relative clause. This misparse should occur across all conditions, but it may have a greater impact on processing in the *Ungrammatical-Interference* condition. Under this misanalysis the RC-internal object would initially be analyzed as a suitable antecedent for an upcoming reciprocal. We return to the ability of such a misparse to affect later parsing decisions in the *Ungrammatical-Interference* condition in the discussion.

PARTICIPANTS

32 self-reported native speakers of Hindi were recruited from the student bodies of IIT, Delhi and Jawaharlal Nehru University in New Delhi (18 male, mean age = 20.1). Participants were compensated Rs. 300 for their participation, which lasted around 35 min.

PROCEDURE

Participants were run on one of two laptop PCs using the Linger software package (Doug Rohde, MIT) in a self-paced word-by-word moving window paradigm (Just et al., 1982). Each trial began with a sentence masked by dashes appearing on the screen. Letters and punctuation marks were masked, but spaces were left unmasked so that word-boundaries were visible. As the participant pressed the spacebar, a new word appeared and the previous word was re-masked. All text appeared in Devanagari font.

A yes/no comprehension question that probed its interpretation followed each sentence (experimental materials can be found at the first author's website). Participants were instructed to read sentences at a natural pace and to respond to the comprehension questions as accurately as possible. Participants responded to questions using the f-key for "yes" and the j-key for "no." If the question was answered incorrectly the word *galat* ("incorrect/wrong") appeared briefly in the center of the screen. Each participant was randomly assigned to one of the lists and the order of the stimuli within the presentation list was randomized for each participant.

ANALYSIS

Data from one participant were excluded due to failure to comply with experimental guidelines. Data from another participant were excluded because the participant's mean accuracy on comprehension questions was close to chance. This resulted in the data of 30 subjects being used for later analysis. Two items were excluded from analysis due to errors.

Statistical analyses were carried out on log-transformed reading times using linear mixed effects regression (Baayen et al., 2008). Reading times from both correct and incorrect trials were included in the analysis. Experimental fixed effects were the simple difference sum-coded factors GRAMMATICALITY and INTERFERENCE and their interaction. All models included random intercepts for both subjects and items. Models with a maximal random effects structure were fit whenever possible (Barr et al., 2013). If a maximal model failed to converge, a model was used that contained only by-subject random slopes for both fixed effects and their interaction.

RESULTS

Comprehension Question Accuracy

Comprehension question accuracy averaged 69.2%. No significant differences were found in average accuracy across conditions (logistic mixed effects model, all zs < 1).

Reading Time Results

Reading times from the post-reciprocal region are given in **Figure 2**.

Pre-reciprocal region. No significant effects were found in the pre-reciprocal region.

Reciprocal region. No significant effects were found in the reciprocal region.

Post-position region. Average reading times were reliably faster in the grammatical conditions than in the ungrammatical conditions (main effect of GRAMMATICALITY: $\hat{\beta} = -0.088$, s.e. $=$ 0.034, $t = -2.92$); see **Figure 3**. Although reading times in the *Ungrammatical-Interference* condition were numerically longer than those in the *Ungrammatical-NoInterference* condition, the GRAMMATICALITY $\times$ INTERFERENCE interaction was not significant ($t = 1.41$). No reliable pairwise differences were observed between ungrammatical conditions ($t < 1$).

Reciprocal+2 region. There were no significant main effects two regions after the critical reciprocal, but the model revealed a marginally significant GRAMMATICALITY $\times$ INTERFERENCE interaction ($\hat{\beta} = 0.105$, s.e. $= 0.054$, $t = 1.96$) two regions after the reciprocal. This interaction reflected the fact that the *Ungrammatical-Interference* condition was read more slowly than any other condition, including the *Ungrammatical-NoInterference* condition. The pairwise comparison between the two ungrammatical conditions revealed the numerical difference between the

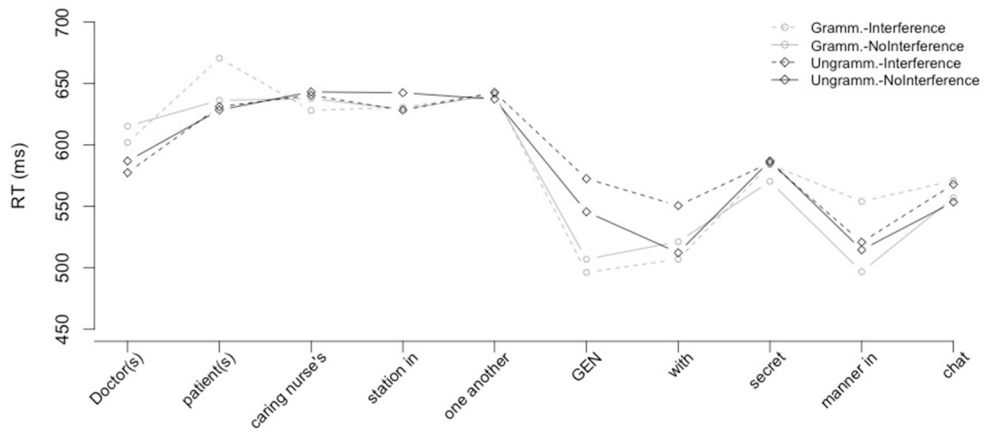


FIGURE 2 | Average word-by-word self-paced reading times for all items in Experiment 1.

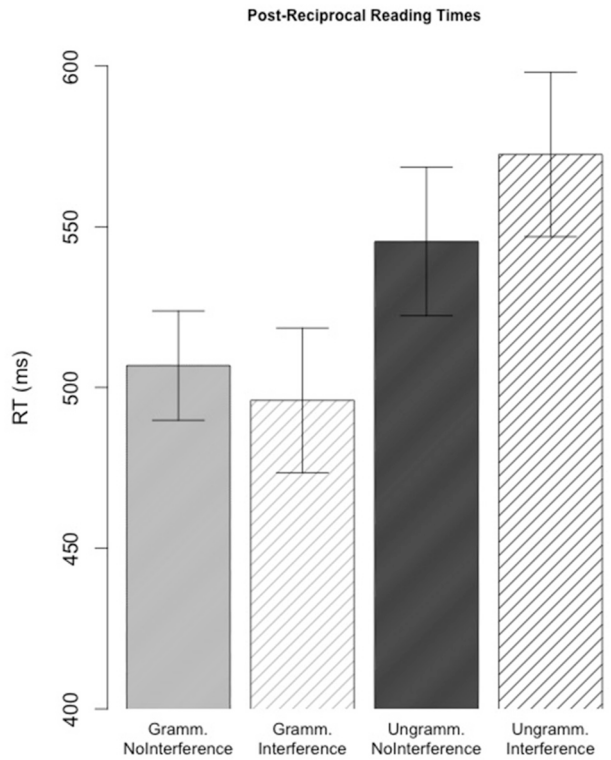


FIGURE 3 | Average post-reciprocal self-paced reading times in Experiment 1. Error bars indicate standard error of the participant mean.

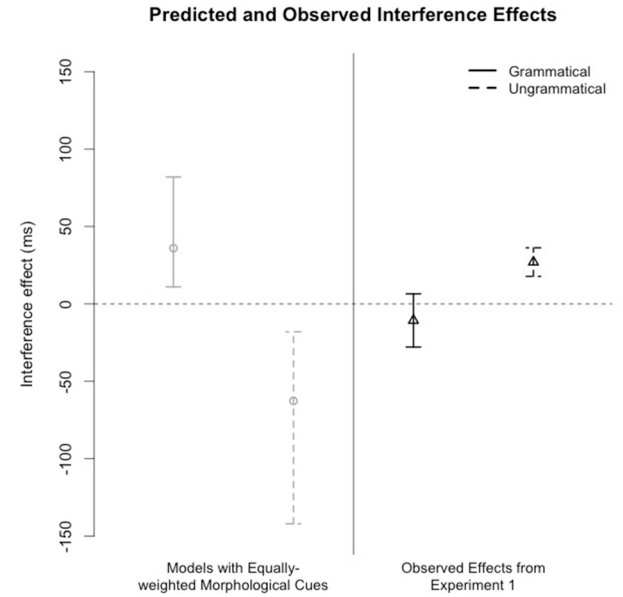


FIGURE 4 | Comparison of predicted interference effects from model simulations and observed interference effects from Experiment 1. Models simulated expected retrieval latencies if morphological and positional cues were assigned equal weights in antecedent retrieval. For the simulated data, error bars represent the middle 95% of the distribution of predicted interference effects. Error bars around the empirical means mark the 95% CI.

two conditions not to be significant ($t = 1.3$). However, given the relatively low power of the current study, it is possible that this interaction would achieve significance with higher power. We return to this interaction effect in the discussion.

Reciprocal+3 till Final region. No significant effects were observed in any subsequent region.

DISCUSSION

The SPR experiment sought to determine whether the processing of a pre-verbal reciprocal in Hindi was subject to facilitatory interference. The study manipulated the number features on a structurally appropriate antecedent for the reciprocal, as well as the features of the structurally inappropriate distractor, as a means of testing whether (equally weighted) morphological cues are used to access a local anaphor's antecedent.

When a structurally appropriate feature-matching antecedent was present to license the pre-verbal reciprocal the regions following the critical reciprocal were read more rapidly than when there was no feature-matching and structurally appropriate antecedent. In contrast to the prediction of the model simulations, we failed to find any evidence of facilitatory interference (see **Figure 4**). In fact, the empirical results trend in the opposite direction; there were clear inhibitory effects. The post-reciprocal region in the *Ungrammatical-Interference* condition was read at a comparable or slightly slower rate than the processing of the reciprocal in the *Ungrammatical-NoInterference* condition. Despite the fact that our study potentially lacks the power to observe an interference effect, we are more secure in our conclusion that there is a lack of facilitatory interference in light of the direction of the numerical trend toward an interaction in the post-reciprocal region.

Two words downstream from the reciprocal, reading times were longest when the local subject did not match the features of the reciprocal but the features of the distractor did match the reciprocal's number features.[2] We discuss this effect below because although it is inconsistent with the predictions of our simulations, it does potentially indicate that the distractor's morphological features may affect overall processing of the reciprocal.

The mechanism by which the distractor exerts inhibitory influence on reciprocal licensing is unclear. It is commonly assumed that inhibitory interference should occur when multiple items in memory match a retrieval cue (e.g., Badecker and Straub, 2002; Lewis and Vasishth, 2005; Van Dyke and McElree, 2011). Yet, we observed inhibition in the absence of a multiple-match configuration: the main subject matched the positional cue and the distractor matched the number cue. This suggests that the mechanism used to explain inhibition in multiple-match cases (e.g., the fan effect in Lewis and Vasishth's, 2005 model), is not the appropriate explanation for our finding. We consider three possible explanations of this inhibitory effect and the role

[2]We expect that this effect would be stronger with a larger sample.

that number features play in guiding initial retrieval under each scenario.

The first possible interpretation of the inhibitory effect links the slightly delayed slowdown to erroneous retrieval of the distractor during initial memory access. The increased reading times in our SPR experiment might reflect initial misretrieval of the distractor based on its morphological overlap with the probe, followed by the increased processing cost of inhibiting that distractor. This line of reasoning has been pursued by Patil et al. (2012) and Chen et al. (2012) to explain inhibitory effects in reflexive licensing. We consider this interpretation unlikely for the present data because we see no evidence of the erroneous retrieval on which the explanation is predicated. In light of subject-verb agreement and NPI licensing effects, we would expect initial misretrieval to result in some degree of facilitation, however fleeting, that would be observable in the self-paced reading times. These facilitatory interference effects consistently yield large effects on reading times in studies of other linguistic dependencies. No such facilitation was observed prior to the point of inhibitory interference in the current study.

The inhibitory effect might also be explained in terms of *cue-confusability*, as defined by Jäger et al. (2014). The proposal rests on the speculation that cues that reliably co-occur in specific retrieval contexts can be confused (less effectively deployed). In reciprocal licensing the clause and plural cues are reliably associated because both cues should be selected whenever a reciprocal is encountered. This contrasts with cue association in reflexive licensing where specific gender cues (e.g., masculine and feminine) and the clause-mate cue co-occur less reliably, e.g., it is not the case that reflexive licensing uniformly uses masculine gender. According to the proposal, confusion is more likely to occur in reciprocal licensing than in reflexive licensing. Although we note that this is a possibility in principle, we believe that the notion of cue-confusability or the mechanism by which confusion creates retrieval interference has not been sufficiently articulated to be thoroughly evaluated.

The third alternative interpretation of the effect connects the slowdown to the influence of an abandoned early garden-path parse that analyzed the distractor as a constituent of the main clause. The previous partial parse could provide an appropriately marked antecedent for the reciprocal, but would fail to provide a coherent global parse. There are no grammatical re-parses of the sentence that would allow the distractor to be reanalyzed as an appropriate antecedent for the reciprocal. We hypothesize that resolving the tension between attempting to license the reciprocal and building a globally grammatical parse of the sentence is the source of the observed interaction. The misparse is expected to intrude on the processing of the reciprocal in the *Ungrammatical-Interference* condition, where consideration of the reparse would result in a structurally appropriate, feature-matching antecedent for the reciprocal.

We favor the interpretation that this inhibitory effect reflects the influence of the mis-parse on repair strategies that are triggered by failure of initial antecedent retrieval (as proposed for similar effects by, Sturt, 2003; Chow et al., 2014). On this interpretation the failure to retrieve an appropriate antecedent for the reciprocal would initiate a more liberal search for a feature-matching phrase, or would attempt to find an alternative parse for the sentence under which the reciprocal could be grammatically bound. These repair procedures are argued to be less constrained by the structure of the previous parse (and therefore structural constraints), perhaps reflecting uncertainty in the structural analysis in light of the error signal. This scenario attributes the increase in reading times to interference, but not interference that occurs during antecedent retrieval. Rather, the locus of interference lies in retrievals associated with syntactic revision and reanalysis. It is also possible that the distractor in the mis-parsed sentence could contribute interference at retrieval time, a possibility that would be consistent with the numerical trend toward an interaction in the post-reciprocal region. We acknowledge that the present study cannot distinguish between these two options.

In sum, our SPR experiment failed to find the characteristic profile of facilitatory interference that has been found in other studies on the construction of subject-verb agreement, NPI-licensing, and control dependencies and is predicted under a cue-based retrieval model that uses morphological cues to access potential antecedents for a local anaphor. Instead, a feature-matching distractor triggered a delayed inhibitory effect when the local subject could not antecede the reciprocal in Hindi. We argued that this process was not an indication of interference during antecedent retrieval, but rather interference during a repair process subsequent to antecedent retrieval.

GENERAL DISCUSSION

The purpose of the present study was to assess whether syntactic cues are given priority over morphological cues in the retrieval of antecedents of pre-verbal reciprocals in Hindi. Investigating the processing of Hindi reciprocals helps to establish whether the absence of facilitatory interference effects from morphologically-matched distractors in previous experiments was due to a confound of anaphor position. We hypothesized that if the absence of interference were solely due to the post-verbal position of the anaphor, and not prioritization of syntactic cues, interference would be observable in the retrieval of an antecedent for a pre-verbal anaphor in Hindi.

In our self-paced reading study native Hindi speaking participants resolved a local reciprocal dependency more quickly when the main clause subject was plural than when no grammatical antecedent was present. The presence of a feature-matching distractor did not induce reliable effects of facilitatory interference when the local subject did not match the reciprocal in features. These findings are consistent with a general lack of facilitation in the licensing of local anaphors found in previous work (e.g., Sturt, 2003; Xiang et al., 2009; Dillon et al., 2013), and with lack of interference during local anaphor licensing more generally (e.g., Nicol and Swinney, 1989; Clackson, 2011). The presence of a feature-matching distractor produced a delayed inhibitory effect when an appropriate antecedent for the reciprocal could not be found. We reasoned that the inhibitory effect in our experiment might have arisen as a result of error-driven repair strategies, and not from participants accessing the distractor during initial antecedent retrieval.

The empirical results of our SPR experiment were compared against the results of a series of simulations that modeled latencies and error rates of a cue-based retrieval process that used equally-weighted morphological and positional cues to retrieve antecedents of a local anaphor. The empirical results did not align with the simulations' prediction that there should be facilitatory interference between ungrammatical conditions.

Overall, the results lend support to the hypothesis that the lack of facilitatory interference in local anaphor antecedent retrieval is not primarily determined by an anaphor's post-verbal position. In particular, the Hindi results appear to be incompatible with a number of the possible ways in which verbal adjacency could influence retrieval of antecedents for local anaphors discussed. The results cast doubt on explanations that rely on recent reactivation of the grammatical antecedent immediately before the reciprocal. In the Hindi materials there is no point at which retrieval of the subject is required between the distractor and when the reciprocal is encountered.

The results are consistent with models of cue-based antecedent retrieval that prioritize syntactic information in one manner or another. As noted in the introduction, a parser could be said to prioritize syntactic cues by assigning them greater weight than morphological cues, or by using syntactic cues exclusively.

Because some dependencies display facilitatory interference effects while others do not, it would appear that retrieval does not consistently prioritize positional cues. One question that arises is how the parser determines when it should prioritize syntactic cues. Rational models often assume that retrieval uses a set of cues and weights that maximizes the probability of retrieving the target, while minimizing the chances of interference. It is important to note that the optimal cue set for meeting both of these goals may change as a function of (i) the dependency being computed and (ii) the local syntactic context. Therefore, strategic considerations that take the local context into account may comprise an important part of the cue selection procedure. We term different solutions that the parser could adopt *retrieval strategies*.

The parser could adopt one of two strategies that make different use of morphological cues during local antecedent retrieval. First, the parser could uniformly prioritize syntactic cues for all instances of local antecedent retrieval regardless of syntactic context. Dillon et al. (2013) proposed that the parser implements such a retrieval strategy. According to these authors, local antecedent retrieval only uses structural cues.

An alternative to this proposal is that the parser could condition the use of morphological cues on the local syntactic context of the anaphor that triggers retrieval, as proposed by Kush (2013). The intuition behind this proposal stems from the observation that in certain environments structural cues alone may not suffice to identify a unique antecedent for a local anaphor. If the subject of the local clause is the anaphor's only co-argument, as it is in (15), then syntactic cues are sufficient to guarantee its retrieval. However, if there exists an additional co-argument that precedes the anaphor as in (16), a syntactic cue like the clause feature would not be able to distinguish the appropriate antecedent (*the boys*) from the structurally appropriate, but feature-mismatching NP *Mary*.

(15) *The boys* spoke with *each other*.
(16) Mary introduced *the boys* to *each other*

Kush (2013) proposed that a parser that could determine the number of clause-mates that preceded a local anaphor might use morphological cues to help guarantee retrieval of an appropriate antecedent. Determining whether the local subject is the anaphor's only clause-mate should be possible by consulting the local syntactic context. When processing English reflexives in direct object position, the anaphor's adjacency to the verb would be sufficient. In Hindi, verbal adjacency cannot be exploited to make such a determination. Kush (2013) proposed that the decision could be made if cue selection had access to the phrase structure rule being used to incrementally parse the input sentence. In cases where the anaphor is the first NP encountered during the incremental parse of the VP, the phrase structure predicted for the VP should not contain co-argument NPs. On the other hand, if the parser encounters a non-subject co-argument that precedes the reciprocal, the PS rule for the VP would reflect its presence and cue selection could determine that the clause index cue would no longer provide diagnostic access to the local subject.

If the parser adopts this retrieval strategy interference effects are predicted to emerge when there are non-subject clause-mates that precede a local anaphor. This proposal is consistent with recent findings from Wagers and colleagues, which suggest that that resistance to interference is, in fact, selectively conditioned on whether the anaphor is encountered after another co-argument (King et al., 2012). Under this interpretation, interference should emerge if a co-argument preceded the reciprocal in Hindi, as in (17). We leave testing this prediction to future work.

(17) *Larke-ne Mary-ko baccoN-ki party me ek-dusre ke-bare-me bataayaa.
Boy-Erg Mary-Acc kids' party in one-another about told.
*The boy told Mary during the kids' party about *each other*.

CONCLUSION

In this paper we asked whether the absence of intrusive licensing during local anaphor antecedent retrieval is restricted to post-verbal anaphors, or whether the lack of interference indicates a more general cross-linguistic state of affairs. We investigated the effect of a feature-matching distractor on the processing of unlicensed pre-verbal reciprocals in Hindi and found no indication of facilitatory interference. The results suggest that antecedent retrieval's ability to accesses the syntactically appropriate subject when licensing a local anaphor does not depend on direct verbal adjacency between the anaphor and its verb. The results appear to be better explained by a cue-based retrieval process that prioritizes, or exclusively uses, structural cues over morphological features. Finally, although we did not find evidence that a feature-matching distractor facilitates the processing of an unlicensed reciprocal, it did appear that a distractor might exert an inhibitory influence on some stage of reciprocal resolution. Future work should test whether this inhibition is a general effect, or whether its appearance is related to properties of the materials used here.

ACKNOWLEDGMENTS

The results from our SPR experiment were previously presented at the CUNY Sentence Processing and GLOW Conferences in 2012. We thank Usha Udaar, Karthik, Pritha Chandra and Ayesha Kidwai for their assistance in recruiting participants and providing space to run Experiment 1. We thank Ashok, Ambrish, and Manu Kush for helping to construct the materials for the SPR experiment. The code used for the simulations was generously provided by Rick Lewis. We thank Brian Dillon, Pedro Alcocer, and Dan Parker for their subsequent revisions and additions to this code. Shravan Vasishth and Umesh Patil also provided useful discussion. This work was supported by NSF IGERT DGE-0801465 to University of Maryland.

REFERENCES

Anderson, J. R. (1989). "A rational analysis of human memory," in *Varieties of Memory and Consciousness: Essays in Honor of Endel Tulving*, eds H. Roediger III and F. Craik (Hillsdale, NJ: Erlbaum), 195–210.

Anderson, J. R. (1990). *The Adaptive Character of Thought*. Hillsdale, NJ: Erlbaum.

Anderson, J. R., and Milson, R. (1989). Human memory: an adaptive perspective. *Psychol. Rev.* 96, 703–719. doi: 10.1037/0033-295X.96.4.703

Anderson, J. R., and Schooler, L. (1991). Reflections of the environment in memory. *Psychol. Sci.* 2, 396–408. doi: 10.1111/j.1467-9280.1991.tb00174.x

Baayen, R. H., Davidson, D. J., and Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *J. Mem. Lang.* 59, 390–412. doi: 10.1016/j.jml.2007.12.005

Badecker, B., and Lewis, R. (2007). "A new theory and computational model of working memory in sentence production: agreement errors as failures of cue-based retrieval," in *Paper Presented at the 20th Annual CUNY Sentence Processing Conference* (San Diego; La Jolla, CA: University of California).

Badecker, W., and Straub, K. (2002). The processing role of structural constraints on the interpretation of pronouns and anaphors. *J.Exp. Psychol. Learn. Mem. Cogn.* 28, 748–769. doi: 10.1037/0278-7393.28.4.748

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: keep it maximal. *J. Mem. Lang.* 68, 255–278. doi: 10.1016/j.jml.2012.11.001

Bhatt, R., and Dayal, V. (2007). Rightward scrambling as rightward remnant movement. *Linguist. Inq.* 38, 287–301. doi: 10.1162/ling.2007.38.2.287

Chen, Z., Jäger, L., and Vasishth, S. (2012). "How structure-sensitive is the parser? Evidence from Mandarin Chinese," in *Empirical Approaches to Linguistic Theory: Studies of Meaning and Structure*, eds B. Stolterfoht and S. Featherston (Berlin: Studies in Generative Grammar, Mouton de Gruyter), 43–62.

Chomsky, N. (1981). *Lectures on Government and Binding*. Berlin: Mouton de Gruyter.

Chow, W.-Y., Lewis, S., and Phillips, C. (2014). Immediate sensitivity to structural constraints in pronoun resolution. *Front. Psychol.* 5:630. doi: 10.3389/fpsyg.2014.00630

Clackson, K. (2011). *Reflexives and Pronouns in Sentence Processing: an Experimental Study of Children and Adults*. Colchester, UK: Doctoral Dissertation, University of Essex.

Dayal, V. (1994). "Binding facts in Hindi and the scrambling phenomenon," in *Theoretical Perspectives on Word Order Issues in South Asian Languages*, eds M. Butt, T. H. King and G. Ramchand (Stanford, CA: CSLI Publications), 237–262.

Dillon, B., Mishler, A., Slogett, S., and Phillips, C. (2013). Contrasting intrusion profiles for agreement and anaphora: experimental and modeling evidence. *J. Mem. Lang.* 69, 85–103. doi: 10.1016/j.jml.2013.04.003

Drenhaus, H., Frisch, S., and Saddy, D. (2005). "Processing negative polarity items: when negation comes through the backdoor," in *Linguistic Evidence: Empirical, Theoretical, and Computational Perspectives*, eds S. Kepser and M. Reis (Berlin: Mouton de Gruyter), 145–165.

Inoue, A. (1991). *A Comparative Study of Parsing in English and Japanese*. PhD dissertation. Mansfield, CT: University of Connecticut.

Jäger, L., Engelmann, F., and Vasishth, S. (2014). "Inhibitory interference in reflexives: Evidence for cue confusability," in *Poster AMLaP 2014* (Edinburgh).

Just, M. A., Carpenter, P. A., and Woolley, J. D. (1982). Paradigms and processes and in reading comprehension. *J. Exp. Psychol. Gen.* 3, 228–238. doi: 10.1037/0096-3445.111.2.228

King, J., Andrews, C., and Wagers, M. (2012). "Do reflexives always find grammatical antecedents for themselves?" in *Poster, 25th Annual CUNY Human Sentence Processing Conference* (New York, NY).

Kush, D. (2013). *Respecting Relations: Memory Access and Antecedent Retrieval in Incremental Sentence Processing*. Doctoral Dissertation, University of Maryland, College Park.

Lasnik, H. (2002). Clause-mate conditions revisited. *Glot Int.* 6, 94–96.

Lewis, R., and Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cogn. Sci.* 29, 375–419. doi: 10.1207/s15516709cog0000_25

Lewis, R., Vasishth, S., and Van Dyke, J. (2006). Computational principles of working memory in sentence comprehension. *Trends Cogn. Sci.* 10, 447–454. doi: 10.1016/j.tics.2006.08.007

Martin, A. E., and McElree, B. (2008). A content-addressable pointer mechanism underlies comprehension of verb-phrase ellipsis. *J. Mem. Lang.* 58, 879–906. doi: 10.1016/j.jml.2007.06.010

Martin, A. E., and McElree, B. (2009). Memory operations that support language comprehension: evidence from verb-phrase ellipsis. *J. Exp. Psychol. Learn. Mem. Cogn.* 35, 1231–1239. doi: 10.1037/a0016271

Mazuka, R., and Itoh, K. (1995). "Can Japanese speakers be led down the garden path?" in *Japanese Sentence Processing*, eds R. Mazuka and N. Nagai (Hillsdale, NJ: Lawrence Erlbaum Associates), 295–329.

McElree, B. (2000). Sentence comprehension is mediated by content-addressable memory structures. *J. Psycholinguist. Res.* 29, 111–123. doi: 10.1023/A:1005184709695

McElree, B., Foraker, S., and Dyer, L. (2003). Memory structures that subserve sentence comprehension. *J. Mem. Lang.* 48, 67–91. doi: 10.1016/S0749-596X(02)00515-6

Miyamoto, E. T. (2003). Reanalysis of clause boundaries in Japanese as a constraint-driven process. *Lang. Speech* 46, 23–50. doi: 10.1177/00238309030460010301

Mohanan, T. (1994). *Argument Structure in Hindi*. Stanford, CA: CSLI Publications.

Nairne, J. S. (1990). A feature model of immediate memory. *Mem. Cognit.* 18, 251–269. doi: 10.3758/BF03213879

Nicol, J., and Swinney, D. (1989). The role of structure in coreference assignment during sentence comprehension. *J. Psycholinguist. Res.* 18, 5–19. doi: 10.1007/BF01069043

Parker, D., Lago, S., and Phillips, C. (2012). "Retrieval interference in the resolution of Anaphoric PRO,"in *Talk, 35th GLOW Conference* (Potsdam).

Patil, U., Vasishth, S., and Lewis, R. (2012). *Retrieval interference in syntactic processing: The case of reflexive binding in English*. Manuscript, Potsdam: University of Potsdam.

Pearlmutter, N. J., Garnsey, S. M., and Bock, K. (1999). Agreement processes in sentence comprehension. *J. Mem. Lang.* 41, 427–456. doi: 10.1006/jmla.1999.2653

Reinhart, T. (1983). *Anaphora and Semantic Interpretation*. London: Croom Helm.

Sturt, P. (2003). The time-course of the application of binding constraints in reference resolution. *J. Mem. Lang.* 48, 542–562. doi: 10.1016/S0749-596X(02)00536-3

Van Dyke, J. A. (2007). Interference effects from grammatically unavailable constituents during sentence processing. *J. Exp. Psychol. Learn. Mem. Cogn.* 33, 407–430. doi: 10.1037/0278-7393.33.2.407

Van Dyke, J. A., and McElree, B. (2006). Retrieval interference in sentence processing. *J. Mem. Lang.* 55, 157–166. doi: 10.1016/j.jml.2006.03.007

Van Dyke, J. A., and McElree, B. (2011). Cue-dependent interference in comprehension. *J. Mem. Lang.* 65, 247–263. doi: 10.1016/j.jml.2011.05.002

Vasishth, S., Brssow, S., Lewis, R., and Drenhaus, H. (2008). Processing Polarity: how the ungrammatical intrudes on the grammatical. *Cogn. Sci.* 32, 685–712. doi: 10.1080/03640210802066865

Wagers, M., Lau, E. F., and Phillips, C. (2009). Agreement attraction in comprehension: representations and processes. *J. Mem. Lang.* 61, 206–237. doi: 10.1016/j.jml.2009.04.002

Xiang, M., Dillon, B., and Phillips, C. (2009). Illusory licensing effects across dependency types: ERP evidence. *Brain Lang.* 108, 40–55. doi: 10.1016/j.bandl.2008.10.002

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 22 April 2014; accepted: 15 October 2014; published online: 05 November 2014.
*Citation: Kush D and Phillips C (2014) Local anaphor licensing in an SOV language: implications for retrieval strategies. Front. Psychol. **5**:1252. doi: 10.3389/fpsyg.2014.01252*

This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology.
Copyright © 2014 Kush and Phillips. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 09 June 2015
doi: 10.3389/fpsyg.2015.00506

Teasing apart retrieval and encoding interference in the processing of anaphors

Lena A. Jäger[1], Lena Benz[1], Jens Roeser[2], Brian W. Dillon[3] and Shravan Vasishth[1]*

[1] Department of Linguistics, University of Potsdam, Potsdam, Germany, [2] Department of Psychology, Nottingham Trent University, Nottingham, UK, [3] Department of Linguistics, University of Massachusetts, Amherst, MA, USA

Two classes of account have been proposed to explain the memory processes subserving the processing of reflexive-antecedent dependencies. Structure-based accounts assume that the retrieval of the antecedent is guided by syntactic tree-configurational information without considering other kinds of information such as gender marking in the case of English reflexives. By contrast, unconstrained cue-based retrieval assumes that all available information is used for retrieving the antecedent. Similarity-based interference effects from structurally illicit distractors which match a non-structural retrieval cue have been interpreted as evidence favoring the unconstrained cue-based retrieval account since cue-based retrieval interference from structurally illicit distractors is incompatible with the structure-based account. However, it has been argued that the observed effects do not necessarily reflect interference occurring at the moment of retrieval but might equally well be accounted for by interference occurring already at the stage of encoding or maintaining the antecedent in memory, in which case they cannot be taken as evidence against the structure-based account. We present three experiments (self-paced reading and eye-tracking) on German reflexives and Swedish reflexive and pronominal possessives in which we pit the predictions of encoding interference and cue-based retrieval interference against each other. We could not find any indication that encoding interference affects the processing ease of the reflexive-antecedent dependency formation. Thus, there is no evidence that encoding interference might be the explanation for the interference effects observed in previous work. We therefore conclude that invoking encoding interference may not be a plausible way to reconcile interference effects with a structure-based account of reflexive processing.

Keywords: anaphors, reflexives, possessives, eye-tracking, German, Swedish, working-memory, interference

OPEN ACCESS

Edited by:
Colin Phillips,
University of Maryland, USA

Reviewed by:
Cristiano Chesi,
Istituto Universitario di Studi Superiori di Pavia, Italy
John E. Drury,
Stony Brook University, USA

****Correspondence:***
Lena A. Jäger,
Department of Linguistics, University of Potsdam, Karl-Liebknecht-Str. 24-25, Potsdam 14476, Germany
lena.jaeger@uni-potsdam.de

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: *12 January 2015*
Paper pending published: *13 March 2015*
Accepted: *09 April 2015*
Published: *09 June 2015*

Citation:
Jäger LA, Benz L, Roeser J, Dillon BW and Vasishth S (2015) Teasing apart retrieval and encoding interference in the processing of anaphors. Front. Psychol. 6:506. doi: 10.3389/fpsyg.2015.00506

1. Introduction

A central task the human sentence processing mechanism has to accomplish is to link two parts of a syntactic dependency, irrespective of how much linguistic material separates the two dependents. Many theories of sentence processing therefore assume that upon encountering the second dependent, the parser triggers a memory retrieval to access the first dependent in order to integrate it with the current node (Gibson, 2000; Lewis and Vasishth, 2005). Interference effects have recently come into focus in sentence processing research because they are taken to be

informative about the more precise nature of the retrieval mechanisms that subserve sentence processing. However, the relationship between empirically observed similarity-based interference effects and theories of retrieval is somewhat indirect, because there are multiple distinct mechanisms that could give rise to similarity-based interference effects in online processing. Indeed, whether or not the observation of interference effects can be interpreted as evidence favoring one or another account of sentence processing depends on the exact mechanisms causing the interference effects. In this article, we will present different mechanisms that have been proposed to account for interference effects in sentence comprehension and present three experiments with different methodologies and languages to tease them apart. We will first give an overview of two kinds of mechanisms, *cue-based retrieval interference* and *encoding interference*, which in the working memory literature have been proposed to underly similarity-based interference. Subsequently, we will turn to the implications for sentence processing and antecedent-retrieval in the processing of reflexives in particular.

Similarity-based interference has long been known to be a major cause of forgetting (Anderson and Neely, 1996). In memory models which represent items as bundles or vectors of features, similarity-based interference is assumed to arise as a function of the degree of overlap between an item's features with the features of other items in memory (Nairne, 1988, 1990; Anderson and Neely, 1996; Anderson and Lebiere, 1998; Anderson et al., 2004; Oberauer and Kliegl, 2006; Lewandowsky et al., 2008). However, the various memory models differ with respect to the mechanisms which they assume to underlie similarity-based interference. Generally speaking, one can distinguish between two kinds of similarity-based interference. On the one hand, similarity-based interference is assumed to affect the encoding or maintenance of an item (Nairne, 1988, 1990; Oberauer and Kliegl, 2006; Lewandowsky et al., 2008). We will refer to this proposal as *encoding interference*. On the other hand, similarity-based interference is assumed to arise during the retrieval of an item (Anderson and Neely, 1996; Anderson and Lebiere, 1998; Anderson et al., 1998, 2004; McElree, 2006; Oberauer and Kliegl, 2006). We will refer to this second proposal as *cue-based retrieval interference*.

Encoding interference is assumed to arise from the competition between the features of similar items that occurs at the moment of encoding or maintaining items in memory. Nairne (1990), for instance, proposed that whenever two items share a feature, they compete for this feature. In a certain proportion of cases, the memory representation of one of these items therefore loses this feature.[1] Hence, this item's memory representation becomes less distinct from other items and, as a result, retrieval probability decreases. An important, but subtle, point here is that even though encoding interference arises at the stage of encoding or maintaining an item in memory, it has an impact on the ease of this item's later *retrieval*. Oberauer and Kliegl (2006), who adopted Nairne (1990)'s concept of feature-overwriting, implemented the idea of an item's memory representation being degraded by decreasing this item's activation level. At the moment of later retrieval, this lower activation level leads to lower retrieval probability and a slow-down in processing times. In their model, the retrieval of an item from working memory is implemented as its gradual activation into the focus layer of the memory system. The processing speed of this gradual activation is defined as a function of this item's activation level prior to retrieval. Thus, if an item's activation level is decreased due to encoding interference from competitor items, a slow-down in the retrieval process is predicted. Note that Oberauer and Kliegl (2006) do not make any predictions about retrieval latencies. Their model is designed to explain data collected in speed-accuracy tradeoff experiments, where they experimentally controlled the time point when retrieval was supposed to happen. In their model, the slow-down in the retrieval process therefore is reflected in a higher proportion of retrieval failures rather than in increased retrieval latencies because participants are forced to interrupt the retrieval process after an experimentally defined time lag. Translating the Oberauer and Kliegl (2006) model to sentence processing, where the participant has more time to carry out retrieval, leads us to the assumption that the slow-down in the retrieval process is reflected in longer retrieval latencies. For the predictions of the experiments reported in this article, we will refer to encoding interference as implemented in the Oberauer and Kliegl (2006) model, with the additional assumption that a slow-down in the retrieval process leads to increased retrieval latencies. In sum, although encoding interference acts at the moment of encoding and maintenance rather than at retrieval, it indirectly affects the success and the speed of the retrieval process because it results in a representation that is more difficult to access.

Cue-based retrieval interference, in contrast, is assumed to arise due to cue-overload at the moment of retrieval. In a content-addressable memory architecture, cue-overload refers to a scenario when the cues used for retrieval do not point to a unique target, but rather match multiple items (Watkins and Watkins, 1975). This is assumed to lead to misretrievals of partially matching distractor items (Anderson and Lebiere, 1998; Anderson et al., 2004; McElree, 2006) and mutual inhibition between the distractors and the target resulting in a higher retrieval latency in case the target and the distractor have one or more retrieval relevant features in common (Anderson and Lebiere, 1998; Anderson et al., 2004).[2] To summarize, encoding interference is predicted to occur whenever items share features, no matter whether these features are used for retrieval or not. Cue-based retrieval interference, in contrast, is predicted to occur when more than one item matches the retrieval features. Inhibition between these items occurs only when they match the same retrieval features, otherwise cue-based retrieval interference

[1] Nairne (1990) did not use the term *encoding interference* but rather *feature-overwriting* to refer to his conception of interference.

[2] Note that the model proposed by McElree (2006) predicts that cue-based retrieval interference is reflected only in retrieval probability, not in retrieval latency. In contrast, the ACT-R architecture developed by Anderson and Lebiere (1998) and Anderson et al. (2004), on which the Lewis and Vasishth (2005) model of sentence processing is based, predicts retrieval interference to be reflected in both, retrieval probability and retrieval latency.

is reflected only in misretrievals (Anderson et al., 2004). Note that encoding interference and cue-based retrieval interference are not mutually exclusive concepts. Indeed, in Oberauer and Kliegl (2006)'s working memory model, both retrieval and encoding interference are assumed and the authors show that their interference model is indeed able to account for a large range of data.

In sentence processing research, early studies investigating interference effects point rather toward encoding than cue-based retrieval interference, but they were not designed to disentangle the two. For example, Gordon et al. (2002) conducted a self-paced reading experiment where participants held a set of nouns in memory while reading the target sentence. The authors report a slow-down in reading times when the noun type (common noun vs. proper name) of the memory load words matched the nouns in the sentence compared to when the memory load nouns and the nouns in the sentence were of different types. These results are further supported by Fedorenko et al. (2006), who also observed similarity-based interference in a memory-load paradigm. Gordon and colleagues report similar results for studies that manipulated similarity between sentence internal nouns rather than memory load (Gordon et al., 2001, 2004, 2006). An example item taken from Gordon et al. (2006) is shown in (1).

(1) **Interference/No interference**
The banker *that* ***the barber/Sophie*** *praised climbed the mountain . . .*

Since in all of these studies, similarity of the nouns was manipulated while the efficiency of the retrieval cues (i.e., the degree to which the retrieval cues uniquely identify the target) remained constant across experimental conditions, the data reported by Gordon and colleagues favor rather encoding than cue-based retrieval interference as an explanation. However, as Van Dyke and McElree (2006) noted, the above cited studies found interference effects only in the region where the critical noun phrase was retrieved (i.e., at the region containing the verb). This might indicate that the observed effect should rather be attributed to cue-based retrieval interference since encoding interference should also affect processing ease at the moment of encoding, i.e., at the moment when the second of the similar nouns is first being encountered. Van Dyke and McElree (2006) conducted a memory load experiment where, in contrast to the memory load experiments reported by Gordon et al. (2002) and Fedorenko et al. (2006), the memory load words were held constant across experimental conditions, but the retrieval cues at the verb were manipulated. The experimental items consisted of object-cleft sentences in which the main clause object preceded the main clause verb (the critical region where retrieval was triggered); for an example taken from Van Dyke and McElree (2006) see (2).

(2) **Interference/No interference**
It was the boat that the guy who lived by the sea ***sailed/ fixed*** *in two sunny days.*

Memory load: *table*, *sink*, *truck*

When the memory load words fit the semantic constraints of the verb, a slow-down in self-paced reading times was observed. These results cannot be attributed to encoding interference since the degree of similarity between the memory load words and the verb's object NP is constant across conditions. Van Dyke and McElree (2006)'s data are thus clear evidence for cue-based retrieval interference playing a role in sentence processing. However, note that the possibility that both retrieval and encoding interference affect sentence processing ease cannot be excluded by Van Dyke and McElree (2006)'s study since their data is clear evidence for cue-based retrieval interference but no evidence *against* encoding interference affecting sentence processing in general.

In recent years, interference effects in the processing of reflexive-antecedent dependencies have drawn considerable attention. The underlying research question was whether unconstrained cue-based retrieval, as proposed by Badecker and Straub (2002) and Patil, Vasishth, and Lewis (unpublished manuscript), or a structure-based access mechanism, as proposed by Nicol and Swinney (1989) and Sturt (2003), subserves the processing of reflexive-antecedent dependencies. Unconstrained cue-based retrieval assumes that all available cues are used to retrieve a reflexive's antecedent. Structure-based accounts, in contrast, assume that structural, i.e., syntactic tree-configurational, information guides the retrieval process. Interference effects in reflexive processing have been generally interpreted in terms of cue-based retrieval interference and taken as evidence for a cue-based retrieval mechanism since *retrieval* interference from syntactically inaccessible constituents is incompatible with the structure-based account. However, as pointed out by Dillon (2011) and Dillon et al. (2013), many of the observed effects—which we will describe more in detail below—can equally well be accounted for by *encoding* interference and hence are not necessarily incompatible with the structure-based account. Indeed, for the kind of materials commonly used to investigate the processing of reflexives (see 3; example taken from Sturt, 2003), encoding interference makes the same predictions for all experimental conditions as the unconstrained cue-based retrieval account implemented in the Lewis and Vasishth (2005) sentence processing model, which is based on the general cue-based architecture Adaptive Control of Thought-Rational (ACT-R) (Anderson and Lebiere, 1998; Anderson et al., 2004) and has been widely used for modeling the processing of reflexives (Dillon, 2011; Dillon et al., 2013; Parker and Phillips, 2014; Kush and Phillips, 2014; Jäger et al., 2015; Patil et al., unpublished manuscript).[3] Thus, for the question of structure-based vs. unconstrained cue-based retrieval in reflexives, it is crucial to disentangle encoding from cue-based retrieval interference. If evidence can be found showing that encoding interference plays a role in the type of materials generally used to investigate the processing of reflexives, this implies that the interference effects that have

[3]The Lewis and Vasishth (2005) model *per se* does not make any commitments with respect to the question which features are used as retrieval cues. Hence it is also possible to implement the structure-based account in this framework by restricting the set of retrieval cues to structural features.

been interpreted as evidence favoring unconstrained cue-based retrieval are equally well compatible with a structure-based account.

(3)

a. **Antecedent-match; distractor-match**
The ***surgeon***$_i$ *who treated* ***Jonathan***$_j$ *had pricked* **himself**$_{i/*j}$ …

b. **Antecedent-match; distractor-mismatch**
The ***surgeon***$_i$ *who treated* ***Jennifer***$_j$ *had pricked* ***himself***$_{i/*j}$ …

c. **Antecedent-mismatch; distractor-match**
The ***surgeon***$_i$ *who treated* ***Jennifer***$_j$ *had pricked* ***herself***$_{i/*j}$ …

d. **Antecedent-mismatch; distractor-mismatch**
The ***surgeon***$_i$ *who treated* ***Jonathan***$_j$ *had pricked* ***herself***$_{i/*j}$ …

Studies investigating interference effects in the processing of reflexives mostly tested sentences in which the reflexive was bound by the local subject which c-commanded the reflexive (*surgeon* in Example 3; henceforth referred to as *antecedent*). We will express the antecedent's conformance to the structural requirements for binding the reflexive by attributing the feature {c-com:+} to it.[4] The interference manipulation was achieved by inserting another noun phrase in a structurally inaccessible position, i.e., not c-commanding the reflexive ({c-com:-}) and hence not qualifying as a binder for the reflexive (*Jonathan/Jennifer* in Example 3; henceforth referred to as *distractor*). A non-structural feature (e.g., gender or number in English reflexives) of this distractor was manipulated. Crucially, the feature which was manipulated might theoretically be used as a retrieval cue. For example, in the processing of English reflexives, the gender feature {gender:masc/fem} marked at the reflexive *himself* or *herself* might be used as a cue to retrieve the antecedent. Thus, if gender is used as a retrieval cue, a gender-matching distractor is predicted to cause cue-based retrieval interference as compared to a distractor which does not match the gender of the reflexive. Therefore, interference effects caused by a feature-matching distractor can be interpreted as evidence favoring an unconstrained cue-based retrieval account. If, in contrast, no effect of a feature-matching distractor is observed, this can be taken as evidence for a structure-based account. This experimental design (or a variation thereof) was used by a large number of studies which aimed to decide whether an unconstrained cue-based retrieval or a structure-based access underlies the processing of reflexive antecedent-dependencies (Nicol and Swinney, 1989; Badecker and Straub, 2002; Sturt, 2003; Xiang et al., 2009; Chen et al., 2012; King et al., 2012; Cunnings and Felser, 2013; Dillon et al., 2013; Clackson and Heyer, 2014; Kush and Phillips, 2014; Parker and Phillips, 2014; Jäger et al., 2015; Patil et al., unpublished manuscript). Some of the cited studies also manipulated feature-match of the structurally accessible antecedent (*surgeon* in Example 3).[5] An effect of antecedent match/mismatch can be accounted for by both unconstrained cue-based retrieval and structure-based accounts.

The results of the above cited studies are mixed. In antecedent-match conditions, increased processing difficulty due to the presence of a cue-matching distractor has been reported by Badecker and Straub (2002), Experiments 3, 4, Chen et al. (2012), Clackson and Heyer (2014), Jäger et al. (2015), Experiment 2, and Patil et al. (unpublished manuscript). By contrast, Sturt (2003), Experiment 1, and Cunnings and Felser (2013), Experiment 2 found a facilitation due to a cue-matching distractor. It should be noted that in Sturt (2003)'s experiment, the effect appeared only delayed and in Cunnings and Felser (2013)'s study, the interference effect was only observed in participants with low working-memory span. Importantly, in a large number of studies, no interference effect in antecedent-match conditions was observed (Nicol and Swinney, 1989; Badecker and Straub, 2002, Experiments 5, 6; Sturt, 2003, Experiment 2; King et al., 2012; Dillon et al., 2013; Kush and Phillips, 2014; Parker and Phillips, 2014; Jäger et al., 2015, Experiment 1). In antecedent-mismatch conditions, a significant processing speed-up due to a cue-matching distractor is reported by King et al. (2012) and Parker and Phillips (2014). The opposite direction of the effect was only observed in Jäger et al. (2015), Experiment 1. The absence of an effect in antecedent-mismatch conditions is reported by Sturt (2003), Xiang et al. (2009) and Dillon et al. (2013). For a literature review of interference effects in reflexives, see Jäger et al. (2015).

As mentioned above, unconstrained cue-based retrieval as implemented in the Lewis and Vasishth (2005) ACT-R model of sentence processing makes precisely the same predictions as encoding interference for sentences like the ones shown in (3). For conditions with a cue-matching antecedent (see 3a,b), the Lewis and Vasishth (2005) model predicts cue-based retrieval interference when the distractor matches the gender of the reflexive (3a). This retrieval interference is predicted to be reflected in inhibition between the antecedent and the distractor because in (3a), but not in (3b), the antecedent (*surgeon*) and the distractor (*Jonathan*) share the gender cue {gender:masc}. Thus, longer retrieval latencies (and hence longer reading times at the reflexive) are predicted in (3a) compared to (3b). Moreover, misretrievals of the partially cue-matching distractor (*Jonathan* in 3a) are predicted. These misretrievals are predicted to be reflected in response-accuracies if the comprehension questions target the reflexive-antecedent dependency. For conditions with a mismatching antecedent (see 3c, d), the unconstrained cue-based retrieval model (Lewis and Vasishth, 2005) also predicts cue-based retrieval interference due to a cue-matching distractor (3c). As in antecedent-match conditions, this retrieval interference is

[4]It should be noted that using {c-com:+} as a feature is a simplification since a tree-configurational relation is not as straightforward to code as a feature of an item as, e.g., gender or number. For a discussion of how tree-configurational information such as c-command could be encoded as an item's feature see Alcocer and Phillips (unpublished manuscript). On a theoretical basis, Kush (2013) argues against the representation of c-command as a feature and discusses how, in online sentence processing, the human parser might distinguish between c-commanding and non-c-commanding antecedents.

[5]In some experiments, only the stereotypical gender of the accessible antecedent was violated (as in 3c,d), whereas in other studies, real feature violations were used resulting in ungrammatical sentences in the antecedent-mismatch conditions.

predicted to be reflected in a higher proportion of misretrievals of the matching distractor. But, in contrast to antecedent-match conditions, no inhibition between the antecedent and the distractor is predicted because they do not share any of the experimentally manipulated retrieval relevant features (in 3c and d, the antecedent and the distractor neither share the gender cue {gender:fem} nor the structural cue {c-com:+}). Since ACT-R predicts faster retrieval latencies in the case of misretrievals as a result of a race-like configuration, the trials with misretrievals are predicted to lead to a decreased mean retrieval latency. Therefore, in the absence of inhibition between the distractor and the antecedent in antecedent-mismatch conditions, faster processing times are predicted when a feature-matching distractor is present.

Encoding interference predicts increased retrieval latencies and a higher proportion of misretrievals as a function of the number of features the target (here the antecedent) shares with other items in memory (Oberauer and Kliegl, 2006).[6] Thus, in conditions with a matching antecedent (see 3a,b), a slow-down and a higher proportion of misretrievals due to a feature-matching distractor (3a) is expected. By contrast, in conditions with a mismatching antecedent (see 3c,d), a slow-down and a higher proportion of misretrievals due to a *feature-mismatching* distractor (3d) is predicted since the mismatching antecedent and the mismatching distractor have the same gender feature {gender:masc}.[7]

To summarize, for materials as the ones presented in (3), both encoding interference and cue-based retrieval interference predict that a matching distractor leads to a processing slow-down in antecedent-match conditions and to a speed-up in antecedent-mismatch conditions. For online reading time measures, both accounts thus make precisely the same predictions and can account for the inhibitory effects in antecedent-match conditions reported by Badecker and Straub (2002), Chen et al. (2012), Clackson and Heyer (2014), Jäger et al. (2015) and Patil et al. (unpublished manuscript) as well as for the facilitatory effects in antecedent-mismatch conditions reported by King et al. (2012) and Parker and Phillips (2014). For retrieval probabilities (to be reflected in response accuracies of adequate comprehension questions), both accounts also make the same predictions for antecedent-match conditions but differ in their predictions for antecedent-mismatch conditions. Hence, if online evidence for encoding interference in reflexives can be found, we need to reconsider the theoretical implications of interference effects in reflexives with respect to the debate about structurally-guided vs. unconstrained cue-based retrieval. In the following, we present two experiments on German and one experiment on Swedish designed to disentangle encoding from cue-based retrieval interference.

[6]To be precise, the number of distractors sharing a certain feature with the target also affects retrieval latencies and retrieval probability because the more distractors share this feature with the target, the higher the probability that one of these distractors "robs" this feature from the memory representation of the target.

[7]Because we set out to determine whether invoking encoding interference is a way to reconcile interference effects with structure-based retrieval, for the predictions of encoding interference we are assuming that only structural retrieval cues are used. If, by contrast, one assumes that gender is used as a retrieval cue, the feature matching distractor (3c) is predicted to be misretrieved more often than the feature mismatching distractor (3d). This prediction is orthogonal to the question of encoding interference, but follows from the basic assumption that an item's retrieval probability depends on its features' match with the retrieval cues. This basic assumption is shared by models of encoding interference (Nairne, 1990; Oberauer and Kliegl, 2006). (Note that this point is unrelated to the cue-based retrieval interference component in the Oberauer and Kliegl, 2006 model which is assumed to cause inhibition between items sharing the same retrieval cues.)

2. Experiment 1: German Reflexives (Self-Paced Reading)

The German reflexive *sich* "himself"/"herself" is an interesting test case for teasing apart encoding from cue-based retrieval interference. The third-person singular reflexive *sich* is gender neutral and, roughly speaking, requires its antecedent to be a c-commanding noun phrase contained in the reflexive's local clause. For more details about the syntactic properties of German reflexives see Everaert (1986), Reinhart and Reuland (1993), Reuland and Reinhart (1995), Reuland (2001), Gast and Haas (2008) and Reuland (2011). Since *sich* is gender neutral and thus gender can be assumed to not be used as a retrieval cue, we do not expect any cue-based retrieval interference from a structurally inaccessible distractor that shares its gender with the antecedent. Encoding interference, in contrast, predicts that a distractor of the same gender as the antecedent leads to a degradation of the antecedent's memory representation resulting in longer processing times when retrieving the antecedent upon encountering the reflexive. Moreover, encoding interference predicts a lower retrieval probability of the antecedent when a gender-sharing distractor is present. We will use the term *gender-overlap* to refer to the situation where the antecedent and the distractor share their gender in order to reserve the term *gender-match* for the match of an item's feature with a retrieval cue as in Example (3) discussed above.

2.1. Materials and Method

2.1.1. Materials

The experimental items consist of a matrix clause whose subject is the antecedent of the third person singular reflexive *sich* (see 4 for an example). The reflexive is the first constituent of a conjoint determiner phrase (*sich und die Kollegen* in 4) which as a whole is the direct object of the matrix verb. The antecedent (*der Dieb/die Diebin* in 4) is modified by an object-extracted relative clause that intervenes between the antecedent and the reflexive. The subject of this relative clause (*der Hehler/die Hehlerin* in 4) does not c-command the reflexive and hence syntactically disqualifies as antecedent. We will refer to this noun phrase as *distractor*. Both the antecedent and the distractor were always animate common nouns with a definite article. King et al. (2012) have shown that interference effects in reflexives are more likely to be detected when the verb, which triggers the retrieval of its subject—which, in turn, is also the reflexive's antecedent—does not directly precede the reflexive. In order to increase the chances of detecting an effect, we chose perfective tense for our materials, because, as opposed to present tense or simple past,

the reflexive precedes the main verb in perfective sentences (for another study on interference effects using pre-verbal reflexives see Kush and Phillips, 2014). Moreover, we inserted a relatively long adverb between the perfective auxiliary *hat* and the reflexive. As in the classical gender-match/mismatch design, we manipulated the antecedent's and the distractor's gender. This resulted in a fully crossed 2 × 2 design with gender of the antecedent (masculine vs. feminine) and interference (gender-overlap vs. no gender-overlap between the distractor and the antecedent) as factors. For our research question, the gender manipulation of the antecedent was not of interest *per se*. It was included in order to experimentally control for lexical properties such as word length or frequency which, due to the nature of the German language, are inseparable from the gender manipulation. We will discuss this issue more in detail in the Results section.

(4) **Masc/Fem antecedent; Masc/Fem distractor**

Der Dieb$_i$***/Die Diebin***$_i$,	*dem/der*	***der Hehler***$_j$***/die Hehlerin***$_j$		*befohlen*	*hat*	*zu*	*stehlen,*	*hat*		
the thief-MASC/the thief-FEM	whom	**the dealer-MASC/the dealer-FEM**		obliged	has	to	steal	has		
überraschenderweise	***sich***$_{i/*j}$	*und*	*die*	*Kollegen*	*angezeigt,*	*berichtete*	*das*	*Hochglanzmagazin.*		
surprisingly	**self**	and	the	colleagues	denounced	reported	the	magazine		

'The thief whom the dealer obliged to steal surprisingly denounced himself/herself and the colleagues, reported the magazine.'

Each sentence was followed by a *yes/no* comprehension question targeting the reflexive-antecedent dependency. One half of the comprehension questions tested whether the antecedent was retrieved successfully (to be answered with *yes*) and the other half tested whether the distractor was misretrieved instead (to be answered with *no*). Question types were balanced across items and held constant within the four conditions of each item.

2.1.2. Participants and Procedure

144 undergraduate students from the University of Potsdam who were all native speakers of German participated in the study for credit or payment of 5 EUR. We chose a relatively large sample size in order to increase statistical power, i.e., reduce Type II error probability. For our research question, high statistical power is particularly important since if encoding interference in the processing of reflexives is absent, a null result is predicted. The number of participants was determined based on a statistical power test assuming an effect of 20 ms and a standard deviation of 75 ms. In order to achieve power of 90%, 149 participants would be needed. Due to the restricted nature of our participant pool, the actual sample size was slightly smaller, which yielded a statistical power of 0.89%. 16 test items and 32 filler sentences were presented in a moving-window self-paced reading paradigm (Just et al., 1982). Items were arranged according to a Latin Square with a different randomization for each participant. Each trial was followed by a *yes/no* comprehension question.

2.2. Results

Statistical analyses were carried out in GNU-R (R Development Core Team, 2011) using linear mixed effects models provided by the lme4 package version 1.0-6 (Bates et al., 2014). Binary dependent variables were modeled using generalized linear mixed models with a logistic link function. For the analyses of comprehension questions and reading times, we fit models testing for a main effect of gender of the antecedent, a main effect of interference (i.e., effect of whether or not the distractor overlapped in gender with the antecedent) and an interaction between the two. All models were fit with random intercepts and slopes for participants and items (Baayen et al., 2008). No correlations between random effects were estimated since in many of the models the correlation matrix of random effects was degenerate.

In German, the feminine form of a noun is usually generated by adding the suffix *-in* and in many nouns, the masculine form is more frequent than the feminine one. Therefore, a correlation between gender and word length and word frequency could not be avoided in the stimuli. More precisely, correlations between the main effect of gender and frequency/length of the antecedent as well as correlations between the interaction antecedent gender × interference and frequency/length of the distractor are expected. Crucially, including the gender manipulation of the antecedent as a fully crossed within-items factor in our design ensured a zero correlation between frequency/length of the antecedent or the distractor with the critical main effect of interference. Along the same lines, correlations between frequency/length of the antecedent and the interaction antecedent gender × interference as well as correlations between frequency/length of the distractor and the main effect of gender of the antecedent cancel out due to the fully-crossed factorial design. To test these assumptions and to obtain estimates for the expected correlations, we computed Pearson-correlations of each of the contrasts to be tested in the linear-mixed model with centered word lengths measured in number of characters and centered log-transformed lemma frequencies taken from dlexDB[8] (Heister et al., 2011) of the antecedent and the distractor (see **Table 1**). As expected, there was a positive correlation ($r = 0.63$) between the main effect of gender of the antecedent and frequency of the antecedent and a negative correlation ($r = -0.44$) between the main effect of gender of the antecedent and word length of the antecedent.

[8]www.dlexdb.de

TABLE 1 | Experiments 1 and 2.

	Length antecedent	Frequency antecedent	Length distractor	Frequency distractor
Interference	0	0	0	0
Gender antecedent	−0.44	0.63	0	0
Gender ant. × Interf.	0	0	−0.24	0.39

Correlations (Pearson correlation coefficient) of word length and log lemma frequency of the antecedent and the distractor with the experimental manipulations (main effect of interference, main effect of gender of the antecedent and their interaction). Word length and log lemma frequencies were centered (z-scores).

TABLE 2 | Experiment 1.

Condition	Accuracy
Gender-overlap - masculine antecedent	71
Gender-overlap - feminine antecedent	73
No gender-overlap - masculine antecedent	75
No gender-overlap - feminine antecedent	77

Mean accuracy scores of question responses in percentage by experimental condition.

Similarly, there was a positive correlation ($r = 0.39$) between the frequency of the distractor and the interaction antecedent gender × interference and a small negative correlation between word length of the distractor and the interaction antecedent gender × interference. Thus, a main effect of gender of the antecedent and the interaction between the two main effects should not be interpreted since they might be confounded with the effects of antecedent/distractor length and frequency.

2.2.1. Comprehension Questions

Comprehension question response accuracies were analyzed using a linear mixed model with a logistic link function. Mean accuracy scores of question responses are provided in **Table 2**. Statistical analyses revealed a main effect of interference: accuracy was lower in conditions with a gender-sharing distractor (estimate $= -0.25$, $SE = 0.12$, $z = -2.02$, $p < 0.05$). Neither the main effect of gender nor the interaction were significant.

2.2.2. Reading Times

An overview of raw reading times for each region of the sentence is provided in **Table A1** in the Appendix. Reading times were analyzed at the reflexive, the following NP together with the preceding conjunction *und* "and" (n+1), the main clause verb (n+2) as well as at the two words preceding the reflexive as a sanity test of the baseline reading times. In order to achieve a close to normal distribution of the model residuals, we analyzed negative reciprocal reading times (Box and Cox, 1964). None of the comparisons reached significance at any region. Modeling log-transformed RTs instead of reciprocal RTs yielded similar results. The output of the linear-mixed models is provided in **Table 3**.

2.3. Discussion

In reading times, we did not find any effect of gender-overlap between the antecedent and the distractor. However, in comprehension questions, we observed lower response accuracies when the distractor overlapped in gender with the antecedent. This effect might be explained by misretrievals due to encoding interference during online processing which, critically, did not affect processing times. Alternatively, the lower response accuracies in the gender-overlap conditions might reflect an offline effect that arises at the moment of answering the comprehension question. Crucially for our research question, we could not find any evidence supporting the idea that encoding interference affects online processing times at the reflexive. With respect to previous studies on reflexives, we can therefore conclude that there is no indication that the interference effects observed in previous studies reflect encoding rather than cue-based retrieval interference.

3. Experiment 2: German Reflexives (Eye-Tracking)

Experiment 2 is a cross-methodological replication of Experiment 1. Already Ronald Fisher, the father of frequentist statistics, emphasized the importance of replication (Fisher, 1937, page 16). Indeed, non-replicable findings are a major problem in experimental psychology and psycholinguistics (Simmons et al., 2011; Asendorpf et al., 2013). Moreover, a potential concern about Experiment 1 is that our conclusions are based on a null result. Although we have addressed this issue by testing a large sample and thus gaining high statistical power, one could still argue that the self-paced reading method is not sensitive enough to detect a potential effect. We therefore tested the same materials as in Experiment 1 in an eye-tracking while reading paradigm, which presumably is a more sensitive method compared to self-paced reading (Staub and Rayner, 2007).

3.1. Materials and Method

3.1.1. Materials

The same stimuli (including fillers) were used as in Experiment 1.

3.1.2. Participants and Procedure

151 undergraduate students from the University of Potsdam with normal or corrected-to-normal vision who were all native speakers of German participated in the experiment against credit or payment of 7 EUR. None of the participants had participated in Experiment 1.

Participants' eye movements (right eye monocular tracking) were recorded with an SR Research Eyelink 1000 eyetracker at a sampling rate of 1000 Hz using a desktop mount camera system with a 35 mm lens. The participant was seated at a height-adjustable table with his/her head stabilized using a forehead/chin-rest. Stimuli were presented on a 22 inch monitor (resolution of 1680 × 1050 pixels) with an eye-to-screen distance of 62 cm and an eye-to-camera distance of 60 cm. As a response pad, a Microsoft Button Box was used. Stimuli were presented using Experiment Builder software provided by SR Research. The experimental items were presented on a light gray background

TABLE 3 | Experiment 1.

	n−1			REFL			n+1			n+2		
Predictor	***coef***	***SE***	***t***	***coef***	***SE***	***t***	***coef***	***SE***	***t***	***coef***	***SE***	***t***
Interference	4e-05	3e-05	1.39	2e-05	2e-05	0.67	0	1e-05	0.18	−1e-05	3e-05	−0.58
Gender antecedent	0	2e-05	−0.08	1e-05	3e-05	0.29	0	1e-05	−0.42	−4e-05	3e-05	−1.25
Gender ant. × Interf.	−4e-05	3e-05	−1.49	−3e-05	2e-05	−1.49	−1e-05	1e-05	−0.96	−1e-05	3e-05	−0.45

Main effects of interference and gender of the antecedent and their interaction on negative reciprocal RTs as dependent variable measured at the adverb preceding the reflexive (n−1), the reflexive (REFL), the coordinate NP following the reflexive (n+1) and the main clause verb (n+2).

TABLE 4 | Experiment 2.

Condition	**Accuracy**
Gender-overlap - masculine antecedent	74
Gender-overlap - feminine antecedent	69
No gender-overlap - masculine antecedent	75
No gender-overlap - feminine antecedent	75

Mean accuracy scores of question responses in percentage by experimental condition.

in black font, font type Times New Roman, font size 14. They were arranged according to a Latin Square and were pseudo-randomized for each participant separately such that every experimental trial was preceded by at least one filler sentence. A nine-point calibration was carried out at the beginning of the experiment and repeated during the experiment, if needed. Each experimental session started with 6 practice trials. At the beginning of each trial, participants had to fixate a drift correction point at the left center of the screen where the first word of the sentence was to appear.

3.2. Results

Linear mixed-effects models were fit with the same predictors as for Experiment 1. As in the analysis of Experiment 1, all models were fit with varying intercepts and slopes for participants and items. No correlations between random effects were estimated since, as in the data of Experiment 1, the correlation matrix of random effects was degenerate in many of the models.

3.2.1. Comprehension Questions

Mean accuracy scores by experimental condition are provided in **Table 4**. We observed a marginal main effect of interference with lower accuracies in conditions where antecedent and distractor had the same gender (estimate $= -0.20$, $SE = 0.10$, $z = -1.95$, $p = 0.05$). This replicates the pattern found in Experiment 1. None of the other effects was significant.

3.2.2. Eye Movements

An overview of raw reading times at each word of the sentence is provided in **Table A2** in the Appendix. The same regions were analyzed as in Experiment 1. Raw fixation durations shorter than 20 ms or longer than 1000 ms (0.25% of the data) were excluded from all analyses. In eye-tracking data, the dependent measures can be partitioned into first-pass, regression-related (proportions of regressions and duration of regressive events) and later-pass measures. Since the exact mapping between syntactic effects and eye-tracking measures is still unclear (Clifton et al., 2007), we analyzed one representative measure from each group. As a first-pass measure, we analyzed first-pass reading time (FPRT, also referred to as gaze duration), which is defined as the sum of all first-pass fixations on a region. As regression related measures, we analyzed first-pass regression-probability (FPRP), i.e., a binary variable coded as 1 if a first-pass regression was initiated from a region, and regression-path duration (RPD), i.e., the sum of all fixation durations starting from the first fixation on a region until leaving this region to the right including all regressive fixations that fall into this time window. As a later-pass measure, we analyzed total-fixation time (TFT), i.e., the sum of all fixations on a region. Strictly speaking, TFT is not a pure late measure but rather the sum of FPRT and re-reading time. However, we chose to report TFT as a representative late measure since TFT is one of the most commonly reported measures in psycholinguistics; we do not analyze re-reading time because the critical region was re-read in only about 20% of the trials leading to very low statistical power. In order to achieve approximately normally distributed residuals, the continuous dependent variables were log-transformed (Box and Cox, 1964).

An overview of the output of the linear mixed-effects models is provided in **Table 5**. At the reflexive (n), the word preceding the reflexive and the region following the reflexive, none of the comparisons reached significance in any of the dependent variables. At region n+2 (i.e., the main clause verb), a significant effect of gender of the antecedent was observed in TFT (longer fixation times in conditions with a feminine antecedent). However, as we have argued in the Results section of Experiment 1, this effect should not be interpreted since it correlates with frequency and length of the antecedent. For our research question, only the main effect of interference is relevant.

Moreover, a *post-hoc* analysis of the region containing the relative clause verb (*zu stehlen* in Example 4) revealed a significant main effect of interference in TFT with longer fixation durations when the antecedent and the distractor overlapped in gender (estimate $= 0.05$, $SE = 0.02$, $t = 2.28$).

3.3. Discussion

Experiment 2 replicated the findings of Experiment 1. As in Experiment 1, no evidence for encoding interference due to gender-overlap between the reflexive's antecedent and a

TABLE 5 | Experiment 2.

DV	Predictor	n−1			REFL			n+1			n+2		
		coef	*SE*	*t or z*	*coef*	*SE*	*t or z*	*coef*	*SE*	*t or z*	*coef*	*SE*	*t or z*
FPRT	Interference	−0.01	0.02	−0.25	−0.01	0.02	−0.31	−0.01	0.03	−0.38	−0.02	0.01	−1.14
	Gender antecedent	0.01	0.02	0.61	−0.03	0.02	−1.35	−0.03	0.03	−1.03	0.01	0.01	0.64
	Gender ant.×Interf.	−0.02	0.02	−1.06	−0.01	0.02	−0.64	−0.04	0.02	−1.91	0	0.02	0.25
RPD	Interference	0.04	0.03	1.59	0.01	0.02	0.44	−0.02	0.03	−0.67	0	0.03	−0.06
	Gender antecedent	0.02	0.02	0.86	−0.02	0.02	−0.76	−0.02	0.02	−1.01	0.01	0.02	0.55
	Gender ant.×Interf.	−0.03	0.02	−1.14	0.03	0.02	1.1	−0.02	0.02	−1.01	0	0.02	−0.22
FPRP	Interference	0.19	0.11	1.64	0.12	0.14	0.83	−0.05	0.11	−0.44	0.13	0.17	0.77
	Gender antecedent	0.09	0.12	0.72	−0.1	0.14	−0.7	−0.05	0.14	−0.34	−0.01	0.13	−0.05
	Gender ant.×Interf.	0.03	0.11	0.22	0.18	0.18	0.98	0.1	0.11	0.85	−0.01	0.14	−0.09
TFT	Interference	0.03	0.02	1.4	−0.01	0.02	−0.36	0.01	0.02	0.37	0	0.02	0.12
	Gender antecedent	0.03	0.02	1.54	0.02	0.02	0.68	0	0.02	0.13	0.04	0.02	2.22*
	Gender ant.×Interf.	0.01	0.02	0.29	0.01	0.02	0.46	−0.04	0.02	−1.93	−0.01	0.02	−0.72

Main effects of interference and gender of the antecedent and their interaction on the dependent variables log-first-pass reading time, log-regression-path duration, first-pass regression probability and log-total fixation time measured at the adverb preceding the reflexive (n−1), the reflexive (REFL), the coordinate NP following the reflexive (n+1) and the main clause verb (n+2). Statistically significant ($\alpha = 0.05$) effects are marked with an asterisk.

structurally inaccessible distractor was observed neither at the reflexive, nor at the pre- or post-critical regions.

At the relative clause verb, however, gender-overlap between the main clause subject, i.e., the antecedent, and the relative clause subject, i.e., the distractor, led to significantly longer total-fixation times. At this region, the relative clause subject needs to be retrieved. Hence, the observed effect, which appears in a similar region as the effects reported by Gordon et al. (2001), might reflect encoding interference. However, it is disconcerting that this effect was observed only in total-fixation time and was not present in Experiment 1, as a *post-hoc* analysis of the self-paced reading data showed. Thus, one might discount this effect as a possible Type I error. If one does not discount the effect, it raises the question why encoding interference affects argument-head dependency completion, but not reflexive-antecedent dependency formation. A possible explanation might be that the encoding interference effect (to the extent that it is not a Type I error) dies out by the time the reflexive is processed.[9] In any case, further replication attempts of this configuration are needed. In sum, it is possible that we are seeing encoding interference at the distractor, but, which is crucial for our research question, this encoding interference does not seem to have any effect at the reflexive.

Taken together, the results of Experiments 1 and 2 are a strong indication that in reflexive-antecedent dependency formation, the sharing of a non-structural feature such as gender does not lead to encoding interference reflected in a processing slow-down. More precisely, it indicates that in materials of the type used in this experiment, encoding interference does not affect retrieval latencies of the antecedent when processing the reflexive. However, the marginal interference effect in offline comprehension accuracies, which had been significant in Experiment 1, indicates that the antecedent was retrieved less often correctly when it shared its gender with the distractor. This can be interpreted as evidence for encoding interference affecting retrieval probability of the antecedent. In sum, neither experiment provides any evidence for the claim that encoding interference affects reading time at the reflexive. However, our offline results suggest that encoding interference might affect retrieval probability of the antecedent. Crucially, even if encoding interference affected retrieval probability of the antecedent or the offline interpretation of the sentence, there is no evidence that it affects the participants' online behavior at the reflexive measured in self-paced reading times or eye-movements. Hence, encoding interference is not a plausible explanation for the *online* effects previous studies have observed in eye-tracking or self-paced reading measures.

4. Experiment 3: Swedish Possessives (Eye-Tracking)

Experiments 1 and 2 yielded converging results: we found no evidence for encoding interference affecting the online processing speed of German reflexives. However, there are still two potential concerns with these results: (i) Our conclusion is based on two null-results, and (ii) we need to cross-linguistically validate our conclusion. In Experiment 3, we addressed these issues by examining the processing of Swedish possessives in an eye-tracking experiment. In

[9]A reviewer noticed that the effect at the relative clause verb occurs in total-fixation time, a measure which can reflect processing difficulty encountered further downstream in the sentence, and therefore might actually reflect processing difficulties at the reflexive which triggers re-readings of the previous materials. However, if this were the case, one would expect an increase in the proportion of regressions or increased regression-path durations at the reflexive. As this is not the case, it is difficult to conclude that the effect observed at the verb reflects processing difficulty associated with the reflexive.

Swedish, there are two kinds of possessives: reflexive possessives that are not gender-marked and pronominal possessives that need to agree in gender with their antecedent. The reflexive possessive *sin* "his"/"her" can only be bound by a c-commanding antecedent inside its local clause. In contrast, the pronominal possessive *hans* "his" must not be bound within its local clause, but requires an antecedent outside its clause domain (see Holmes and Hinchliffe, 1994 and Kaiser, 2003, p. 209). In a 2 × 2 factorial design, we manipulated anaphor type (pronominal possessive vs. reflexive possessive) and interference, i.e., whether or not a structurally inaccessible distractor shared the gender of the antecedent. For this design, encoding interference predicts increased processing difficulty in the gender-overlap conditions compared to the no-gender-overlap conditions, regardless of anaphor type. Cue-based retrieval interference, in contrast, predicts an interaction between anaphor type and interference: increased processing difficulty due to a gender-sharing distractor is predicted for the gender-marked pronominal possessives but not for the gender-unmarked reflexive possessives. This is because only in pronominal possessives, the gender-marked anaphor can trigger a retrieval process where gender is used as a retrieval cue. When both the antecedent and the distractor match the gender cue, cue-based retrieval interference predicts inhibition between the antecedent and distractor and a higher proportion of misretrievals of the distractor (Lewis and Vasishth, 2005). Thus, the present experiment allows us to directly pit encoding and cue-based retrieval interference against each other. In contrast to Experiments 1 and 2, cue-based retrieval interference predicts an interaction rather than a null-result.

(5) a. **Pronominal possessives; gender-overlap/no gender-overlap**

Åke$_i$	*säger*	*att*	***Alf***$_j$***/Ann***$_j$	*jobbade*	*med*	***hans***$_{i/*j}$	*sysslingar*	*på*	*helgerna.*
Åke-MASC	says	that	**Alf-MASC/Ann-FEM**	worked	with	**his**	cousins	at	the weekend

'Åke says that Alf/Ann worked with his cousins at the weekend.'

b. **Reflexive possessives; gender-overlap/no gender-overlap**

Åke$_i$	*som*	***Alf***$_j$***/Ann***$_j$	*tackade*	*ringer*	***sina***$_{i/*j}$	*sysslingar*	*på*	*kvällen.*
Åke-MASC	whom	**Alf-MASC/Ann-FEM**	thanked	calls	**his**	cousins	in	the evening

'Åke, whom Alf/Ann thanked, calls his cousins in the evening.'

4.1. Materials and Method

4.1.1. Materials

The conditions with pronominal possessives (see 5a for an example item) consist of a superordinate clause whose subject is the antecedent (*Åke* in 5a) and a subordinate clause containing the distractor (*Alf* or *Ann* in 5a) which either matches or mismatches the gender of the antecedent and the gender-marked pronominal possessive (*hans* "his" in 5a). The conditions with reflexive possessives (see 5b for an example item) consist of a main clause containing the antecedent (*Åke* in 5b) and the gender-unmarked reflexive possessive (*sina* "his"/"her" in 5b). The distractor (*Alf* or *Ann* in 5b) is the subject of an appositive relative clause intervening between the antecedent and the reflexive possessive. As Swedish does not code masculine and feminine as grammatical gender, and the number of nouns with inherent gender such as *boy* or *girl* is very limited, both the antecedent and the distractor were proper names in all experimental sentences. Indeed, it is crucial for our research question to extend the findings of Experiments 1 and 2 to proper names, which differ from common nouns with respect to their referential properties (Longobardi, 1994; Elbourne, 2005), since several of the studies reporting interference effects in reflexives actually employed proper names (e.g., Badecker and Straub, 2002).

The nouns used as antecedents and distractors are all highly frequent, gender unambiguous Swedish first names taken from Statistics Sweden, a database which contains the 100 most frequently given and used male and female first names in Sweden.[10] Antecedents and distractors are all matched for word length (numbers of characters) within each item. Half of the items have a feminine antecedent and the other half a masculine antecedent. The possessed noun phrase (*sysslingar* in 5) is always a plural noun.

Two types of comprehension questions were designed. The first type probed for the correct interpretation of the anaphor-antecedent dependency. 50% of these questions were to be answered with *yes*. The second question type targeted various parts of the sentence, but not the interpretation of the anaphor. Again, 50% of these questions were to be answered with *yes*.

4.1.2. Participants and Procedure

35 native speakers of Swedish currently living in Berlin or Potsdam with normal or corrected-to-normal vision participated in the experiment against payment of 5 EUR (plus 6.20 EUR to cover travel expenses). The sample size was smaller compared to Experiments 1 and 2 due to logistic limitations, but we tested a larger number of experimental items compared to Experiments 1 and 2. Participants' eye movements (right eye monocular tracking) were recorded while reading 48 experimental sentences

[10] http://www.scb.se/BE0001-EN; we used the data of 2012.

and 70 filler sentences. The general technical set-up was the same as in Experiment 2. Stimuli were arranged in a Latin Square and pseudo-randomized such that each experimental trial was preceded by at least one filler sentence. Each trial was followed by a comprehension question. Two thirds of the comprehension questions targeted the correct interpretation of the anaphor and one third targeted other parts of the sentence. The experiment started with 5 practice trials to familiarize participants with the procedure.

4.2. Results

On all dependent variables, we fit linear mixed-effects models with main effects of anaphor type (pronominal vs. reflexive possessive), interference (whether or not the distractor had the same gender as the antecedent) and their interaction as predictors. When the interaction reached significance, nested contrasts testing for an interference effect within each anaphor type were fit. All models were fit with varying intercepts for participants and items. No varying slopes were fit because the generalized likelihood-ratio test showed that they did not improve the model fit. The pattern of results was not affected by whether or not varying slopes were fit. For the interpretation of results, it should be kept in mind that the effect of anaphor type is not of theoretical relevance to our research question. As the two levels of anaphor type differ lexically at the pre-critical and the critical region, a main effect of anaphor type does not have any useful interpretation.

4.2.1. Comprehension Questions

Mean accuracy scores by experimental condition and question type (i.e., whether or not the comprehension question targeted the anaphor) are provided in **Table 6**. We ran a linear-mixed effects model with a logistic link function with main effects of anaphor type, interference and question type and their interactions including the three-way interaction between all main effects as predictors. The model output is summarized in **Table 7**. The main effect of interference and the interaction between interference and question type reached significance. Moreover, a marginal three-way interaction between interference, anaphor type and question type was observed. A second model in which we applied nested contrasts testing for an interference effect within each level of anaphor type and question type[11] showed that the interactions were caused by a highly significant interference effect that was present only in questions targeting the anaphor in pronominal possessives (estimate $= -1.16$, $SE = 0.25$, $z = -4.62$, $p < 0.0001$). In sum, in questions targeting the anaphor-antecedent dependency, the presence of a gender matching distractor led to lower response accuracies in sentences with pronominal possessives but not in sentences with reflexive possessives. In questions not targeting the anaphor-antecedent dependency, no effects were observed.

TABLE 6 | Experiment 3.

Condition	Accuracy	
	Anaphor	Other
Pronominal - gender-overlap	75	82
Pronominal - no gender-overlap	90	82
Reflexive - gender-overlap	85	80
Reflexive - no gender-overlap	86	81

Mean accuracy scores of comprehension questions in percentage by experimental condition and question type, i.e., whether the question targeted the anaphor-antecedent dependency or another element of the sentence.

4.2.2. Eye Movements

An overview of raw reading times at each region of the sentence is provided in **Table A3** in the Appendix. We analyzed the pre-critical region containing the verb (plus postposition), the critical region containing the pronominal or reflexive possessive and the post-critical region containing the possessed noun. The same dependent variables were analyzed as in Experiment 2. Continuous dependent variables were log-transformed in order to achieve approximately normally distributed residuals.

An overview of the output of the linear mixed-effects models is provided in **Table 8**. The effect of anaphor type reached significance across regions and dependent variables. However, as mentioned above, this effect was not of interest to our research question: conditions with pronominal and reflexive possessive differ from each other in syntactic structure, distractor position, lexicon, word length and number of words contained in the pre-critical region. At the pre-critical and the critical region, no other effect reached significance in any dependent variable. At the post-critical region, a significant interaction between anaphor type and interference was observed in FPRP. Pairwise comparisons revealed that this interaction was driven by a significant interference effect in pronominal possessives. When the distractor shared the gender of the antecedent and hence matched the gender-cue, less first-pass regressions were observed (estimate $= -0.44$, $SE = 0.18$, $z = -2.47$, $p < 0.05$). In order to test whether this facilitation due to a gender-matching distractor reflected misretrievals of the latter, we re-ran the models on comprehension question response accuracies for trials

TABLE 7 | Experiment 3.

Predictor	*coef*	*SE*	*z*
Interference	−0.17	0.07	−2.42*
Anaphor type	−0.01	0.07	−0.17
Question type	0.12	0.10	1.11
Interference × Anaphor type	−0.12	0.07	−1.61
Interference × Question type	−0.15	0.07	−2.04*
Anaphor type × Question type	−0.06	0.07	−0.86
Interference × Anaphor type × Question type	−0.14	0.07	−1.95

Analysis of comprehension questions: Main effects of interference, anaphor type and question type together with their interactions. Statistically significant ($\alpha = 0.05$) effects are marked with an asterisk.

[11]The model predictors were main effects of anaphor and question type, interaction between anaphor type and question type and the four pairwise comparisons (interference effects in pronominal and reflexive possessives in question targeting the anaphor and questions not targeting the anaphor.)

TABLE 8 | Experiment 3.

DV	Predictor	n−1			REFL/PRON			n+1		
		coef	*SE*	*t or z*	*coef*	*SE*	*t or z*	*coef*	*SE*	*t or z*
FPRT	Interference	0.02	0.02	0.75	0.01	0.02	0.57	−0.03	0.02	−1.29
	Anaphor type	0.03	0.02	1.43	−0.04	0.02	−1.97	0.05	0.02	1.99
	Anaph. type×Interf.	0.03	0.02	1.28	−0.01	0.02	−0.41	0.02	0.02	0.71
RPD	Interference	−0.01	0.03	−0.18	−0.02	0.03	−0.67	0	0.04	−0.05
	Anaphor type	−0.25	0.03	−8.11*	−0.13	0.03	−4.04*	−0.07	0.04	−1.94
	Anaph. type×Interf.	0.01	0.03	0.29	0.01	0.03	0.17	−0.02	0.04	−0.53
FPRP	Interference	−0.08	0.13	−0.62	−0.11	0.13	−0.85	−0.19	0.12	−1.56
	Anaphor type	−1.26	0.13	−9.4*	−0.32	0.13	−2.45*	−0.4	0.12	−3.36*
	Anaph. type×Interf.	−0.08	0.13	−0.63	0.09	0.13	0.7	−0.25	0.12	−2.1*
TFT	Interference	0.04	0.03	1.46	0.02	0.03	0.64	0.01	0.03	0.5
	Anaphor type	−0.07	0.03	−2.44*	0.09	0.03	3.15*	0.03	0.03	0.89
	Anaph. type×Interf.	0.05	0.03	1.7	−0.01	0.03	−0.4	0.02	0.03	0.65

Main effects of interference, anaphor type and their interaction at the pre-critcal region n−1, the reflexive/pronominal possessive (REFL/PRON) and the post-critical region n+1. The dependent measures are log-first-pass reading time, log-regression-path duration, first-pass regression probability and log-total fixation time. Statistically significant ($\alpha = 0.05$) effects are marked with an asterisk.

with and without a first-pass regression from the post-critical region separately.

In trials without a first-pass regression from n+1, the interference effect in pronominal possessives in questions targeting the critical dependency (i.e., the effect observed in the overall data) was highly significant (estimate $= -1.19$, $SE = 0.28$, $z = -4.21$, $p < 0.0001$). By contrast, in trials with a first-pass regression initiated at n+1, this effect did not reach significance (estimate $= -0.94$, $SE = 0.57$, $z = -1.66$, $p = 0.09$). This *post-hoc* analysis clearly shows that the interference effect in response accuracies in pronominal possessives was driven by trials in which no first-pass regression was initiated, i.e., by the trials responsible for the facilitation observed in FPRP.

4.3. Discussion

We did not find any evidence for encoding interference affecting processing times of Swedish anaphor-antecedent dependencies. Together with the results of Experiments 1 and 2, this suggests that in materials with a classical gender-match/mismatch manipulation, encoding interference does not affect retrieval latencies of the antecedent. In comprehension questions, we did not see evidence for encoding interference affecting retrieval probability of the antecedent either. This is in contrast to the pattern observed in response accuracies of Experiments 1 and 2.

Evidence for interference occurring at the moment of retrieval was observed in online and offline measures. The lower proportion of first-pass regressions initiated at the region directly after the gender-marked pronominal possessive in conditions with a gender-matching distractor indicates a processing facilitation due to a cue-matching distractor. Such a facilitation can be explained in terms of misretrievals of the gender-matching distractor under the assumption that misretrievals go along with shorter retrieval latencies. The lower response accuracies in comprehension questions targeting the retrieval of the antecedent support this explanation. Indeed, the *post-hoc* analysis of response accuracies for trials with and without a first-pass regression from the post-critical region clearly shows that the facilitation observed in first-pass regressions is directly connected to misretrievals of the gender-matching distractor.

The cue-based ACT-R model of sentence processing (Lewis and Vasishth, 2005) predicts misretrievals of the gender-matching distractor. These misretrievals are predicted to lead to shorter retrieval latencies, i.e., a processing facilitation, in the respective trials. However, the ACT-R model also predicts inhibition between the gender-matching distractor and the antecedent leading to longer retrieval latencies of the antecedent. Overall, the predicted direction of the interference effect therefore depends on the concrete parameter setting of the model. With the default parameter setting, inhibitory interference (i.e., the opposite effect than the one in the data) is predicted. If one assumes a particularly high activation of the distractor, ACT-R predicts the observed pattern because the highly activated distractor is misretrieved in a considerable proportion of the trials, which leads to a speed-up in the observed mean retrieval latencies (Jäger et al., 2015). Indeed, facilitation in a configuration similar to our materials has been observed in previous studies (Sturt, 2003; Cunnings and Felser, 2013; Laurinavichyute et al., 2015; Patil et al., unpublished manuscript). An argument favoring the assumption that the distractor is highly activated in our materials is that, similar to the other experiments reporting facilitation, the distractor is in subject position. Moreover, the distractor has a recency advantage over the antecedent as it is linearly closer to the retrieval site. Indeed, ACT-R predicts a recency advantage which follows from the assumption that an item's activation level decreases as a function of the passage of time (decay) and intervening material (interference). In sum, under the plausible assumption that the distractor is highly activated in our materials, cue-based

retrieval interference as implemented in the ACT-R model can account for the observed pattern. Hence, the interference effect in pronominal possessives can be interpreted as evidence favoring a cue-based retrieval mechanism. However, it should be kept in mind that pronominal possessives are not subject to Binding Principle A (Chomsky, 1981). Hence, the observed effects cannot be interpreted as evidence against theories of sentence processing claiming that Principle A is immune to interference from structurally illicit antecedents (Nicol and Swinney, 1989; Phillips et al., 2011; Dillon et al., 2013).

An alternative explanation that can account for the facilitation leading to misretrievals of the gender-matching distractor in pronominal possessives but not in reflexive possessives builds on the fact that we are comparing reflexive possessives which are subject to Binding Principle A with pronominal possessives which are subject to Binding Principle B. As mentioned above, pronominal possessives must not be bound in their local domain (Binding Principle B, see Chomsky, 1981). In the syntax-semantic literature about the interpretation of pronouns, it has been proposed that in the presence of a local c-commanding noun phrase which matches the gender feature of the anaphor (as the gender-matching distractor in the pronominal possessives conditions of Experiment 3), local binding is preferred over a non-local antecedent (Fox, 1998; Heim and Kratzer, 1998). This leads to a temporary violation of Binding Principle B. Only after the local binder has successfully been inhibited, the actual search for the structurally licit antecedent is initiated (Grodzinsky and Reinhart, 1993; Reinhart, 2000; Reuland, 2011). If in our materials, the syntactically local binder of the pronominal possessive (i.e., the distractor) is accessed in a first stage of the retrieval process, in a certain proportion of the trials, this local binder might be misretrieved in case it matches the gender of the pronominal possessive and the search for the antecedent terminates already after this first stage. Such a scenario would explain the misretrievals reflected in response accuracies and also the speed-up in trials where misretrievals occurred. This model correctly predicts that facilitatory interference should be observed only with Principle B pronouns, not with Principle A reflexives since in reflexives, the local binder is the licit antecedent. Crucially, the absence of an effect in our reflexive possessive conditions is not explained by them being unmarked for gender but rather by their syntactic binding properties.

To summarize, we found no evidence for encoding interference affecting the processing of Swedish possessives. We did observe evidence for retrieval interference in gender-marked pronominal possessives. The presence of a gender-matching distractor led to facilitated processing, presumably as a consequence of misretrievals of the latter in a certain proportion of trials. Although this pattern can be explained in terms of unconstrained cue-based retrieval, it is also consistent with the view that comprehending a pronoun constrained by Principle B requires comprehenders to temporarily consider and inhibit coreference with the local subject (the distractor in our materials). However, it should be noted that recent evidence from English pronouns reported by Chow et al. (2014) is inconsistent with the idea of first accessing and subsequently inhibiting a local antecedent. In none of their five reading experiments did they observe a facilitatory effect on pronoun resolution from a feature-matching local antecedent.

5. General Discussion

We set out to find evidence for encoding interference in the processing of reflexives. With respect to the current debate about structure-based vs. unconstrained cue-based retrieval subserving the processing of reflexives, the question whether encoding interference can be observed in reflexives is crucial because, as has been argued by Dillon (2011), encoding interference provides an alternative explanation for interference effects in reflexives which originally have been attributed to cue-based retrieval interference and hence taken as evidence for unconstrained cue-based retrieval (Badecker and Straub, 2002; Chen et al., 2012; Jäger et al., 2015; Patil et al., unpublished manuscript).

In order to decide whether encoding interference is present in the processing of reflexives, we conducted two experiments on the German reflexive *sich*. In contrast to previous studies, where encoding and cue-based retrieval interference made the same predictions, the gender-unmarked *sich* allowed us to pit against each other the predictions of retrieval and encoding interference. Cue-based retrieval interference predicts no effect of gender of a structurally inaccessible distractor whereas encoding interference predicts a slow-down when the gender of the distractor matches the gender of the antecedent. Neither with self-paced reading nor with eye-tracking did we find any indication for an online interference effect caused by a gender-sharing distractor, although the statistical power of our experiments was considerably higher than the one of previous experiments reporting interference effects in reflexives. We conducted a third experiment on Swedish possessives to cross-linguistically validate our finding. The interaction between interference and anaphor type provided further support for the conclusion that sharing the gender feature with a distractor does not lead to encoding interference in the processing of reflexives. Although we did not find any evidence that encoding interference affected online processing ease, response accuracies in the comprehension questions of Experiment 1 indicate that encoding interference might have caused misretrievals of the gender-sharing distractor. However, this effect was only marginal in Experiment 2 and could not be replicated in Experiment 3. Critically, these supposed misretrievals observed in Experiment 1 are not reflected in online processing measures. In sum, there is no evidence for encoding interference affecting online processing measures. Therefore, there is no evidence for the proposal that online interference effects reported in previous studies on reflexives arise from encoding interference. This finding therefore provides support for the assumption that interference effects observed in reflexive processing arise at the moment of retrieval rather than at the encoding stage. In other words, encoding interference is not a plausible explanation for reconciling interference effects with a structure-based account of reflexive processing. Thus, taken together with the interference effects reported in previous studies on reflexive processing, our findings favor an unconstrained cue-based retrieval architecture.

Lastly, we want to emphasize that our results should not be interpreted as evidence for the absence of encoding interference in sentence processing *per se*. Indeed, the effect at the relative clause verb in Experiment 2 might reflect encoding interference. The presence of encoding interference *as such* is in principle not incompatible with a content-addressable architecture since content-addressability is an architectural mechanism concerning the *retrieval*, but not the *encoding* or the *maintenance* of an item in working memory.

More generally, our findings provide support for a content-addressable memory architecture subserving language comprehension. This adds to a growing body of evidence from various kinds of syntactic dependencies such as filler-gap (McElree et al., 2003) and subject-verb dependencies (Van Dyke and Lewis, 2003; Van Dyke and McElree, 2006, 2011; Van Dyke, 2007; Wagers et al., 2009; Dillon et al., 2013), the licensing of negative-polarity items (Vasishth et al., 2008) and verb-phrase ellipsis (Martin and McElree, 2008), suggesting that the parser uses a cue-based retrieval mechanism to process these dependencies. One fundamental question in sentence processing research is whether the human parser uses qualitatively different retrieval mechanisms in the processing of different kinds of dependencies. Indeed, proponents of the structure-based account of reflexive processing have argued that the retrieval mechanisms mediating the processing of reflexives differ qualitatively from the ones used, e.g., in the processing of subject-verb dependencies (Phillips et al., 2011; Dillon et al., 2013). Hence, evidence for cue-based retrieval subserving the processing of reflexives is one important piece of evidence toward a content-addressable model of working memory underlying sentence processing in general, which not only invokes qualitatively similar working memory mechanisms to explain the processing of different kinds of linguistic dependencies, but, even beyond that, locates the language processing system within a general cognitive architecture where independently motivated working memory mechanisms operate on linguistic representations.

Acknowledgments

This work was partly funded by the Studienstiftung des deutschen Volkes by awarding a scholarship to the first author. Publication of this article was supported by the Deutsche Forschungsgemeinschaft and the Open Access Publishing Fund of the University of Potsdam.

Supplementary Material

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpsyg.2015.00506/abstract

References

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., and Qin, Y. (2004). An integrated theory of the mind. *Psychol. Rev.* 111, 1036–1060. doi: 10.1037/0033-295X.111.4.1036

Anderson, J. R., Bothell, D., Lebiere, C., and Matessa, M. (1998). An integrated theory of list memory. *J. Mem. Lang.* 38, 341–380. doi: 10.1006/jmla.1997.2553

Anderson, J. R., and Lebiere, C. (1998). *Atomic Components of Thought*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Anderson, M. C., and Neely, J. H. (1996). "Interference and inhibition in memory retrieval," in *Memory*, eds E. L. Bjork and R. Bjork (San Diego, CA: Academic Press), 237–313.

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J., Fiedler, K., et al. (2013). Recommendations for increasing replicability in psychology. *Eur. J. Pers.* 27, 108–119. doi: 10.1002/per.1919

Baayen, R. H., Davidson, D. J., and Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *J. Mem. Lang.* 59, 390–412. doi: 10.1016/j.jml.2007.12.005

Badecker, W., and Straub, K. (2002). The processing role of structural constraints on the interpretation of pronouns and anaphors. *J. Exp. Psychol. Learn. Mem. Cogn.* 28, 748–769. doi: 10.1037/0278-7393.28.4.748

Bates, D. M., Maechler, M., Bolker, B., and Walker, S. (2014). *lme4: Linear Mixed-Effects Models Using Eigen and S4*. R package version 1.0-6.

Box, G. E. P., and Cox, D. R. (1964). An analysis of transformations. *J. R. Stat. Soc.* 26, 211–252.

Chen, Z., Jäger, L., and Vasishth, S. (2012). "How structure-sensitive is the parser? Evidence from mandarin chinese," in *Empirical Approaches to Linguistic Theory: Studies of Meaning and Structure, Studies in Generative Grammar*, eds B. Stolterfoht and S. Featherston (Berlin: Mouton de Gruyter), 43–62.

Chomsky, N. (1981). *Lectures on Government and Binding*. Dordrecht: Foris.

Chow, W.-Y., Lewis, S., and Phillips, C. (2014). Immediate sensitivity to structural constraints in pronoun resolution. *Front. Psychol.* 5:630. doi: 10.3389/fpsyg.2014.00630

Clackson, K., and Heyer, V. (2014). Reflexive anaphor resolution in spoken language comprehension: structural constraints and beyond. *Front. Psychol.* 5:904. doi: 10.3389/fpsyg.2014.00904

Clifton, C. J., Staub, A., and Rayner, K. (2007). "Eye movements in reading words and sentences," in *Eye Movements: A Window on Mind and Brain*, eds R. P. Van Gompel, M. H. Fischer, W. S. Murray, and R. L. Hill (Oxford, UK; Amsterdam: Elsevier Inc.), 341–372.

Cousineau, D. (2005). Confidence intervals in within-subjects designs: a simpler solution to Loftus and Masson's method. *Tutorial Quant. Methods Psychol.* 1, 42–45.

Cunnings, I., and Felser, C. (2013). The role of working memory in the processing of reflexives. *Lang. Cogn. Process.* 28, 188–219. doi: 10.1080/01690965.2010.548391

Dillon, B. (2011). *Structured Access in Sentence Comprehension*. Ph.D. thesis, University of Maryland, College Park, MD.

Dillon, B., Mishler, A., Sloggett, S., and Phillips, C. (2013). Contrasting intrusion profiles for agreement and anaphora: experimental and modeling evidence. *J. Mem. Lang.* 69, 85–103. doi: 10.1016/j.jml.2013.04.003

Elbourne, P. D. (2005). *Situations and Individuals, Vol. 14 of Current Studies in Linguistics*. Cambridge, MA: MIT Press.

Everaert, M. (1986). *The Syntax of Reflexivation*. Dordrecht: Foris.

Fedorenko, E., Gibson, E., and Rohde, D. (2006). The nature of working memory capacity in sentence comprehension: evidence against domain specific resources. *J. Mem. Lang.* 54, 541–553. doi: 10.1016/j.jml.2005.12.006

Fisher, R. (1937). *The Design of Experiments, 2nd Edn.* Edinburgh, UK: Oliver and Boyd.

Fox, D. (1998). "Locality in variable binding," in *Is the Best Good Enough? Optimality and Competition in Syntax*, eds P. Barbosa, D. Fox, P. Hagstrom, M. McGinnis, and D. Pesetsky (Cambridge, MA: MIT Press), 129–155.

Gast, V., and Haas, F. (2008). "Reflexive and reciprocal readings of anaphors in German and other European languages," in *Reciprocals and Reflexives: Theoretical and Typological Explorations*, eds E. König and V. Gastpages (Berlin: Mouton de Gruyter), 307–346.

Gibson, E. (2000). "Dependency locality theory: a distance-based theory of linguistic complexity," in *Image, Language, Brain*, eds A. Marantz, Y. Miyashita, and W. O'Neil (Cambridge, MA: MIT Press), 95–126.

Gordon, P. C., Hendrick, R., and Johnson, M. (2001). Memory interference during language processing. *J. Exp. Psychol. Learn. Mem. Cogn.* 27, 1411–1423. doi: 10.1037/0278-7393.27.6.1411

Gordon, P. C., Hendrick, R., and Johnson, M. (2004). Effects of noun phrase type on sentence complexity. *J. Mem. Lang.* 51, 97–114. doi: 10.1016/j.jml.2004.02.003

Gordon, P. C., Hendrick, R., Johnson, M., and Lee, Y. (2006). Similarity-based interference during language comprehension: evidence from eye tracking during reading. *J. Exp. Psychol. Learn. Mem. Cogn.* 32, 1304–1321. doi: 10.1037/0278-7393.32.6.1304

Gordon, P. C., Hendrick, R., and Levine, W. H. (2002). Memory-load interference in syntactic processing. *Psychol. Sci.* 13, 425–430.

Grodzinsky, Y., and Reinhart, T. (1993). The innateness of binding and coreference. *Linguist. Inq.* 24, 69–101. doi: 10.1111/1467-9280.00475

Heim, I., and Kratzer, A. (1998). *Semantics in Generative Grammar*. Oxford, UK: Blackwell.

Heister, J., Würzner, K.-M., Bubenzer, J., Pohl, E., Hanneforth, T., Geyken, A., et al. (2011). dlexDB – eine lexikalische Datenbank für die psychologische und linguistische Forschung. *Psychol. Rundsch.* 62, 10–20. doi: 10.1026/0033-3042/a000029

Holmes, P., and Hinchliffe, I. (1994). *Swedish. A Comprehensive Grammar*. New York, NY: Routledge.

Jäger, L., Engelmann, F., and Vasishth, S. (2015). Retrieval interference in reflexive processing: experimental evidence from Mandarin, and computational modeling. *Front. Psychol.* 6:617. doi: 10.3389/fpsyg.2015.00617

Just, M. A., Carpenter, P. A., and Woolley, J. D. (1982). Paradigms and processes in reading comprehension. *J. Exp. Psychol. Gen.* 111, 228–238. doi: 10.1037/0096-3445.111.2.228

Kaiser, E. (2003). "Encoding (non)locality in anaphoric relations," in *Generative Approaches to Finnic and Saami Linguistics*, eds D. Nelson and S. Manninen (Stanford, CA: CSLI Publications), 269–293.

King, J., Andrews, C., and Wagers, M. (2012). "Do reflexives always find a grammatical antecedent for themselves?," in *25th Annual CUNY Conference on Human Sentence Processing* (New York, NY: The CUNY Graduate Center), 67.

Kush, D. (2013). *Respecting Relations: Memory Access and Antecedent Retrieval in Incremental Sentence Processing*. Ph.D thesis, University of Maryland, College Park, MD.

Kush, D., and Phillips, C. (2014). Local anaphor licensing in an SOV language: implications for retrieval strategies. *Front. Psychol.* 5:1252. doi: 10.3389/fpsyg.2014.01252

Laurinavichyute, A., Jäger, L., Akinina, Y., Benz, L., and Dragoy, O. (2015). "Teasing apart retrieval and encoding interference in Russian reflexives," in *28th Annual CUNY Conference on Human Sentence Processing* (Los Angeles, CA: University of Southern California), 140.

Lewandowsky, S., Geiger, S. M., and Oberauer, K. (2008). Interference-based forgetting in verbal short-term memory. *J. Mem. Lang.* 59, 200–222. doi: 10.1016/j.jml.2008.04.004

Lewis, R. L., and Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cogn. Sci.* 29, 375–419. doi: 10.1207/s15516709cog0000/25

Longobardi, G. (1994). Reference and proper names: a theory of N-movement in syntax and logical form. *Linguist. Inq.* 25, 609–665.

Martin, A. E., and McElree, B. (2008). A content-addressable pointer mechanism underlies comprehension of verb-phrase ellipsis. *J. Mem. Lang.* 58, 879–906. doi: 10.1016/j.jml.2007.06.010

McElree, B. (2006). "Accessing recent events," in *The Psychology of Learning and Motivation: Advances in Research and Theory*, Vol. 46, ed B. H. Ross (San Diego, CA; London, UK: Elsevier Inc.), 155–200.

McElree, B., Foraker, S., and Dyer, L. (2003). Memory structures that subserve sentence comprehension. *J. Mem. Lang.* 48, 67–91. doi: 10.1016/S0749-596X(02)00515-6

Morey, R. (2008). Confidence intervals from normalized data: a correction to Cousineau (2005). *Tutorial Quant. Methods Psychol.* 4, 61–64.

Nairne, J. S. (1988). A framework for interpreting recency effects in immediate serial recall. *Mem. Cogn.* 16, 343–352. doi: 10.3758/BF03197045

Nairne, J. S. (1990). A feature model of immediate memory. *Mem. Cogn.* 18, 251–269. doi: 10.3758/BF03213879

Nicol, J., and Swinney, D. (1989). The role of structure in coreference assignment during sentence comprehension. *J. Psycholinguist. Res.* 18, 5–19. doi: 10.1007/BF01069043

Oberauer, K., and Kliegl, R. (2006). A formal model of capacity limits in working memory. *J. Mem. Lang.* 55, 601–626. doi: 10.1016/j.jml.2006.08.009

Parker, D., and Phillips, C. (2014). "Selective priority for structure in memory retrieval," in *27th Annual CUNY Conference on Human Sentence Processing* (Columbus, OH: Ohio State University), 100.

Phillips, C., Wagers, M. W., and Lau, E. F. (2011). "Grammatical illusions and selective fallibility in real-time language comprehension," in *Experiments at the Interfaces, Vol. 37 of Syntax and Semantics*, ed J. Runner (Bingley, UK: Emerald Group Publishing Limited), 147–180.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available online at: http://www.R-project.org/

Reinhart, T. (2000). "Strategies of anaphora resolution," in *Interface Strategies*, eds H. Bennis, M. Everaert, and E. J. Reuland (Amsterdam: Royal Academy of Sciences), 295–325.

Reinhart, T., and Reuland, E. (1993). Reflexivity. *Linguist. Inq.* 24, 657–720.

Reuland, E. (2001). Primitives of binding. *Linguist. Inq.* 32, 439–492. doi: 10.1162/002438901750372522

Reuland, E. (2011). *Anaphora and Language Design*. Cambridge, MA: MIT Press.

Reuland, E., and Reinhart, T. (1995). "Pronouns, anaphors and case," in *Studies in Comparative Germanic Syntax*, eds H. Haider, S. Olsen, and S. Vikner (Dordrecht: Kluwer), 241–269.

Simmons, J. P., Nelson, L. D., and Simonsohn, U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* 22, 1359–1366. doi: 10.1177/0956797611417632

Staub, A., and Rayner, K. (2007). "Eye movements and on-line comprehension processes," in *The Oxford Handbook of Psycholinguistics*, eds M. G. Gaskell and G. Altmann (Oxford, UK: Oxford University Press), 327–342.

Sturt, P. (2003). The time-course of the application of binding constraints in reference resolution. *J. Mem. Lang.* 48, 542–562. doi: 10.1016/S0749-596X(02)00536-3

Van Dyke, J. (2007). Interference effects from grammatically unavailable constituents during sentence processing. *J. Exp. Psychol. Learn. Mem. Cogn.* 33, 407–430. doi: 10.1037/0278-7393.33.2.407

Van Dyke, J., and Lewis, R. L. (2003). Distinguishing effects of structure and decay on attachment and repair: a cue-based parsing account of recovery from misanalyzed ambiguities. *J. Mem. Lang.* 49, 285–316. doi: 10.1016/S0749-596X(03)00081-0

Van Dyke, J., and McElree, B. (2011). Cue-dependent interference in comprehension. *J. Mem. Lang.* 65, 247–263. doi: 10.1016/j.jml.2011.05.002

Van Dyke, J., and McElree, B. (2006). Retrieval interference in sentence comprehension. *J. Mem. Lang.* 55, 157–166. doi: 10.1016/j.jml.2006.03.007

Vasishth, S., Bruessow, S., Lewis, R. L., and Drenhaus, H. (2008). Processing polarity: how the ungrammatical intrudes on the grammatical. *Cogn. Sci.* 32, 685–712. doi: 10.1080/03640210802066865

Wagers, M. W., Lau, E. F., and Phillips, C. (2009). Agreement attraction in comprehension: representations and processes. *J. Mem. Lang.* 61, 206–237. doi: 10.1016/j.jml.2009.04.002

Watkins, O. C., and Watkins, M. J. (1975). Buildup of proactive inhibition as a cue-overload effect. *J. Exp. Psychol. Hum. Learn. Mem.* 104, 442–452. doi: 10.1037/0278-7393.1.4.442

Xiang, M., Dillon, B., and Phillips, C. (2009). Illusory licensing effects across dependency types: ERP evidence. *Brain Lang.* 108, 40–55. doi: 10.1016/j.bandl.2008.10.002

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2015 Jäger, Benz, Roeser, Dillon and Vasishth. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Appendix

TABLE A1 | Experiment 1.

	Der/Die	*Dieb/-in*	*dem/der*	*der/die*	*Hehler/-in*	*befohlen*	*hat*	*zu stehlen*	*hat*	*überraschenderweise*	*sich*	*und die Kollegen*	*angezeigt*
FPRT													
Gend.-overlap - masc. ant.	448 (8)	547 (12)	460 (11)	412 (6)	505 (19)	523 (15)	499 (14)	669 (40)	525 (15)	482 (8)	444 (10)	1221 (16)	506 (13)
Gend.-overlap - fem. ant.	462 (12)	622 (17)	493 (10)	439 (8)	552 (16)	599 (28)	528 (16)	611 (24)	552 (19)	503 (13)	438 (6)	1233 (13)	530 (15)
No gend.-overlap - masc. ant.	482 (15)	562 (22)	459 (8)	426 (10)	535 (14)	571 (17)	515 (17)	614 (23)	517 (15)	485 (9)	446 (8)	1224 (11)	535 (12)
No gend.-overlap - fem. ant.	453 (10)	611 (15)	485 (9)	443 (9)	534 (13)	538 (12)	523 (19)	594 (22)	517 (16)	479 (9)	430 (6)	1218 (12)	539 (16)

Means and standard errors of raw reading times in ms for each region by experimental condition. Between-participants variance has been removed using Cousineau (2005)'s normalization with Morey (2008)'s correction factor.

TABLE A2 | Experiment 2.

	Der/Die	*Dieb/-in*	*dem/der*	*der/die*	*Hehler/-in*	*befohlen*	*hat*	*zu stehlen*	*hat*	*überraschenderweise*	*sich*	*und die Kollegen*	*angezeigt*
FPRT													
Gend.-overlap - masc. ant.	257 (5)	351 (9)	288 (6)	275 (7)	333 (10)	344 (7)	272 (6)	431 (11)	301 (7)	350 (10)	286 (6)	246 (5)	348 (8)
Gend.-overlap - fem. ant.	244 (6)	481 (15)	338 (10)	293 (8)	372 (12)	347 (7)	273 (6)	424 (11)	295 (7)	357 (10)	300 (6)	251 (5)	341 (6)
No gend.-overlap - masc. ant.	250 (5)	364 (13)	296 (6)	257 (8)	402 (11)	341 (7)	270 (6)	417 (10)	295 (8)	360 (9)	292 (5)	252 (6)	356 (7)
No gend.-overlap - fem. ant.	243 (5)	468 (15)	342 (10)	308 (9)	314 (10)	344 (7)	283 (7)	421 (11)	305 (7)	351 (9)	299 (6)	252 (5)	347 (6)
RPD													
Gend.-overlap - masc. ant.	NA	408 (11)	336 (10)	379 (13)	434 (16)	389 (12)	384 (29)	748 (33)	373 (25)	487 (21)	375 (18)	348 (19)	449 (17)
Gend.-overlap - fem. ant.	NA	581 (17)	408 (12)	448 (16)	639 (22)	472 (20)	396 (27)	830 (43)	433 (32)	530 (27)	362 (14)	327 (12)	451 (19)
No gend.-overlap - masc. ant.	NA	415 (13)	337 (9)	326 (12)	501 (15)	413 (13)	337 (15)	677 (28)	387 (28)	471 (16)	350 (15)	330 (11)	463 (18)
No gend.-overlap - fem. ant.	NA	563 (16)	421 (16)	478 (16)	578 (22)	472 (19)	383 (24)	753 (38)	413 (37)	450 (17)	375 (15)	358 (19)	438 (15)
FPRP													
Gend.-overlap - masc. ant.	NA	15 (1)	7 (1)	22 (2)	17 (2)	6 (1)	12 (1)	28 (2)	8 (1)	20 (2)	11 (1)	15 (1)	14 (1)
Gend.-overlap - fem. ant.	NA	19 (2)	13 (1)	26 (2)	29 (2)	12 (1)	11 (1)	28 (2)	10 (1)	18 (2)	10 (1)	15 (1)	14 (1)
No gend.-overlap - masc. ant.	NA	17 (2)	8 (1)	16 (1)	15 (1)	10 (1)	10 (1)	26 (2)	7 (1)	17 (2)	9 (1)	14 (1)	13 (1)
No gend.-overlap - fem. ant.	NA	21 (2)	13 (1)	30 (2)	32 (2)	13 (1)	10 (1)	29 (2)	9 (1)	16 (2)	11 (1)	16 (1)	13 (1)
TFT													
Gend.-overlap - masc. ant.	346 (10)	542 (16)	464 (14)	406 (16)	514 (19)	549 (16)	352 (11)	663 (18)	355 (12)	554 (16)	365 (11)	309 (8)	462 (11)
Gend.-overlap - fem. ant.	370 (11)	838 (25)	603 (21)	495 (19)	713 (23)	580 (15)	351 (12)	662 (18)	342 (12)	543 (17)	351 (9)	308 (9)	444 (11)
No gend.-overlap - masc. ant.	336 (9)	535 (18)	473 (14)	324 (14)	618 (19)	530 (14)	346 (11)	617 (16)	353 (13)	528 (15)	358 (9)	306 (9)	465 (11)
No gend.-overlap - fem. ant.	366 (10)	772 (23)	594 (20)	515 (19)	582 (22)	551 (16)	353 (11)	613 (17)	348 (10)	530 (17)	362 (9)	310 (8)	436 (10)

Means and standard errors of raw first-pass reading time (FPRT), regression-path duration (RPD) and total-fixation time (TFT) in ms and first-pass regression probability (FPRP) in percentages for each region by experimental condition. From continuous dependent variables, between-participants variance has been removed using Cousineau (2005)'s normalization with Morey (2008)'s correction factor.

TABLE A3 | Experiment 3.

	Åke Åke	*säger som*	*att Alf/Ann*	*Alf/Ann tackade*	*jobbade med ringer*	*hans sina*	*sysslingar sysslingar*	*på helgerna på kvällen*
FPRT								
Pron - gend.-overlap	381 (10)	403 (14)	257 (9)	292 (10)	421 (13)	287 (6)	327 (10)	650 (21)
Pron - no gend.-overlap	393 (11)	388 (12)	277 (11)	288 (10)	393 (11)	285 (6)	334 (11)	673 (22)
Refl - gend.-overlap	405 (11)	286 (10)	295 (10)	503 (17)	391 (12)	310 (11)	308 (11)	673 (21)
Refl - no gend.-overlap	397 (12)	293 (10)	300 (9)	512 (16)	397 (13)	297 (8)	327 (12)	659 (24)
RPD								
Pron - gend.-overlap	NA	553 (21)	421 (26)	432 (26)	505 (21)	400 (22)	549 (42)	3636 (133)
Pron - no gend.-overlap	NA	513 (18)	405 (34)	413 (24)	521 (26)	421 (26)	525 (32)	3370 (115)
Refl - gend.-overlap	NA	443 (20)	435 (22)	669 (31)	749 (42)	500 (37)	656 (56)	4120 (163)
Refl - no gend.-overlap	NA	443 (19)	455 (25)	659 (28)	738 (40)	568 (49)	643 (55)	3882 (149)
FPRP								
Pron - gend.-overlap	NA	19 (2)	20 (2)	16 (2)	11 (2)	16 (2)	17 (2)	80 (2)
Pron - no gend.-overlap	NA	18 (2)	16 (2)	15 (2)	13 (2)	16 (2)	24 (2)	81 (2)
Refl - gend.-overlap	NA	27 (2)	19 (2)	15 (2)	31 (2)	19 (2)	28 (2)	81 (2)
Refl - no gend.-overlap	NA	25 (2)	21 (2)	14 (2)	31 (2)	22 (2)	26 (2)	79 (2)
TFT								
Pron - gend.-overlap	856 (31)	1121 (40)	444 (23)	713 (29)	1010 (36)	601 (21)	685 (24)	1067 (30)
Pron - no gend.-overlap	799 (33)	1047 (33)	468 (25)	634 (25)	917 (30)	579 (19)	675 (24)	1128 (34)
Refl - gend.-overlap	1104 (46)	781 (32)	930 (38)	1412 (45)	1039 (34)	544 (20)	660 (24)	1133 (35)
Refl - no gend.-overlap	991 (34)	777 (29)	890 (36)	1398 (49)	1051 (36)	532 (20)	663 (24)	1130 (38)

Means and standard errors of raw first-pass reading time (FPRT), regression-path duration (RPD) and total-fixation time (TFT) in ms and first-pass regression probability (FPRP) in percentages for each region by experimental condition. From continuous dependent variables, between-participants variance has been removed using Cousineau (2005)'s normalization with Morey (2008)'s correction factor.

frontiers in PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 24 September 2014
doi: 10.3389/fpsyg.2014.01025

The structure-sensitivity of memory access: evidence from Mandarin Chinese

Brian Dillon[1], Wing-Yee Chow[2], Matthew Wagers[3], Taomei Guo[4]*, Fengqin Liu[5] and Colin Phillips[6]

[1] Department of Linguistics, University of Massachusetts, Amherst, MA, USA
[2] Basque Center on Cognition, Brain and Language, Donostia – San Sebastián, Spain
[3] Department of Linguistics, University of California Santa Cruz, Santa Cruz, CA, USA
[4] State Key Laboratory of Cognitive Neuroscience and Learning, Center for Collaboration and Innovation in Brain and Learning Sciences, Beijing Normal University, Beijing, China
[5] State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University, Beijing, China
[6] Department of Linguistics, Maryland Language Science Center, University of Maryland, College Park, MD, USA

Edited by:
Claudia Felser, University of Potsdam, Germany

Reviewed by:
Stephani Foraker, Suny Buffalo State, USA
Nayoung Kwon, Konkuk University, South Korea

***Correspondence:**
Taomei Guo, State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University, Beijing 100875, China
e-mail: guotm@bnu.edu.cn

The present study examined the processing of the Mandarin Chinese long-distance reflexive *ziji* to evaluate the role that syntactic structure plays in the memory retrieval operations that support sentence comprehension. Using the multiple-response speed-accuracy tradeoff (MR-SAT) paradigm, we measured the speed with which comprehenders retrieve an antecedent for *ziji*. Our experimental materials contrasted sentences where *ziji*'s antecedent was in the local clause with sentences where *ziji*'s antecedent was in a distant clause. Time course results from MR-SAT suggest that *ziji* dependencies with syntactically distant antecedents are slower to process than syntactically local dependencies. To aid in interpreting the SAT data, we present a formal model of the antecedent retrieval process, and derive quantitative predictions about the time course of antecedent retrieval. The modeling results support the *Local Search hypothesis*: during syntactic retrieval, comprehenders initially limit memory search to the local syntactic domain. We argue that Local Search hypothesis has important implications for theories of locality effects in sentence comprehension. In particular, our results suggest that not all locality effects may be reduced to the effects of temporal decay and retrieval interference.

Keywords: working memory, reflexive processing, speed-accuracy trade-off, Mandarin Chinese, sentence processing

INTRODUCTION

One fundamental question for models of sentence comprehension is the question of how comprehenders are able to construct long-distance linguistic dependencies reliably and rapidly in comprehension. Long-distance dependencies occur whenever two non-adjacent elements in a sentence must be syntactically and/or semantically integrated with each other. For example, in a sentence like "*William took a terrible yet interesting photo of himself*," the relationship between the reflexive anaphor *himself* and its antecedent *William* is constructed across multiple intervening words. Recent models of sentence comprehension have advanced the hypothesis that this sort of syntactic dependency formation minimally requires the use of memory retrieval mechanisms to access temporally distant syntactic encodings. On this view, to interpret the reflexive in the sentence above, comprehenders must retrieve a representation of the antecedent from memory. Moreover, it has been argued that the memory retrieval mechanisms that underlie sentence comprehension share a number of key features with domain-general retrieval mechanisms (McElree, 2000; McElree et al., 2003; Van Dyke and Lewis, 2003; Lewis and Vasishth, 2005; Lewis et al., 2006). This view receives support from mounting evidence that comprehenders rely on a cue-based, direct access retrieval mechanism during syntactic comprehension. Cue-based retrieval mechanisms allow direct access to syntactic encodings by matching retrieval cues to the features of all item representations in memory in a parallel fashion. Items whose features provide a close match to the retrieval cues are then retrieved for further processing (for a discussion of implementations of this idea, see Clark and Gronlund, 1996).

Despite the growing evidence in favor of cue-based retrieval in sentence comprehension, these models still face a number of difficult theoretical questions. One central question concerns the nature of the retrieval cues used to retrieve syntactic dependents during processing (Van Dyke and McElree, 2011; Dillon et al., 2013; Kush, 2013). Existing evidence suggests that comprehenders use both semantic and syntactic cues to guide retrieval (Van Dyke and McElree, 2006, 2011; Van Dyke, 2007). Furthermore, there is some evidence that syntactic cues may be given priority over semantic or morphological cues, although this may depend on the kind of dependency being formed (Van Dyke and McElree, 2011; Dillon et al., 2013). However, very little is known about the nature of the syntactic cues that guide retrieval operations. In the present paper, we address this question by asking whether syntactic cues refer only to the attributes of individual syntactic

encodings, such as their case or thematic role (*item information*), or whether syntactic cues distinguish constituents based on their hierarchical or linear distance from the retrieval site (*position information*). We hypothesize that the cues that guide memory retrieval during parsing do include positional syntactic information, and furthermore, that comprehenders use positional information as retrieval cues to prioritize retrieval of constituents within the local syntactic domain (what we will refer to as the *Local Search hypothesis*).

The goal of the present paper is to evaluate the Local Search hypothesis by examining the speed with which comprehenders process reflexive dependencies in Mandarin Chinese, where reflexive anaphors may be bound either inside or outside of their local clause. We then develop a formal model of a local search retrieval process, and derive quantitative predictions about the processing time necessary to recover *ziji*'s antecedent if the parser assumes a local search strategy. To preview our conclusion, the results of our investigation support the main predictions of the Local Search hypothesis. We argue that the Local Search hypothesis offers important insight into a widely-observed preference for local dependencies over distant dependencies in sentence comprehension (Kimball, 1973; Hawkins, 1994; Gibson, 1998; Bartek et al., 2011; *a.o.*). In particular, the results presented here support theories that attribute these *locality effects* to a substantive bias to search syntactically local domains at retrieval, rather than theories that attribute locality effects entirely to effects of decay or interference of items in working memory (MacDonald et al., 1994; Lewis and Vasishth, 2005; Lewis et al., 2006).

CUE-BASED RETRIEVAL IN SENTENCE PROCESSING

There are two main sources of empirical evidence that implicate the use of a cue-based retrieval mechanism in parsing. The first source of evidence comes from studies of interference effects in online processing. In a cue-based, direct access memory architecture, the process of matching the retrieval cues against all memory encodings allows rapid access to task-relevant encodings, but is susceptible to interference effects. If there is a good match between the retrieval cues and more than one item in memory, then access to the target memory can be impeded. Similarity-based interference effects have been widely documented in studies of sentence comprehension (Gordon et al., 2001, 2004; Van Dyke and Lewis, 2003; Drenhaus et al., 2005; Lewis and Vasishth, 2005; Lewis et al., 2006; Van Dyke and McElree, 2006, 2011; Van Dyke, 2007). In addition, cue-based retrieval architectures naturally account for the phenomenon of *illusory licensing*. Illusory licensing occurs when comprehenders appear to use a grammatically unavailable constituent to license a syntactically dependent element (negative polarity item licensing: Vasishth et al., 2008; Xiang et al., 2009; subject-verb agreement: Wagers et al., 2009; Dillon et al., 2013). In a cue-based retrieval architecture, this arises because a syntactically illicit constituent may be misretrieved during the search for a licensor, which in turn leads to illusory licensing of the dependent element that triggered the retrieval.

A second important line of evidence in favor of cue-based retrieval mechanisms comes from studies of the time course of memory access. In a cue-based, direct access memory architecture, only items that match the retrieval cues are contacted at retrieval, and so retrieval times are predicted to be constant over search sets of different sizes. This prediction has been supported by speed-accuracy tradeoff studies (SAT) of item recognition. In an SAT study, participants are trained to respond at a number of varying response deadlines. This allows the experimenter to derive an SAT function that tracks behavioral accuracy as a function of time. This function measures the complete time course of information processing. Importantly, SAT permits the experimenter to make separate measurements of processing speed and processing accuracy, two aspects of information processing that are confounded in simple reaction time (RT) paradigms (Wickelgren, 1977). SAT studies of recognition judgments in list memory tasks have provided support for direct access models of recognition memory by showing that the number of elements in the list does not affect processing speed (McElree and Dosher, 1989). This finding is not consistent with search models, which retrieve representations based on their location in memory. Search may proceed either in serial or parallel, but it crucially involves performing explicit comparison processes over a positionally defined search set (see Townsend and Ashby, 1983). For this reason, search models predict that memory access times should grow either as a function of the size of the search space, or as a function of the position of the target in a serial or ordered search. This prediction reflects the fact that the more sampling operations are necessary to recover the intended target, the longer time should be required for retrieval. In contrast to item recognition, the retrieval of explicit order information does appear to recruit this sort of iterative search process (McElree and Dosher, 1993; Gronlund et al., 1997).

In language comprehension, research using the SAT technique has demonstrated that memory access time does not grow with the size of the search space (McElree, 2000; McElree et al., 2003; Foraker and McElree, 2007; Martin and McElree, 2008, 2009). For instance, McElree et al. (2003) examined the processing of object cleft-constructions as in (1), in which the clefted object is separated from the associated verb by 1, 2, or 3 clauses:

a. It was the scandal that the celebrity relished.
b. It was the scandal that model believed that the celebrity relished.
c. It was the scandal that the model believed that the journalist reported that the celebrity relished. (1)

McElree and colleagues used the SAT paradigm to measure comprehension accuracy at various time points from the offset of the final verb in (1), manipulating the hierarchical (and linear) distance between the filler (*the scandal*) and the verb that hosts its gap (*relished*). The results suggested that the length manipulation impacted how accurately comprehenders were able to retrieve the *wh*-filler, as reflected in their asymptotic accuracy rates on a plausibility detection task, but that it did not impact the speed of this retrieval. McElree et al. argued that these results favor a cue-based direct access retrieval mechanism over search-based retrieval mechanisms, on the assumption that their length manipulation increased the size of the search set for the critical retrieval of the *wh*-filler. Similar results were observed for the

comprehension of verb phrase ellipsis (Martin and McElree, 2008, 2009), sluicing (Martin and McElree, 2011), and pronominal reference (Foraker and McElree, 2007).

Although previous SAT studies have consistently found that the structural distance between two elements in a dependency does not impact the speed of forming the dependency, the technique has been shown to have the power to detect other sorts of processing slowdowns that occur during sentence comprehension. For example, SAT studies have shown that processing slowdowns obtain in cases of syntactic reanalysis (McElree et al., 2003; Bornkessel et al., 2004), cases of potential lexical ambiguity (Foraker and McElree, 2007), and configurations that require multiple retrieval operations (McElree et al., 2003). Lastly, although length *per se* has not been shown to modulate processing speed, in certain cases the type of intervening material has been shown to contribute to slowed processing. McElree et al. (2003) also reported that the time necessary to process a subject-verb dependency is slowed by an intervening relative clause, but not an intervening prepositional phrase (see also Wagers and McElree, 2009).

LOCALITY EFFECTS IN A CUE-BASED ARCHITECTURE

The adoption of a cue-based, direct access architecture for syntactic processing requires a reexamination of existing theories of locality effects in sentence processing. It is widely observed that local syntactic dependencies are easier to process or, in cases of ambiguity, preferred over longer syntactic dependencies (Kimball, 1973; Frazier, 1978; Just and Carpenter, 1992; Hawkins, 1994; Gibson, 1998; Lewis and Vasishth, 2005; Lewis et al., 2006; Bartek et al., 2011; *inter alia*). Because cue-based, direct access mechanisms do not need to execute a serial search of a parse to retrieve a syntactic dependent or a pronominal antecedent, locality effects cannot emerge as a property of the access mechanism without making further assumptions. Instead, the advantage for local dependencies reflects two factors: time-based decay and interference processes (Van Dyke and Lewis, 2003; Lewis and Vasishth, 2005; Lewis et al., 2006; Bartek et al., 2011; see also Frazier and Clifton, 1998). Decay and interference both serve to degrade the availability of more distant syntactic constituents, thus making processing of longer dependencies more difficult. Decay does so by affecting the activation of constituents: more local constituents will have higher activation values by virtue of being accessed more recently. The effect of interference is more indirect. When dependencies are longer, then there are likely to be more constituent encodings in memory. When there are more items in memory, the degree of similarity-based interference is likely to be greater.

These explanations for locality effects in sentence processing stand in contrast to the explanation offered by accounts that attribute locality effects to a parsing strategy that preferentially attaches incoming constituents within a local syntactic domain (Kimball, 1973; Frazier, 1978; Berwick and Weinberg, 1984; Frazier and Clifton, 1989, 1996; Gibson et al., 1996; Gibson, 1998; Sturt et al., 1999, 2000). Although these accounts vary widely in their details, they might all be called *search-based* accounts of syntactic retrieval. The core claims of search-based accounts include (i) the parser distinguishes local vs. distant syntactic domains through positional syntactic information, and (ii) the search for a potential syntactic dependent proceeds by first searching within some local syntactic domain, the size of which may vary across theories.

Although it may appear that the claims of these search-based accounts are incompatible with the locality account advanced by cue-based parsing models, this is not so. The core claims of search-based accounts may in fact be integrated with cue-based retrieval models if we suppose that positional syntactic information is available to guide retrieval operations, and that at retrieval the parser uses this positional information to limit retrieval to a local syntactic domain. We call this the *Local Search* hypothesis:

> *Local Search hypothesis*: The parser uses positional syntactic information during the retrieval of syntactic dependents, and positional cues serve to restrict retrieval to constituents in some local syntactic domain. (2)

According to the Local Search hypothesis, locality effects in sentence processing reflect in part a parsing strategy that prioritizes the retrieval of syntactically local constituents. In other words, the Local Search hypothesis claims that locality effects in sentence processing do not merely emerge from effects of decay and interference, but instead they reflect a strategy for the retrieval of syntactic dependents.

Within existing cue-based parsing models, it is not generally assumed that positional syntactic information is available to guide retrieval. Lewis and Vasishth (2005) propose that positional syntactic information, either hierarchical or linear, plays no role in the memory retrieval operations that guide attachment operations (cf. the *no serial order* hypothesis; see also Lewis et al., 2006). For example, while the parser may be able to use item information such as case to identify the subject of a sentence, it cannot use positional syntactic information to distinguish the encoding of the local subject from a more distant subject. On this account, the inability to distinguish distant and local subjects on the basis of their syntactic position provides an explanation of the well-known center embedding difficulty in terms of retrieval difficulty: when the parser needs to retrieve a subject for an embedded verb, there are too many similar subject encodings in memory to permit the parser to retrieve the grammatically appropriate subject, and positional information cannot be recruited to help with this process (Lewis and Vasishth, 2005; Lewis et al., 2006). In contrast to the Local Search hypothesis, these models claim that locality effects in sentence processing are entirely reducible to the effects of interference and temporal decay.

This claim is compatible with the SAT results reviewed above, which suggest that neither the linear nor structural position of the retrieval target directly impacts retrieval times. It is tempting to conclude from this that positional syntactic information is not used to guide syntactic retrieval operations. However, this conclusion would be premature on the basis of these data alone. These studies largely investigated configurations where there was only one grammatically licit position that could serve as the target of the retrieval. For instance, when the verb initiates the retrieval for the clefted object filler in (1), *wh*-feature cues could unambiguously select the filler as the target of retrieval. For this

reason, they provide no direct test of whether syntactic position information plays a role in helping the retrieval mechanism to distinguish between multiple, syntactically accessible targets in different syntactic domains.

A potential exception is Martin and McElree (2011), who investigated the processing of sluiced sentences as in (3):

a. Distant VP: Michael drank coffee and typed something, but he didn't tell me **what**.
b. Recent VP: Michael typed something and drank coffee, but he didn't tell me **what**. (3)

Martin and McElree hypothesized that the processing of the sluiced *wh*-phrase *what* requires comprehenders to retrieve an antecedent VP from memory, which is then used to construct the elided clause (i.e., the IP) at the sluice site. Martin and McElree manipulated the distance between the antecedent VP (*typed something* in 3) and the sluiced *wh*-phrase. In addition, they attempted to manipulate the size of the antecedent search set by manipulating whether a competitor VP was present in a coordination structure (*drank coffee* in 3). Nonetheless, they observed that only accuracy, not processing speed, was negatively impacted by the presence of multiple VP antecedents. They additionally observed that the recency of the antecedent VP only impacted retrieval accuracy. They argued that this pattern of results provided a strong data point in favor of content-addressable retrieval operations over syntactically structured search operations. However, because the two candidate antecedents were coordinated, this study may not have effectively manipulated the syntactic locality of the antecedent VP. Both potential antecedents were in a structurally similar position in the preceding clause. For this reason, this study leaves unresolved the question of how structural locality impacts retrieval.

THE CURRENT STUDY

In the present study we evaluate the Local Search hypothesis by investigating the processing of the Mandarin Chinese long-distance reflexive *ziji*. The anaphor *ziji* is an example of the cross-linguistically well-attested class of long-distance reflexives, reflexive pronouns that may be bound outside of their local clause. Thus, unlike the English reflexives *himself* and *herself*, *ziji* does not require that its antecedent be in the same clause, as seen in (4), where subscript indices are used to indicate possible coreference:

Zhangsan$_j$ *shuo Lisi*$_i$ *nongshang-le ziji*$_{i/j}$
Zhangsan says Lisi harm-perf self
"Zhangsan says that Lisi harmed him/herself" (4)

In (4), it can be seen that *ziji* may be bound either by the local subject *Lisi* or the matrix subject *Zhangsan*. Like many long-distance reflexives, *ziji* imposes a number of constraints on potential antecedents (Büring, 2005; Huang et al., 2009). There are significant syntactic constraints placed on antecedents: they must be subjects whose clausal projection dominates the clause that contains *ziji* (Huang and Liu, 2001). In addition to these syntactic constraints, there are a number of discourse-pragmatic constraints on the use of *ziji*. Antecedents must be animate and sentient, and must be prominent in the current discourse (Xue et al., 1994; Huang and Liu, 2001). In the absence of an appropriate antecedent in the immediate sentential context, *ziji* has been claimed to refer to the speaker, presumably as a reflex of the prominent discourse status that is automatically afforded to the speaker (Kuno, 1972; Huang and Liu, 2001). Though there are ongoing debates about the exact nature of *ziji*'s licensing conditions (Huang et al., 2006), it is uncontroversial that resolving the antecedent-anaphor dependency requires the comprehender to systematically exclude structurally unacceptable referents from consideration.

Although *ziji* can in principle take either local or long-distance antecedents, previous research suggests that there is a preference for local antecedents over more distant antecedents in online comprehension. For example, Li and Zhou (2010) provide ERP evidence that long-distance binding of *ziji* elicits a larger P300/600 response relative to local or ambiguous binding of *ziji*, suggesting greater processing difficulty associated with recovering long-distance interpretations. In addition, cross-modal priming studies have shown that probes associated with local antecedents are recognized more quickly than probes to long-distance antecedents upon encountering *ziji* (Gao et al., 2005; Liu, 2009). Chen et al. (2012) also present self-paced reading evidence that comprehenders read local *ziji*-antecedent dependencies more quickly than long-distance dependencies. These studies establish a preference for local binding over long-distance binding in comprehension, but without any direct time course evidence it is unclear whether this preference reflects a difference in retrieval speed or retrieval accuracy for local antecedents.

In this study we investigate the processing of local and long-distance interpretations of *ziji* as in (5). We take the embedded and matrix clauses in (4) to constitute distinct syntactic domains for the purposes of finding *ziji*'s antecedent.

a. *Zhangsan*$_i$ *shuo fengbao*$_j$ *hai-le ziji*$_{i/j}$
Zhangsan says storm harm-perf *ziji*
"Zhangsan said the storm harmed him."
b. *Xiaoshuo*$_i$ *shuo Zhangsan*$_j$ *hai-le ziji*$_{i/j}$
novel says Zhangsan harm-perf *ziji*
"The novel said Zhangsan harmed himself." (5)

Because *ziji* requires an animate and sentient antecedent, only *Zhangsan* in (5a,b) is a grammatically licensed antecedent. Of critical interest is the long-distance configuration (5a), where the local subject is inanimate and thus semantically inappropriate as an antecedent for *ziji*. The critical empirical question in this comparison is whether comprehenders will show delayed access to the matrix antecedent in (5a). If the match on semantic cues outweighs the effect of dependency locality, and if it grants reliable direct access to the matrix antecedent, then the only difference between (5a) and (5b) should be the amount of time the antecedent has decayed in memory. Previous findings show that decay alone does not impact the speed of retrieval (see e.g., McElree et al., 2003). Thus, if semantic cue match outweighs the effect of locality in this configuration, we predict no difference

in retrieval speeds between local and long-distance interpretations of *ziji*. Instead, we should see only a difference in processing accuracy between the two configurations in (5), such that local antecedents are more accurately retrieved.

The Local Search hypothesis, however, does predict a difference in retrieval speeds. If the parser initially uses cues that limit retrieval to the local clausal domain, then on a significant portion of trials the comprehender should misretrieve the local subject in (5a) despite its poor fit to *ziji*'s semantic cues. This would require the comprehender to engage costly reanalysis processes to recover the more distant antecedent, leading to slowed retrieval times in (5a). In Experiment 1, we used a variant of the SAT technique known as multiple response SAT (MR-SAT) to estimate the speed of processing *ziji* in these two configurations to determine whether local and long-distance *ziji* dependencies are associated with different retrieval speeds.

EXPERIMENT 1

Experiment 1 employed the multiple-response speed-accuracy tradeoff procedure (MR-SAT; Wickelgren et al., 1980) to estimate the time course of retrieving an antecedent for *ziji* in sentences such as (5). MR-SAT is an attractive technique to use in studying language comprehension because it dissociates processing speed from processing accuracy (McElree, 2000; McElree et al., 2003; Foraker and McElree, 2007; Martin and McElree, 2008, 2009, 2011). In a MR-SAT paradigm, participants are required to make acceptability judgments at pre-specified response latencies. This provides a measure of how accuracy grows over time, and thus provides a direct measure of the time course of processing. In contrast, single RT paradigms are limited in how informative they are about time course of processing. Because participants can trade speed and accuracy in many standard judgment tasks (Wickelgren, 1977), merely estimating a point RT per condition (or a single RT/accuracy pair) can obscure differences between the probability of successfully completing a process and the speed with which that process reaches completion. In contrast, the full time course summarized in an SAT function allows the researcher to separately estimate the speed and the accuracy of memory retrieval. In the present case, we are concerned with the nature of any difficulty observed with non-local *ziji* interpretations as in (5a). Prior work suggests that retrieval difficulty associated with temporal decay or linear distance is associated with a decrease in retrieval accuracy, rather than retrieval speed (McElree, 2000; McElree et al., 2003; Foraker and McElree, 2007; Martin and McElree, 2008, 2009). Based on these results, we do not expect to observe differences in retrieval speed purely as a function of decay or recency.

METHOD

Participants

Twenty college students from Beijing Normal University participated in the experiment. Data from 3 participants were excluded for reasons that are detailed below. The remaining 17 participants included 10 females, and had a mean age of 23.5 years. Each participant completed six 1-h experimental sessions spaced at least a day apart, in addition to a 1-h practice session for familiarization with the MR-SAT procedure. All participants were native Mandarin Chinese speakers and had normal or corrected-to-normal vision. Following an IRB-approved protocol, all participants gave informed consent and were paid 35 RMB per hour for their participation in the experiment.

Materials

Our critical experimental materials were Mandarin sentences that contained a main clause verb that selects for a sentential complement (e.g., "*biaoshi*," say). The embedded complement clause was always transitive, and the embedded object was always the sentence-final word. Two features of the stimuli were manipulated orthogonally, in a crossed 3 × 3 experimental design. One was the position of a syntactically prominent animate subject; it was either the subject of the main clause (*long distance animate)*, the subject of the local (embedded) clause (*local animate*), or not present (*no antecedent*). In addition we manipulated the form of the embedded object NP, which was either the long-distance reflexive *ziji*, a contextually plausible definite NP, or a contextually implausible definite NP.

Four of the nine resulting conditions formed the critical experimental conditions (**Table 1**). Based on the position of the animate subject, *ziji* either took a long-distance antecedent (Long-distance animate, *ziji* condition) or a local antecedent (Local animate, *ziji* condition). In the control conditions *ziji* was replaced with a full NP that was a plausible object of the embedded verb (e.g., *the batsman*). The inclusion of these control conditions helps to ensure that any differences in processing speed or accuracy observed in the critical *ziji* conditions are specific to retrieval processes associated with *ziji,* rather than other properties of the sentence frame. In the critical experimental conditions, sentences were acceptable across all four conditions.

In the local *ziji* and the corresponding control conditions, the main clause subject NP was always an inanimate noun that described a form of written or spoken media (e.g., *book*, *documentary*, *memo*) to ensure compatibility with the meaning of the main clause verbs (e.g., *say*) while being an unacceptable antecedent of *ziji*. None of the inanimates used in any position could be construed metonymically; metonymic interpretations of inanimates (i.e., *the newspaper* being used to refer to the employees of the newspaper) may be used as antecedents for *ziji.* In order to ensure that participants do not have ceiling performance in our task (McElree, 2006), a temporal adverbial clause was interpolated between the embedded subject and the embedded verb. In all conditions, an animate NP was used as the subject of the temporal adverbial phrase. However, since it occupied a position that is not structurally higher than *ziji,* it is not a grammatical antecedent for *ziji*.

In addition to these critical four conditions, the implausible object conditions contained a contextually implausible embedded object (e.g., "The auto-biography says that the coach underestimated *the glasses* when the team was doing poorly.") and the no animate conditions did not contain an animate NP in either the matrix or embedded subject position (e.g., "The auto-biography says that the report underestimated *ziji* when the team was doing poorly."). These extra conditions provided unacceptable counterparts to the critical conditions, either because of a

Table 1 | Summary of the critical conditions in Experiment 1.

Condition	Example
Long-distance animate, *ziji*	张教练 表明 那篇报导 [在 团队 未能 发挥 水准 的时候] 低估了 **自己** Coach Zhang say [that report (when team not perform well-time) underestimate **ziji**] "Coach Zhang says that that report underestimated ***self*** when the team was doing poorly."
Local animate, *ziji*	回忆录 表明 张教练 [在 团队 未能 发挥 水准 的时候] 低估了 **自己** Auto-biography say [coach Zhang (when team not perform well-time) underestimate **ziji**] "The auto-biography says that coach Zhang underestimated ***self*** when the team was doing poorly."
Long-distance animate, control	张教练 表明 那篇报导 [在 团队 未能 发挥 水准 的时候] 低估了 **那位击球手** Coach Zhang say [that report (when team not perform well-time) underestimate **that batsman**] "Coach Zhang says that that report underestimated ***the batsman*** when the team was doing poorly."
Local animate, control	回忆录 表明 张教练 [在 团队 未能 发挥 水准 的时候] 低估了 **那位击球手** Auto-biography say [coach Zhang (when team not perform well-time) underestimate **that batsman**] "The auto-biography says that coach Zhang underestimated ***the batsman*** when the team was doing poorly."

local implausibility, or because *ziji* did not have an antecedent available. There were two reasons for including these additional conditions, despite the fact that they were not part of the primary experimental manipulation. First, they provided unacceptable sentences that could be used in *d*-prime scaling. More importantly, the inclusion of the implausible object and no antecedent *ziji* conditions ensured that within the experiment neither the presence of *ziji* nor the acceptability of the sentence was predictable from the sentence context. Because it is typical for a subject in an SAT experiment to see all conditions of an experimental item, the inclusion of these additional conditions was critical to ensure that participants could not use familiarity with the sentence context to anticipate their response in advance of the sentence final word.

Forty sets of the 9 sentence types (5 acceptable and 4 unacceptable) were generated. The resulting 360 sentences were equally distributed in 6 presentation lists, one for each of the 6 sessions, to minimize the repetition of content material within a session. Thus, across the six sessions, each participant saw each experimental item in each of its 9 conditions. Crucially, no two instances of *ziji* sentences from the same item set were presented within the same session. Within a session, each participant viewed 206 sentences, of which 60 were drawn from the current study. Since only one third of target sentences contained *ziji*, the critical *ziji* conditions comprised around 10% of all sentences within and across sessions. The order of presentation within a session was randomized.

Procedure

Stimulus presentation, timing, and response collection were all carried out on a personal computer using the Linger software (Rohde, 2003). Each trial began with a 500 ms fixation cross presented in the center of the screen. Each word appeared in the center of the screen for 400 ms, followed by 200 ms of blank screen. All words were presented using simplified Chinese characters, and the last word of each sentence was marked with a period (。). At the onset of the final word, a series of 18 auditory response cues (50 ms, 1000 Hz tone) was initiated. The cues occurred every 350 ms, and the final word of the sentence remained on the screen. Participants were asked to decide for each sentence whether it was an acceptable, coherent sounding sentence or not (in Mandarin: *tōngshùn he héshì)*. Participants were trained to initially respond by pressing both response keys simultaneously to indicate an undecided response, and to respond at every tone. They were then trained to switch their response to either the "accept" or "reject" key as soon as they could. Importantly, they were also trained to modify their responses if their assessment changed. During the 1-h practice session, participants were told that some of the sentences were complex, but nevertheless were meaningful sentences, and explicit feedback was given about acceptable and unacceptable sentences in the experiment. Each participant performed six 1-h experimental sessions, and in each they saw one of the lists of materials. The order of lists was randomized across participants.

Data analysis

To derive the full time-course information, d' scores were calculated by comparing an acceptable and an unacceptable condition at each of the response tones. The resultant series of d'-values at each time point t was fit using a shifted exponential function:

$$d' = \lambda\left(1 - e^{-\beta(t-\delta)}\right), t > \delta,$$
$$d' = 0 \quad , otherwise \tag{6}$$

The SAT function in Equation (6) describes the growth of accuracy over time using three parameters: *asymptote* (λ), *rate* (β), and *intercept* (δ). By regressing the non-linear SAT function against the time course data collected in the experiment, we may make inferences about the effect of experimental manipulations on each of the parameters. The initial period of chance performance is described by the intercept parameter (δ), which indicates the point at which the SAT function departs from chance performance (0 in d' units). The next portion of the function is characterized by a period of increasing accuracy; the rate of growth in this portion of the SAT function is described by the

rate parameter (β). The last portion of the function reflects terminal accuracy in the behavioral judgment, and it is reflected in the asymptote parameter of the SAT function (λ). The intercept and the rate together index the speed of the process, while the asymptote indexes the terminal accuracy of the process. The processing speed may also be evaluated by considering a composite measure known as the *speed* of the SAT function ($\beta^{-1} + \delta$). By parameterizing the SAT function in this way, we can separately estimate the speed of processing (as reflected in the intercept, rate, or speed measures) and the accuracy of processing (as reflected in the asymptote). Differences in the intercept or rate parameters indicate a difference in processing speed between two conditions; differences in the asymptote parameter indicate a difference in processing accuracy.

d' is the standard measure of discrimination (assuming equal-variance Gaussian distributions): $d' = \Phi(\text{hits}) - \Phi(\text{false alarms})$ (Macmillan and Creelman, 2004; Φ represents the inverse of the cumulative distribution function of the standard normal). However, in the models reported below, we only report a *pseudo* d' measure that does not correct for the false alarm rate [$d' = \Phi(\text{hits})$]. We adopted this analysis because the somewhat high acceptability of the no antecedent condition (see below) made it inappropriate for the construction of a discriminative d' measure. For reference, a pseudo d' score of 2.5 represents perfect performance in our experiment, and a pseudo d' score of 0.84 represents a hit rate of 0.80.

It is important to note that our pseudo d' measure does not correct for any response bias that participants may have. In this respect, our analysis differs significantly from the approach adopted in previous SAT work, which has generally used d' as the dependent measure to ensure that any time course differences are not simply due to differences in response bias across conditions. However, we note that our critical conditions constitute a 2×2 crossed factorial design (presence of *ziji* by position of animate antecedent). This design allows us to account for any response bias introduced by two major features of our stimuli: the configuration of the sentence prior to the critical final word, and the presence of *ziji* in final position. If response bias varies as a function of the sentence context, then this bias should be shared by both *ziji* and control conditions. Likewise, if there is response bias associated with a sentence final *ziji*, as opposed to a sentence final lexical NP, then this bias should be shared by both *ziji* conditions. Thus, any interactions of *ziji* and position of the animate subject in our design cannot be the result of response bias introduced by either of these two configurations.

Data analysis proceeded in two steps: a model selection analysis and a parameter estimation analysis (Liu and Smith, 2009). In the model selection analysis, the best fit SAT model was determined using the adjusted R^2-statistic (in Equation 7) using a hierarchical model-testing scheme over the averaged data, an approach pursued in prior work on SAT in sentence comprehension (McElree, 2000; McElree et al., 2003; Foraker and McElree, 2007; Martin and McElree, 2008, 2009). However, we note that for multiple-response SAT, determining the number of independent data points n is not a trivial problem, because of the lack of independence between responses on any trial. Because of the uncertainty concerning the number of truly independent data points that underlie any one MR-SAT function, it is difficult to straightforwardly apply model fitting metrics such as adjusted R^2, the AIC, and the BIC. In the parameter estimation analyses, only fully saturated models that allow all parameters to vary by condition are considered, and any differences between the critical conditions on the parameters of interest are assessed using familiar hypothesis testing measures over individual parameter estimates. This analysis follows the recommendations of Lorch and Myers (1990) for dealing with regression analyses in the context of a repeated measures experiment. In order to obtain parameter estimates, we used the R statistical computing environment to fit non-linear regressions of the SAT function Equation (6) against the pseudo d' score (see McElree and Griffith, 1998; McElree, 2000; McElree et al., 2003; Martin and McElree, 2008, 2009, 2011). We used the `nls()` function with an adaptive non-linear least squares algorithm (Dennis et al., 1981) to determine the least squares fit of the SAT function to the data.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (d_i - \hat{d}_i)^2/(n-k)}{\sum_{i=1}^{n} (d_i - \bar{d})^2/(n-k)} \tag{7}$$

Prior to modeling the d' scores, analysis was performed on empirical pseudo d' measures by participants. This was obtained by taking the average rate of acceptance over the last four response points in each condition to determine the empirical hit rate, and calculating d' as described above. Hit rates that reflected perfect performance were smoothed by subtracting 0.0125 from the hit rate [1/(2N) smoothing (Macmillan and Creelman, 2004)].

Where appropriate, behavioral measures and parameter estimates from the SAT function in Equation (6) were further analyzed by entering them into a 2×2 repeated-measures ANOVA crossing dependency type (*ziji* vs. *control*) and the position of the animate argument (*long-distance* vs. *local*).

Of the twenty participants run, data from two participants were excluded due to unreliable dynamics estimates. The empirical d' scores from these participants appeared better fit by a sigmoidal rather than exponential function, leading to unrealistically large and unreliable differences in the critical conditions in the crucial intercept and rate parameters when fit with the SAT function in Equation (6). The data from one further participant were rejected due to lower than 60% correct responses on both critical *ziji* conditions. These participants' empirical d' were not included in any analyses below.

RESULTS

Accuracy and empirical d′ analysis

For the four critical experimental conditions, acceptance rates were high. Average acceptance was 87% for long-distance *ziji* conditions and 83% for local *ziji* conditions. The rates of acceptance for the long distance and local control conditions were 91 and 88%, respectively.

In contrast, the average acceptance was 47% percent for no antecedent *ziji* conditions, and the unacceptable control conditions each had an average acceptance rate of 2%. In addition, the rate of acceptance of the no antecedent acceptable control condition was 92%.

Table 2 presents the mean empirical pseudo d' for *ziji* and control conditions. The data were analyzed using a 2×2 repeated-measures ANOVA with dependency type and animate position as within-participant factors. This analysis revealed a marginal main effect of position of animate argument [$F_{(1,\ 16)} = 3.6$, $p < 0.1$], as well as a marginal effect of dependency type [$F_{(1,\ 16)} = 4.3$, $p < 0.1$]. The interaction of animate position and dependency type was not significant ($F < 0.1$). However, planned comparisons between the long-distance and local conditions within *ziji* and control conditions did not reveal any reliable effects [*ziji*: $t_{(16)} = 1.02$, $p = 0.32$; control: $t_{(16)} = 1.6$, $p = 0.12$].

Time course analysis

Competitive fits of the shifted exponential function in Equation (6) were conducted to assess differences in asymptotic accuracy, rate, and intercept across conditions for each participant. Model fits were conducted separately for control and *ziji* conditions. Because the empirical d' analysis revealed only marginal differences between conditions in accuracy, it is not clear whether competitive model fits are justified in allowing the asymptote parameter to vary freely between conditions. In light of this, we fitted two SAT functions to each data set: one model whose asymptote parameter was fixed to the empirical pseudo d' obtained by averaging over the final four response latencies, and one where the asymptote parameter was allowed to vary. We report results from the free parameter models, but we note that fitting the models with fixed asymptotes did not yield qualitatively different results.

Model-fitting analyses pitted nested models against each other on adjusted R^2 Equation (7), following McElree et al. (2003) and Liu and Smith (2009). Fits to the across-participants average for the critical *ziji* conditions revealed a small advantage for models that allocated separate intercept parameters (δ) for local and long-distance conditions (2λ-1β-2δ, R^2:0.986) and models that models that posit separate rate (β) parameters (2λ-2β-1δ, R^2:0.985) over models that posited shared rate and intercept parameters for the two conditions (2λ-1β-1δ, R^2:0.982). This difference reflected a small rate advantage for local *ziji* condition over long-distance *ziji* condition (LD β: 0.96 s^{-1}, Local β: 1.26 s^{-1}; LD δ: 0.75 s, Local δ: 0.58 s). These models were in turn a better fit to the data than any model that contained only a single asymptote for both conditions (max $R^2 = 0.974$). Control conditions showed no improvement in fit for additional rate or intercept parameters (2λ-1β-1δ, R^2:0.996; 2λ-2β-1δ, R^2:0.996; 2λ-1β-2δ, R^2:0.996). The average data for *ziji* and control conditions, along with best-fit models on the adjusted R^2 metric, are presented in **Figure 1**.

Of critical interest is whether the fits to the average data reflect a reliable trend across individuals. It is possible that the SAT function reflected in the average is not in fact representative of a pattern observed in any individual subject (Liu and Smith, 2009), and in the present case, there was a very small difference between models with different dynamics parameters and those without. In order to assess the reliability of parameter estimates across participants, each individual's d' data were fit with the SAT function separately for each of the four critical conditions. As before, fits were conducted both with fixed and free asymptote parameters, and these two types of models did not yield qualitatively different results. Thus, we report only the results from models

Table 2 | Mean empirical pseudo d'-values, obtained by averaging accuracies over final four response latencies.

	ziji	Control
Long-distance animate	1.26 (±0.12)	1.53 (±0.12)
local animate	1.11 (±0.13)	1.33 (±0.13)

By-participant standard error in parentheses.

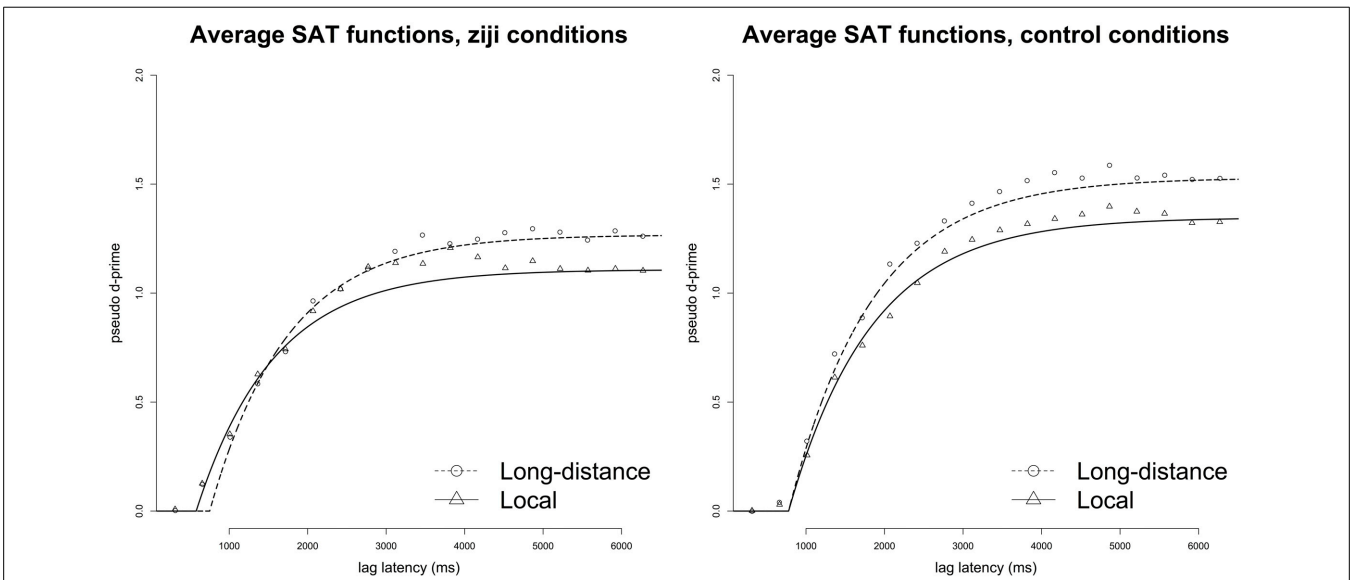


FIGURE 1 | Time course data (points) and best-fit SAT functions (lines) to average pseudo d' scores in Experiment 1, for *ziji* and control conditions.

with free asymptote parameters. Fits to individual participants revealed differences in whether dynamics differences between the critical conditions were reflected in the SAT function's rate, its intercept, or both. Because of these differences, here we additionally present and analyze the speed measure ($\beta^{-1} + \delta$). This composite measure allows us to quantify processing speed in a uniform way across individuals in the face of this variation. The results of this analysis for the critical *ziji* conditions are presented in **Table 3**.

The individual parameter estimates were entered into a 2 × 2 repeated-measures ANOVA with dependency type and animate position as within-participant factors. ANOVAs revealed an interaction of dependency type and animate position both for rate parameters β [$F_{(1,\ 16)} = 4.8,\ p < 0.05$] and for the composite speed measure $\beta^{-1} + \delta$ [$F_{(1,\ 16)} = 8.2,\ p < 0.05$]. In addition, there was a main effect of dependency type on speed measures [$F_{(1,\ 16)} = 6.1,\ p < 0.05$] and on rate parameters [$F_{(1,\ 16)} = 3.6,\ p < 0.1$], reflecting faster processing of *ziji* conditions. Additionally, a significant effect of dependency type was observed for the asymptote parameter λ [$F_{(1,\ 16)} = 4.9,\ p < 0.05$]. There were no significant effects on the intercept parameter δ for either fixed or free asymptote models.

Planned pairwise comparisons revealed that these interactions were driven by differences in the *ziji* conditions in the speed measure $\beta^{-1} + \delta$ [$t_{(16)} = -2.6, p < 0.05$]. This analysis revealed only marginal effects on the rate parameter β [$t_{(16)} = 1.9,\ p < 0.1$]. There were no effects of antecedent position in the control conditions for either speed [$t_{(16)} = 1.6, p = 0.12$] or rate [$t_{(16)} = -0.1$, $p = 0.89$]. On average, the speed measure for local *ziji* condition was 294ms faster (95% CI: 52–538 ms) than long-distance *ziji* condition.

DISCUSSION

Both individual and average data suggest a time-course advantage for local *ziji* conditions compared to LD *ziji* conditions. In competitive model fits to the average time course data, there was a slight advantage for models that allocated different rate parameters to long-distance and local *ziji* conditions; no such difference was observed for control comparisons. An analysis of the parameter estimates for individual participants showed that for the critical *ziji* comparison, the local condition was processed significantly faster than the long-distance condition, as reflected in the rate parameter (β) and the composite speed measure ($\beta^{-1} + \delta$). No difference was observed in control conditions. In ANOVA analyses of empirical d' scores, there was no significant difference for either *ziji* or control comparisons in asymptotic accuracy.

Follow-up experiment

One unexpected finding in Experiment 1 was the high acceptance rate of the no antecedent *ziji* condition, which participants accepted on 47% of trials. It has been claimed that in the absence of an overt, syntactically prominent antecedent, *ziji* can refer to the speaker (Huang and Liu, 2001). However, post-experiment debriefing suggested that some speakers also interpreted *ziji* as coreferential with an implicit author of the inanimate main clause subjects such as *book* or *speech*. It is also possible that the subject contained in the temporal adjunct in our experimental sentences contributed to retrieval interference, and was misinterpreted as

Table 3 | By-subject and average parameter estimates for critical ziji comparisons, along with average parameter estimates for control comparisons.

	LD R^2	Loc R^2	Asymptote (λ, d')		Rate (β, s^{-1})		Intercept (δ, s)		Speed (β^{-1}+δ, s)	
			LD	Local	LD	Local	LD	Local	LD	Local
Average	0.99	0.98	1.30	1.15	0.95	1.27	0.72	0.74	1.76	1.53
S1	0.98	0.96	0.95	0.72	3.51	15.00	0.29	0.30	0.58	0.37
S2	0.75	0.94	1.66	1.55	2.81	3.90	0.93	0.89	1.29	1.15
S3	0.97	0.98	2.00	0.95	0.72	1.38	0.88	0.96	2.27	1.69
S4	0.99	0.86	0.95	1.75	3.87	1.95	0.65	0.68	0.90	1.20
S5	1.00	0.98	0.68	0.30	0.86	15.00	0.28	0.54	1.44	0.61
S6	0.99	0.98	0.93	0.59	1.36	3.57	1.08	1.25	1.81	1.53
S7	0.95	0.96	1.89	2.21	2.08	1.77	0.57	0.58	1.05	1.15
S8	1.00	0.93	1.05	2.26	3.81	0.97	1.58	1.26	1.84	2.29
S9	0.91	0.96	1.76	0.44	1.09	4.65	1.66	1.92	2.58	2.14
S10	0.97	0.97	0.57	0.99	1.67	3.27	1.18	1.44	1.78	1.75
S11	0.98	0.99	1.65	0.97	0.60	1.49	0.15	0.28	1.81	0.95
S12	0.94	0.87	2.10	1.64	0.89	1.84	1.65	1.64	2.77	2.18
S13	1.00	0.99	0.40	0.76	1.26	1.33	1.73	0.99	2.52	1.74
S14	0.99	0.99	0.77	0.48	0.87	1.60	0.82	0.26	1.96	0.88
S15	0.98	0.92	1.38	1.71	2.14	1.23	1.00	1.19	1.47	2.00
S16	0.99	0.99	1.37	0.62	1.06	1.14	0.77	0.76	1.71	1.64
S17	0.98	0.99	2.03	1.67	1.01	3.32	0.60	0.81	1.59	1.11
Control	0.99	0.99	1.57	1.39	0.97	0.91	0.77	0.79	1.80	1.89

an antecedent for *ziji* on some trials. We conducted a follow-up experiment to determine the interpretations that comprehenders might assign to each of our three conditions.

The follow-up experiment used the three *ziji* conditions from Experiment 1. Twenty-four of the original 40 item sets were selected at random, and were distributed in a Latin Square fashion into three experimental lists. Each list was presented as a short questionnaire on the online experimental platform IbexFarm (Drummond, 2011). Each sentence was presented on the screen, and participants were instructed to choose their preferred interpretation of *ziji*'s antecedent from five options: the *main clause* subject, the *local* subject, the *interfering* subject contained inside the temporal adjunct, the *speaker* of the sentence, or *none*. When the main clause subject was inanimate (e.g., *book*), the main clause subject response option referred to the implicit author (e.g., *the book's author*).

Seventeen native Mandarin speakers were recruited via the Internet. The results are presented in **Table 4**. In order to test for differences across conditions in the proportion of responses, each response category was converted into a binary variable that was 1 if a response was in the category, and 0 otherwise. Each response category was analyzed using logistic mixed effects models with crossed random intercepts for subjects and items and random slopes of condition for subjects. Two Helmert contrasts were employed for the condition factor: a *locality* contrast that compared local and LD conditions, and an *antecedent* contrast that contrasted the no antecedent condition with the average of the LD and local conditions.

This analysis revealed that the *no antecedent* condition had significantly more *none* responses ($\beta = 0.64$, Wald's $z = 3.4$, $p < 0.05$) and *interferer* responses ($\beta = 1.53$, Wald's $z = 6.1$, $p < 0.05$) than the other two conditions. The LD condition had significantly more matrix subject responses than the local conditions ($\beta = -3.5$, Wald's $z = -8.0$, $p < 0.05$), and the local condition had significantly more embedded subject responses than the LD condition ($\beta = 3.4$, Wald's $z = 8.8$, $p < 0.05$). In addition, there was a significant effect of *antecedent* on embedded subject responses ($\beta = -0.9$, Wald's $z = -5.2$, $p < 0.05$), reflecting the low proportion of local subject responses in the no antecedent condition. No other effects were significant.

The follow-up experiment confirms that participants overwhelmingly select a structurally prominent, animate antecedent for *ziji* when there is one available. This replicates the judgments reported in the literature on Mandarin long-distance reflexives. Additionally, the results show that in the absence of a semantically appropriate and syntactically accessible antecedent, comprehenders nonetheless ultimately prefer a sentence-internal antecedent: the interfering subject is selected on 31% of trials, and on 35% of trials participants coerce an animate antecedent from the matrix subject.

Table 4 | Average proportion of interpretations reported on critical ziji comparisons in the follow-up experiment.

	Long-distance	Local	Interfering	Speaker	None
LD antecedent	0.87	0.06	0.03	0.00	0.04
Local antecedent	0.03	0.91	0.03	0.00	0.03
No antecedent	0.35	0.12	0.31	0.04	0.18

DISCUSSION

The results of the follow-up experiment confirm that comprehenders prefer to select syntactically prominent, animate antecedents for *ziji* in our materials. The results of Experiment 1 show that comprehenders are measurably slower to access long-distance antecedents for *ziji* than local antecedents. The fact that dependency distance impacted retrieval speed in our SAT experiment contrasts with previous SAT findings. Previous work on SAT in language comprehension suggests that distance does not affect the dynamics parameters in the SAT function (McElree, 2000; McElree et al., 2003; Foraker and McElree, 2007; Martin and McElree, 2008, 2009). This makes it unlikely that the faster access times we observe to local antecedents reflect a simple effect of temporal distance or recency.

Another way that long-distance and local antecedent configurations differ is in the type of interference contributed by the semantically inappropriate antecedent. In long-distance conditions, the semantically inappropriate antecedent intervenes between the target antecedent and the anaphor, and so generates retroactive interference (RI) in the process of retrieving the target antecedent. Conversely, in local antecedent configurations, the long-distance antecedent precedes the target, and so generates proactive interference (PI) that may disrupt the anaphor's retrieval of its antecedent. The difference in the type of interference created by the semantically inappropriate antecedent may be critical: Öztekin and McElree (2007) observed that in recognition memory tasks, the presence of PI has an effect on retrieval dynamics, leading to slower retrieval times. However, recent SAT work has directly investigated the effects of PI and RI on retrieval processes in language comprehension (Van Dyke and McElree, 2011). Van Dyke and McElree (2011) suggest that RI contributes more difficulty in dependency completion in sentence comprehension than does PI, but crucially, they show that the type of interference (PI/RI) does not impact retrieval speeds in multiple-response SAT. Instead, they observe only that RI configurations lower asymptotic accuracy relative to PI configurations. In light of these results, it appears unlikely that the speed differences that we observed were due to the type of interference generated by the inappropriate antecedent.

A MODEL OF THE LOCAL SEARCH HYPOTHESIS

We have suggested that neither recency alone nor the type of interference (RI/PI) was the source of the observed differences in retrieval times in Experiment 1. Instead, we argue that these results suggest that comprehenders consider or misretrieve the local subject position when the target antecedent is syntactically distant, which then leads to slowed retrieval of long-distance antecedents. We propose that this arises because locality outweighs semantic cues when retrieving an antecedent for *ziji*. There are two potential explanations of this locality effect in our data. According to the Local Search hypothesis, this effect reflects the use of cues that restrict retrieval operations to a local syntactic domain. However, it is possible that this locality effect reflects more misretrievals of a semantically inappropriate local subject

simply because it has a relatively high resting activation. There are two reasons to suspect that the local subject might have higher resting activation prior to reaching the anaphor *ziji*. First, it is more recent, and so will have undergone less temporal decay. Second, the embedded subject forms a dependency with the verb that precedes *ziji*. The process of retrieving the subject to form this dependency may boost the embedded subject's resting activation prior to encountering the anaphor.

To distinguish between a Local Search account and an account that attributes the slowed processing of LD *ziji* to the heightened activation of the local subject, we formalize the predictions of both accounts with a simple quantitative model of the antecedent retrieval process for *ziji*. Our model incorporates the declarative memory component of the ACT-R framework (Anderson and Lebiere, 1998), which implements a direct access cue-based retrieval process that is subject to temporal decay and retrieval interference. An attractive feature of this model is that it has been used in a number of successful models of cue-based parsers (Lewis and Vasishth, 2005; Lewis et al., 2006; Vasishth et al., 2008). Our goal in modeling the antecedent retrieval process using ACT-R is to estimate the effect of a local search strategy on SAT retrieval dynamics above and beyond the effects of interference and decay.

We define our retrieval models in terms of the set of cues (the retrieval probe) used to retrieve an antecedent from memory. The *unrestricted* retrieval model limits the probe to item information only. In our implementation, this includes category identity (*NP*), a case feature (+*Nominative*), which serves to identify subjects, and an animacy feature (+*Animate*). The latter two cues implement the syntactic and semantic constraints on *ziji*'s antecedent. The *local search* retrieval model includes these features plus a feature (+*Local*) that distinguishes the local clause from other clauses. In terms of our stimuli, this feature is used to distinguish the embedded clause from both the matrix clause and the adjunct clause. This feature implements the core claim of the Local Search hypothesis: that the parser uses positional information to restrict search to the local syntactic domain at retrieval, creating a retrieval process that explicitly prioritizes retrieval within a local syntactic domain (here taken to be the local clause).

Our model assumes that the process of finding *ziji*'s antecedent involves a series of serially executed, cue-based retrievals from a content-addressable memory store (consistent with the processing assumptions of Lewis and Vasishth, 2005). Once an item is retrieved, it is evaluated as the antecedent of *ziji*. If the retrieved item is rejected as an antecedent for *ziji*, then the processor samples another potential antecedent from the linguistic context, without replacement. We assume that the processor samples antecedents in this way until an appropriate antecedent is found. Under this model, cue match, temporal decay, and interference all influence the average number of sampling operations that are required to recover the correct antecedent for *ziji*. The more sampling operations are executed during the retrieval of an antecedent, the slower the speed of the SAT function that tracks this process.

To fit this model to the empirical data, we first determine the probability of successfully retrieving the target antecedent on each successive sampling operation. We determine these probabilities by simulation, using the equations that define declarative memory in ACT-R. The ACT-R component of the simulations reported below was developed by Badecker and Lewis (2007) using the R programming language (R Core Team, 2013). Under this model, the parser retrieves the item in memory with the highest activation value, where activation is a function of the match to retrieval cues and the resting activation of all items in memory. Formally, the activation of a memory item i (A_i) is the sum of its resting activation B_i, the match between the item and each of the J retrieval cues in the probe (S_j), and random noise (ϵ):

$$A_i = B_i + \sum_j W_j S_{ji} + \epsilon \tag{8}$$

The weight associated with each retrieval cue W_j is the total amount of goal activation available G divided by the number of retrieval cues. The resting activation of item i is a function of temporal decay (controlled by the decay parameter d) over all M intervals t_m since the item was last retrieved or created:

$$B_i = ln\left[\sum_m t_m^{-d}\right] \tag{9}$$

The match of an item i to the retrieval probe is the sum of a weighted associative boost for each cue S_j in the retrieval probe that matches the features of item i. The weight of a feature W_j is assumed to be equal across all cues in the probe. The associative boost that a given cue adds to an item it matches is reduced by the *fan* of that cue, or the number of items in memory that match that cue:

$$S_{ji} = S - \ln(fan_j) \tag{10}$$

Lastly, a small amount of stochastic noise is added to every item's activation level. On any given trial a noise value is drawn from a logistic distribution with a mean of zero and a variance that is controlled by a noise parameter s.

$$\epsilon \sim logistic(0, \sigma^2) \tag{11}$$

$$\sigma^2 = \frac{\pi^2}{3} s^2 \tag{12}$$

For all predictions reported below, we simulated the model's predictions on a range of parameter settings, and report the mean predicted values across all parameter settings (following the approach in Dillon et al., 2013). Our choices of possible parameter settings were based on the settings reported in Lewis and Vasishth (2005)[1]. One exception was the scaling parameter F, which was set to yield a mean retrieval time of 90 ms. This was chosen because it provided a close fit to the estimated retrieval time of 85 ms in the SAT paradigm found by McElree et al. (2003). The times between the creation

[1]The parameter settings considered were: $F = \{0.08, 0.10, 0.12\}$; $d = \{0.4, 0.5, 0.6\}$, $S = \{0.75, 1.0, 1.25\}$; $G = \{0.75, 1.0, 1.25\}$. Crossing all possible parameter settings resulted in 81 unique parameterizations of the model.

of antecedent representations and the retrieval associated with *ziji* were calculated directly from the experimental presentation parameters. In addition, an intermediate retrieval of the local subject at the embedded verb was simulated, which provided a boost in the embedded subject's resting activation prior to the point when *ziji* was encountered.

For both retrieval models, the probability of retrieving the target and each of the distractor NPs under these conditions was estimated using Monte Carlo simulation and averaging across all parameter settings considered. From this distribution we simulated the average number of sampling operations necessary to recover the target antecedent for both retrieval models. The resulting distributions are presented in **Table 5**. Under local search models, the local target is reliably retrieved after only one sampling operation on 56% of trials, whereas the modal number of sampling operations required to access the long-distance antecedent under the search model is 3 (occurring on 55% of trials). In the unrestricted search models, there is a lower probability of success with a single retrieval: local antecedents are retrieved on the first trial in 35% of trials for unrestricted models, and long distance antecedents are retrieved on 27% of trials. The lower probability of success for unrestricted models reflects the additional interference from the syntactically illicit distractor NP that occurs without positional cues to retrieval. On unrestricted models the number of sampling operations necessary to recover the target antecedent does not appear to differ substantially for local and long-distance antecedents. This pattern suggests that the increased resting activation of the local subject does not by itself lead to a substantially increased rate of retrieval errors during the retrieval of a long-distance antecedent. Instead, the search model results suggest that a semantically inappropriate local subject is most likely to be misretrieved when the search probe contains positional cues that select the local subject.

Next, we calculated the distribution of finishing times for the search process under the serial sampling model we have proposed. We simulated the distribution of finishing times for a retrieval process with n sampling iterations by simulating the sum of n retrievals from the ACT-R model given above. For retrievals beyond the first, an additional 50 ms was added, reflecting the additional processing necessary to evaluate the retrieved item[2]. In ACT-R, the retrieval latency T_i is a function of activation and a scaling parameter F (see Footnote 1):

$$T_i = Fe^{-Ai} \tag{13}$$

Table 5 | Probability distribution over the average number of sampling operations necessary to recover ziji's antecedent for the critical experimental conditions, for each of the candidate retrieval models.

P (number of samples = *X*)	1	2	3
LD antecedent, unrestricted	0.270	0.385	0.345
Local antecedent, unrestricted	0.353	0.437	0.210
LD antecedent, local search	0.190	0.258	0.552
Local antecedent, local search	0.560	0.351	0.089

[2]Fifty milliseconds is the time necessary to execute a single production step in ACT-R.

Inspection of the resulting finishing time distributions showed that they were well-fit by gamma distributions. Therefore, we modeled the overall predicted finishing time distribution for a given retrieval model as a mixture of gamma distributions, with each component reflecting the distribution of finishing times for a process with n sampling iterations. The mixing probabilities on each component were provided by the distribution in **Table 5**. With this mixture distribution, we could then follow the modeling approach advanced by McElree (1993). To do this, we used the resulting mixture to model the probability that the retrieval process will have completed by any time t as the cumulative distribution of this mixture, offset by a constant base encoding time δ (McElree, 1993):

$$P(T \leq t) = \frac{\beta^\alpha}{(\alpha - 1)} \int_0^{t-\delta} e^{-\beta t'} t'^{\alpha-1} dt' \qquad t > \delta, \text{ else } 0. \tag{14}$$

This cumulative distribution was then used to estimate the probability of responding with a hit at each time point t. This was calculated following the method described in McElree (1993). In particular, we assumed that all unfinished processes at time t contributed a hit 50% of the time, reflecting a guess on the part of the participant. We additionally assumed that on 5% of trials the target antecedent was rejected, leading to a miss response. The predicted proportion of hits at each time point was then transformed using the inverse cumulative normal distribution. Finally, the SAT function was fit to the predicted curves for each retrieval model and parameter setting, and the speed measure $\beta^{-1} + \delta$ was estimated for each predicted curve. We define the *locality advantage* as the predicted speed to access a long-distance antecedent minus the predicted speed to access a local antecedent, given a set of model parameters and a retrieval probe. The predicted locality advantages were calculated for both retrieval models, under all parameter settings. The predicted locality advantages were then compared to the empirical locality advantage in speed observed in Experiment 1.

Figure 2 provides a comparison of the empirical locality advantage with the predicted locality advantages for unrestricted and local search models. It can be seen that the local search model provides a good fit to the SAT data. On average, local search models predicted a locality advantage of 143 ms, approximately half of the observed empirical estimate of 294 ms from Experiment 1. However, the unrestricted search model predicts a much smaller speed advantage for local antecedents (39 ms). We tested the fit of each candidate retrieval model to the data by comparing the distribution of predicted locality advantage effects to the distribution of the mean locality effect estimated in Experiment 1. From these distributions, we calculated Bayes factors using the model comparison approach advocated by Gallistel (2009). This comparison gives 5:1 odds in favor of the local search model over the unrestricted model, providing "substantial" evidence in favor of the Local Search model (Jeffreys, 1961).

The modeling results suggest that the local search model provides a better explanation of our experimental data than does an account in which the locality advantage is simply attributed to the heightened activation of the local subject. We note that the model

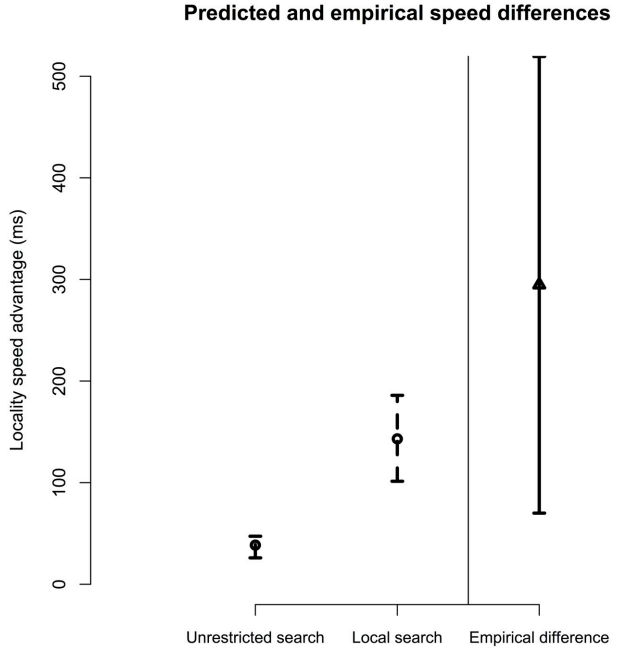


FIGURE 2 | Comparison of empirical speed advantage for local *ziji* antecedents and search model's predictions, given both unrestricted and local search retrieval probes. For model predictions, error bars represent the 95% central range of predicted locality advantage over parameter settings. For empirical data, error bars represent 95% confidence interval of true speed advantage for local antecedents by participants.

does confirm that an unrestricted model of antecedent retrieval does predict a small speed difference in the SAT function, due to the interaction of temporal decay, RI, and reactivation of the local subject prior to the anaphor. However, given the modeling assumptions here, these factors alone were not sufficient to allow the model to capture the findings of Experiment 1. By providing evidence against these plausible alternative explanations for the results of Experiment 1, the findings from the computational simulations lend additional support to the Local Search hypothesis.

DISCUSSION

SUMMARY OF RESULTS

The current study presented time-course data from the MR-SAT paradigm on the processing of the Mandarin Chinese long-distance reflexive *ziji*. Non-linear regressions using the SAT function revealed that the parameters that describe the speed of processing (specifically, rate and speed) were significantly faster for sentences containing a local antecedent for *ziji* than for sentences with a long-distance antecedent for *ziji*. Control conditions without any anaphoric dependency showed no difference in speed or rate parameters. We observed only marginal differences in accuracy. Sentences with long-distance animate subjects were accepted at slightly higher rates for *ziji* and control conditions alike, and control conditions were accepted at a higher rate than *ziji* conditions. A follow-up experiment evaluated the interpretations that comprehenders assigned to *ziji* to a subset of our experimental materials. These results confirmed that participants overwhelmingly interpreted a local animate subject as the preferred antecedent for *ziji* when it was present, and likewise when there was a long-distance animate subject present. However, when there was no syntactically licit animate subject in the sentence, participants either rejected *ziji* as antecedentless, interpreted *ziji* as coreferent with an implicit possessor argument in the highest subject position, or interpreted it as coreferent with an animate subject embedded inside a temporal adjunct clause.

To aid in interpreting these data, we fit the predictions of two retrieval models to the SAT data. The Local Search model implemented a retrieval process that used positional syntactic cues to restrict retrieval to the local clause. The Unrestricted model used only item information to access potential antecedents. We showed that the Local Search model provided a closer fit to the empirical data than the Unrestricted model, using plausible parameter estimates.

LOCALITY IN RETRIEVAL

The slower time course to access the matrix subject suggests that comprehenders initially access the local subject position when retrieving an antecedent for *ziji*, even if that position does not contain an acceptable antecedent. The results of our simulations suggest that this misretrieval of the local subject is not merely due to a higher resting activation for local subject positions compared to more distant subject positions. Instead, the models suggest that comprehenders attempt to use retrieval cues to limit search to the local syntactic domain. This supports the key claims of the Local Search hypothesis: comprehenders attempt to limit retrieval to the local clause, even for dependencies that are not strictly clause-bounded. This suggests that in at least some cases, locality effects in processing do not simply reflect decay and interference processes. In some cases, they additionally reflect a search strategy that favors the retrieval of syntactically local dependents.

One interesting finding from Experiment 1 is the individual variation in the retrieval dynamics observed across participants. Four of the 17 participants showed substantially faster retrieval of the long-distance antecedent than the local antecedent. For these participants, the average speed advantage seen for long-distance antecedents was 343 ms. This variation raises the possibility that the positional cues used to retrieve an antecedent are under strategic control, such that these four participants were able to prioritize retrieval of the highest subject over the local subject. Additionally, two of the remaining 13 participants showed a substantial rate advantage for the local conditions, driven by extremely fast retrieval speeds for local *ziji* conditions. The extremely rapid growth of these participants' SAT functions suggests that they may have adopted a distinct strategy for determining whether *ziji* was licensed in our experiment, perhaps one based on familiarity with an animate referent rather than full retrieval of an antecedent. Although we believe it is important to understand the variation observed across our participants, we caution that these suggestions are for the moment highly speculative. Further research is necessary to determine the exact ways in which memory search strategies are subject to strategic and individual variation.

We presented a model of the Local Search strategy that relies on a direct access memory architecture. On this model, the slowdown for retrieving the matrix antecedent reflects the fact that comprehenders must execute multiple retrieval operations to recover the distant antecedent in the face of a substantive locality bias in retrieval. However, the data are also compatible with a serial scan mechanism that operates over syntactic structures. This is compatible with previous claims about the mechanisms that allow the recovery of order or positional information in retrieval (McElree and Dosher, 1993; Gronlund et al., 1997). On this model, the present results do not reflect misretrieval of the local subject, but rather a backwards process of traversing the parse until an acceptable antecedent is encountered. Although existing SAT data provide evidence against the use of serial search processes for a number of linguistic dependencies, it is possible that a serial search process is applied uniquely to syntactic binding dependencies. Indeed, Berwick and Weinberg (1984) make an argument on computational grounds for just such a serial, backwards search process for the retrieval of a bound anaphor's antecedent. However, our present results do not distinguish between these two distinct mechanisms.

Although we have argued that our simulations point to a role for a Local Search strategy in memory access, it is true that this argument rests on a number of modeling assumptions that we made. It is possible that the SAT results reflect an overwhelming activation advantage for the local subject that is not captured in our implemented retrieval model, which might lead to slowed access to the distant subject even without the use of positional cues. One way this might occur is if the local subject were available in the focus of attention at the point of processing the anaphor, thus obviating the need for any retrieval process (McElree, 2006; Jonides et al., 2008). This interpretation seems less likely in light of findings that indicate that the focus of attention is extremely limited in size and scope, possibly corresponding to just one task-relevant encoding (McElree and Dosher, 1989; McElree, 1998). If only one element occupies focal attention before *ziji* is processed, it is likely to be the verb, although it is difficult to generalize from findings about the scope of attention in recognition memory tasks to sentence processing. The data on the capacity of the focus of attention is somewhat sparser for connected linguistic representations, which have considerably richer structure than lists. However, it has been shown that opening a new clause displaces the contents of focal attention (McElree et al., 2003; Wagers and McElree, 2009), and so the adverbial clause that intervened between the subject and the verb in Experiment 1 is likely to have displaced the local subject from active memory.

A second possibility is that the local subject is reactivated at the verb that precedes *ziji.* Although our model accounted for a process of local subject reactivation prior to the anaphor, it is possible that the boost given to the local subject due to this reactivation is substantially larger than our model allows for. At present we cannot rule out this possibility, but we believe that it is unlikely on empirical grounds. In particular, data from the cross-modal lexical priming paradigm show that a subject is not strongly activated above baseline while processing its verb (Nicol and Swinney, 1989, 2003). Studies that have contrasted activation of the local subject position before and after reflexive anaphors demonstrate that reactivation of the local subject is contingent on the construction of an anaphoric dependency; processing the verb alone is not sufficient to boost activation, nor is activation observed in post-verbal positions that do not contain a reflexive anaphor (see a review in Nicol and Swinney, 2003).

Finally, it should be noted that our model assumes that all retrieval cues are equally diagnostic. Although this is a plausible assumption that is common in ACT-R modeling and elsewhere (see also Clark and Gronlund, 1996; Lewis and Vasishth, 2005), recent research into how retrieval cues are combined in sentence processing does raise the possibility that syntactic and semantic cues are not equally weighted. In particular, Van Dyke and McElree (2011) argue that syntactic cues are more highly weighted than semantic cues in comprehension, and Dillon et al. (2013) argue that cue weight may vary as a function of grammatical dependency. Further work is necessary to determine whether different cues to antecedent retrieval for *ziji* are in fact differentially weighted, and if so, how differential cue weighting would influence the conclusions of the present research.

CONCLUSION

The present study examined the time-course of antecedent retrieval for the Mandarin Chinese long-distance anaphor *ziji*. It was found that *ziji* is processed more quickly with a local antecedent than with a long-distance antecedent. A computational model of the retrieval process supports the conclusion that the locality advantage observed when retrieving *ziji*'s antecedent reflects an explicit local search strategy: when retrieving an antecedent, comprehenders prioritize retrieval of items within the local clause. These results suggest that locality effects in sentence processing cannot be entirely reduced to the effects of temporal decay and interference in memory.

ACKNOWLEDGMENTS

This work was supported in part by NSF EAPSI-091387 to Brian Dillon, by NSF IGERT DGE-0801465 to the University of Maryland, by NSF BCS-0848554 to Colin Phillips, and by a Natural Science Foundation of China grant (31170970) to Taomei Guo. We would like to thank Brian McElree for his extensive advice on the current work, as well as Rick Lewis for his help in performing the ACT-R simulations reported here. We are grateful to the members of the Department of Linguistics at the University of Maryland, Stephani Foraker, Lyn Frazier, Roger Levy, Julie Van Dyke, Shravan Vasishth, and Ming Xiang for useful discussions of the material presented here, and to Mehmeti Abdullah, Julianne Chaloux, Peiyao Chen, Yangsook Park, and Yu Kai for assistance in collecting data and preparing experimental materials.

REFERENCES

Anderson, J. R., and Lebiere, C. (1998). *The Atomic Components of Thought.* Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Badecker, B., and Lewis, R. (2007). "A new theory and computational model of working memory in sentence production: agreement errors as failures of cue-based retrieval," in *Paper Presented at the 20th Annual CUNY Sentence Processing Conference* (San Diego; La Jolla: University of California).

Bartek, B., Lewis, R. L., Vasishth, S., and Smith, M. R. (2011). In search of on-line locality effects in sentence comprehension. *J. Exp. Psychol. Learn. Mem. Cogn.* 37, 1178. doi: 10.1037/a0024194

Berwick, R., and Weinberg, A. (1984). *The Grammatical Basis of Linguistic Performance*. Cambridge: MIT Press.

Bornkessel, I., McElree, B., Schlesewsky, M., and Friederici, A. D. (2004). Multi-dimensional contributions to garden path strength: dissociating phrase structure from relational structure. *J. Mem. Lang.* 51, 495–522. doi: 10.1016/j.jml.2004.06.011

Büring, D. (2005). *Binding Theory*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511802669

Chen, Z., Jäger, L., and Vasishth, S. (2012). How structure-sensitive is the parser? Evidence from Mandarin Chinese. *Empir. Approach. Linguist. Theory Stud. Mean. Struct.* 111, 43.

Clark, S. E., and Gronlund, S. D. (1996). Global matching models of recognition memory: how the models match the data. *Psychon. Bull. Rev.* 3, 37–60. doi: 10.3758/BF03210740

Dennis, J. E. Jr., Gay, D. M., and Walsh, R. E. (1981). An adaptive nonlinear least-squares algorithm. *ACM Trans. Math. Soft.* 7, 348–368. doi: 10.1145/355958.355965

Dillon, B., Mishler, A., Sloggett, S., and Phillips, C. (2013). Contrasting intrusion profiles for agreement and anaphora: experimental and modeling evidence. *J. Mem. Lang.* 69, 85–103. doi: 10.1016/j.jml.2013.04.003

Drenhaus, H., Saddy, D., and Frisch, S. (2005). "Processing negative polarity items: when negation comes through the backdoor," in *Linguistic Evidence — Empirical, Theoretical, and Computational Perspectives*, eds S. Kepser and M. Reis (Berlin: Mouton de Gruyter), 145–165. doi: 10.1515/9783110197549.145

Drummond, A. (2011). *IbexFarm (Version 0.2.7) [Software]*. Available online at: http://spellout.net/ibexfarm/

Foraker, S., and McElree, B. (2007). The role of prominence in pronoun resolution: availability versus accessibility. *J. Mem. Lang.* 56, 357–383. doi: 10.1016/j.jml.2006.07.004

Frazier, L. (1978). *On Comprehending Sentences: Syntactic Parsing Strategies*. Doctoral dissertation, University of Connecticut.

Frazier, L., and Clifton, C. (1989). Successive cyclicity in the grammar and the parser. *Lang. Cogn. Process.* 4, 93–126. doi: 10.1080/01690968908406359

Frazier, L., and Clifton, C. (1996). *Construal*. Cambridge: MIT Press.

Frazier, L., and Clifton, C. (1998). "Sentence reanalysis, and visibility," in *Reanalysis in Sentence Processing*, eds J. D. Fodor and F. Ferreira (Netherlands: Springer), 143–176. doi: 10.1007/978-94-015-9070-9_5
The editors are Janet Dean Fodor and Fernanda Ferreira

Gallistel, C. R. (2009). The importance of proving the null. *Psychol. Rev.* 116, 439. doi: 10.1037/a0015251

Gao, L., Liu, Z., and Huang, Y. (2005). Who is *ziji*? An experimental research on Binding Principle. *Linguist. Sci.* 2, 39–50.

Gibson, E. (1998). Linguistic complexity: locality of syntactic dependencies. *Cognition* 68, 1–76. doi: 10.1016/S0010-0277(98)00034-1

Gibson, E., Pearlmutter, N., Canseco-Gonzalez, E., and Hickok, G. (1996). Recency preferences in the human sentence processing mechanism. *Cognition* 59, 23–59. doi: 10.1016/0010-0277(95)00687-7

Gordon, P. C., Hendrick, R., and Johnson, M. (2001). Memory interference during sentence processing. *Psychol. Sci.* 13, 425–430. doi: 10.1111/1467-9280.00475

Gordon, P. C., Hendrick, R., and Johnson, M. (2004). Effects of noun phrase type on sentence complexity. *J. Mem. Lang.* 51, 97–114. doi: 10.1016/j.jml.2004.02.003

Gronlund, S. D., Edwards, M. B., and Ohrt, D. D. (1997). Comparison of the retrieval of item versus spatial position information. *J. Exp. Psychol. Learn. Mem. Cogn.* 23, 1261–1274. doi: 10.1037/0278-7393.23.5.1261

Hawkins, J. A. (1994). *A performance Theory of Order and Constituency*. Cambridge, UK: Cambridge University Press.

Huang, C.-T. J., Cole, P., and Hermon, G. (2006). "Long-distance reflexives: an east asian perspective," in *The Blackwell Companion to Syntax*, Vol. III, eds M. Everaert, H. van Riemsdijk, R. Goedemans, and B. Hollebrandse (Blackwell), 21–84.

Huang, C.-T. J., and Liu, L. (2001). "Logophoricity, attitudes, and *ziji* at the interface," in *Long Distance Reflexives, Syntax and Semantics*, Vol. 33, eds P. Cole, C.-T. J. Huang, and G. Hermon (New York, NY: Academic Press), 141–195.

Huang, C. -T. J., Li, Y. -H. A., and Li, Y. (2009). *The Syntax of Chinese*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9781139166935

Jeffreys, H. (1961). *The Theory of Probability, 3rd Edn*. Oxford, UK: Oxford University Press.

Jonides, J., Lewis, R. L., Nee, D. E., Lustig, C. A., Berman, M. G., and Moore, K. S. (2008). The mind and brain of short-term memory. *Annu. Rev. Psychol.* 59, 193–224. doi: 10.1146/annurev.psych.59.103006.093615

Just, M. A., and Carpenter, P. A. (1992). A capacity theory of comprehension: individual differences in working memory. *Psychol. Rev.* 99:122. doi: 10.1037/0033-295X.99.1.122

Kimball, J. (1973). Seven principles of surface structure parsing in natural language. *Cognition* 2, 15–47. doi: 10.1016/0010-0277(72)90028-5

Kuno, S. (1972). Pronominalization, reflexivization, and direct discourse. *Linguist. Inq.* 3, 161–195.

Kush, D. (2013). *Respecting Relations.* Doctoral dissertation, University of Maryland, College Park.

Lewis, R., and Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cogn. Sci.* 29, 375–419 doi: 10.1207/s15516709cog0000_25

Lewis, R., Vasishth, S., and Van Dyke, J. (2006). Computational principles of working memory in sentence comprehension. *Trends Cogn. Sci.* 10, 447–454. doi: 10.1016/j.tics.2006.08.007

Liu, C., and Smith, P. (2009). Comparing time-accuracy curves: beyond goodness-of-fit measures. *Psychon. Bull. Rev.* 16, 190–203. doi: 10.3758/PBR.16.1.190

Li, X., and Zhou, X. (2010). Who is *ziji*? ERP responses to the Chinese reflexive pronouns during sentence comprehension. *Brain Res.* 1331, 96–104. doi: 10.1016/j.brainres.2010.03.050

Liu, Z. (2009). The cognitive process of Chinese reflexive processing. *J. Chin. Linguisti.* 37, 1–27.

Lorch, R. F., and Myers, J. L. (1990). Regression analyses of repeated measures data in cognitive research. *J. Exp. Psychol. Learn. Mem. Cogn.* 16:149. doi: 10.1037/0278-7393.16.1.149

MacDonald, M. C., Pearlmutter, N. J., and Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychol. Rev.* 101:676. doi: 10.1037/0033-295X.101.4.676

Macmillan, N. A., and Creelman, C. D. (2004). *Detection Theory: A User's Guide*. New York, NY: Cambridge University Press.

Martin, A. E., and McElree, B. (2008). A content-addressable pointer mechanism underlies comprehension of verb-phrase ellipsis. *J. Mem. Lang.* 58, 879–906. doi: 10.1016/j.jml.2007.06.010

Martin, A. E., and McElree, B. (2009). Memory operations that support language comprehension: evidence from verb-phrase ellipsis. *J. Exp. Psychol. Learn. Mem. Cogn.* 35, 1231. doi: 10.1037/a0016271

Martin, A. E., and McElree, B. (2011). Direct-access retrieval during sentence comprehension: evidence from sluicing. *J. Mem. Lang.* 64, 327–343. doi: 10.1016/j.jml.2010.12.006

McElree, B. (1993). The locus of lexical preference effects in sentence comprehension: a time-course analysis. *J. Mem. Lang.* 32, 536–571. doi: 10.1006/jmla.1993.1028

McElree, B. (1998). Attended and non-attended states in working memory: accessing categorized structures. *J. Mem. Lang.* 38, 225–252. doi: 10.1006/jmla.1997.2545

McElree, B. (2000). Sentence comprehension is mediated by content-addressable memory structures. *Journal of Psycholinguist. Res.* 29, 111–123. doi: 10.1023/A:1005184709695

McElree, B. (2006). Accessing recent events. *Psychol. Learn. Motiv.* 46, 155–200. doi: 10.1016/S0079-7421(06)46005-9

McElree, B., and Dosher, B. (1989). Serial position and set size in short-term memory: time course of recognition. *J. Exp. Psychol. Gen.* 18, 346–373.

McElree, B., and Dosher, B. (1993). Serial retrieval processes in the recovery of order information. *J. Exp. Psychol. Gen.* 122, 291–315.

McElree, B., Foraker, S., and Dyer, L. (2003). Memory structures that subserve sentence comprehension. *J. Mem. Lang.* 48, 67–91. doi: 10.1016/S0749-596X(02)00515-6

McElree, B., and Griffith, T. (1998). Structural and lexical constraints on filling gaps during sentence processing: a time-course analysis. *J. Exp. Psychol. Learn. Mem. Cogn.* 24, 432–460.

Nicol, J. L., and Swinney, D. (2003). "The psycholinguistics of anaphora," in *Anaphora: A Reference Guide*, ed A. Barss (Malden, MA: Blackwell Publishing), 72–104.

Nicol, J., and Swinney, D. (1989). The role of structure in coreference assignment during sentence comprehension. *J. Psycholinguist. Res.* 18, 5–19.

Öztekin, I., and McElree, B. (2007). Proactive interference slows recognition by eliminating fast assessments of familiarity. *J. Mem. Lang.* 57, 126–149. doi: 10.1016/j.jml.2006.08.011

R Core Team. (2013). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available online at: http://www.R-project.org/

Rohde, D. (2003). *Linger (Version 2.94) [Software]*. Available online at: http://tedlab.mit.edu/∼dr/Linger/

Sturt, P., Pickering, M. J., and Crocker, M. W. (1999). Structural change and reanalysis difficulty in language comprehension. *J. Mem. Lang.* 40, 136–150.

Sturt, P., Pickering, M. J., and Crocker, M. W. (2000). Search strategies in syntactic reanalysis. *J. Psycholinguist. Res.* 29, 183–194. doi: 10.1023/A:1005145028351

Townsend, J. T., and Ashby, F. G. (1983). *Stochastic Modeling of Elementary Psychological Processes*. New York, NY: Cambridge University Press.

Van Dyke, J. A. (2007). Interference effects from grammatically unavailable constituents during sentence processing. *J. Exp. Psychol. Learn. Mem. Cogn.* 33, 407–430. doi: 10.1037/0278-7393.33.2.407

Van Dyke, J. A., and Lewis, R. L. (2003). Distinguishing effects of structure and decay on attachment and repair: a cue-based parsing account of recovery from misanalyzed ambiguities. *J. Mem. Lang.* 49, 285–316. doi: 10.1016/S0749-596X(03)00081-0

Van Dyke, J. A., and McElree, B. (2006). Retrieval interference in sentence processing. *J. Mem. Lang.* 55, 157–166. doi: 10.1016/j.jml.2006.03.007

Van Dyke, J. A., and McElree, B. (2011). Cue-dependent interference in comprehension. *J. Mem. Lang.* 65, 247–263. doi: 10.1016/j.jml.2011.05.002

Vasishth, S., Brüssow, S., Lewis, R., and Drenhaus, H. (2008). Processing polarity: how the ungrammatical intrudes on the grammatical. *Cogn. Sci.* 32, 685–712. doi: 10.1080/03640210802066865

Wagers, M., and McElree, B. (2009). "*Focal attention and the timing of memory retrieval in language comprehension,*" *Talk given at the Architectures and Mechanisms for Language Processing Conference* (Barcelona).

Wagers, M. W., Lau, E. F., and Phillips, C. (2009). Agreement attraction in comprehension: representations and processes. *J. Mem. Lang.* 61, 206–237. doi: 10.1016/j.jml.2009.04.002

Wickelgren, W. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta Psychologica* 41, 67–85.

Wickelgren, W. A., Corbett, A., and Dosher, B. (1980). Priming and retrieval from short-term memory: a speed-accuracy tradeoff analysis. *J. Verb. Learn. Verb. Behav.* 19, 387–404.

Xiang, M., Dillon, B., and Phillips, C. (2009). Illusory licensing effects across dependency types: ERP evidence. *Brain Lang.* 108, 40–55. doi: 10.1016/j.bandl.2008.10.002

Xue, P., Pollard, C., and Sag, I. (1994). "A new perspective on Chinese *ziji*," in *Proceedings of the 13th Annual Meeting of the West Coast Conference on Formal Linguistics* (University of Chicago Press and CSLI Publication), 432–447.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 26 June 2014; accepted: 28 August 2014; published online: 24 September 2014.

*Citation: Dillon B, Chow W-Y, Wagers M, Guo T, Liu F and Phillips C (2014) The structure-sensitivity of memory access: evidence from Mandarin Chinese. Front. Psychol. **5**:1025. doi: 10.3389/fpsyg.2014.01025*

This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology.

Copyright © 2014 Dillon, Chow, Wagers, Guo, Liu and Phillips. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 05 January 2016
doi: 10.3389/fpsyg.2015.01966

The Relationship Between Anaphor Features and Antecedent Retrieval: Comparing Mandarin *Ziji* and *Ta-Ziji*

Brian Dillon[1]*, *Wing-Yee Chow*[2] and *Ming Xiang*[3]

[1] Department of Linguistics, University of Massachusetts Amherst, Amherst, MA, USA, [2] Department of Linguistics, University College London, London, UK, [3] Department of Linguistics, University of Chicago, Chicago, IL, USA

In the present study we report two self-paced reading experiments that investigate antecedent retrieval processes in sentence comprehension by contrasting the real-time processing behavior of two different reflexive anaphors in Mandarin Chinese. Previous work has suggested that comprehenders initially evaluate the fit between the morphologically simple long-distance reflexive "ziji" and the closest available subject position, only subsequently considering more structurally distant antecedents (Gao et al., 2005; Liu, 2009; Li and Zhou, 2010; Dillon et al., 2014; cf. Chen et al., 2012). In this paper, we investigate whether this *locality bias* effect obtains for other reflexive anaphors in Mandarin Chinese, or if it is associated specifically with the morphologically simple reflexive *ziji*. We do this by comparing the processing of *ziji* to the processing of the morphologically complex reflexive *ta-ziji* (lit. s/he-self). In Experiment 1, we investigate the processing of *ziji*, and replicate the finding of a strong locality bias effect for *ziji* in self-paced reading measures. In Experiment 2, we investigate the processing of the morphologically complex reflexive *ta-ziji* in the same structural configurations as Experiment 1. A comparison of our experiments reveals that *ta-ziji* shows a significantly weaker locality bias effect than *ziji* does. We propose that this results from the difference in the number of morphological and semantic features on the anaphor *ta-ziji* relative to *ziji*. Specifically, we propose that the additional retrieval cues associated with *ta-ziji* reduce interference from irrelevant representations in memory, allowing it to more reliably access an antecedent regardless its linear or structural distance. This reduced interference in turn leads to a diminished locality bias effect for the morphologically complex anaphor *ta-ziji*.

Keywords: sentence processing, Mandarin Chinese, long-distance reflexives, working memory, referential processing

OPEN ACCESS

Edited by:
Claudia Felser, University of Potsdam, Germany

Reviewed by:
Lena Ann Jäger, University of Potsdam, Germany
Andrea Eyleen Martin-Nieuwland, University of Edinburgh, UK

***Correspondence:**
Brian Dillon
brian@linguist.umass.edu

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: *27 July 2015*
Accepted: *08 December 2015*
Published: *05 January 2016*

Citation:
Dillon B, Chow W-Y and Xiang M (2016) The Relationship Between Anaphor Features and Antecedent Retrieval: Comparing Mandarin Ziji and Ta-Ziji. Front. Psychol. 6:1966. doi: 10.3389/fpsyg.2015.01966

INTRODUCTION

Anaphoric expressions such as pronouns (e.g., *him*, *he*), reflexives (e.g., *himself*), and anaphoric definite descriptions (e.g., *the boy*) have been widely studied in both linguistic and psycholinguistic traditions. Linguists have long been concerned with how the interpretation and syntactic distribution of referring expressions are determined (Chomsky, 1981; Heim, 1982; Elbourne, 2008; *a.o.*). Psycholinguists have studied anaphoric expressions both as a window into how

comprehenders organize a text, and as a window into the working memory mechanisms that support sentence-level and discourse-level language comprehension (Kintsch, 1975; Gernsbacher, 1989; Greene et al., 1992; Myers and O'Brien, 1998; Foraker and McElree, 2011; Kush, 2013; Sturt, 2013; Dillon, 2014; *a.o*).

In the present work, we pursue a research question at the intersection of these two traditions. We ask how the form of a referring expression is related to the processing mechanisms that comprehenders use to assign it a referent. To this end, we contrast the processing of two closely related reflexive anaphors in Mandarin Chinese. Specifically, we compare the processing behavior of the morphologically simple reflexive *ziji* to that of the morphologically complex reflexive *ta-ziji*. In two self-paced reading experiments, we investigate the degree to which each anaphor exhibits a *locality bias*, a processing advantage (or preference) for syntactically local antecedents over more distant ones. Our empirical goal is to investigate the effect of morphological complexity on the processing of an anaphoric expression, with special attention to how morphology modulates the degree to which an anaphor will exhibit locality biases in processing. We interpret our results with respect to a theoretical model of anaphoric processing developed in other work (Dillon et al., 2014). To foreshadow our empirical and theoretical conclusions: our experimental findings suggest that morphologically complex anaphors in Mandarin Chinese show a diminished locality bias in comparison to morphologically simple ones, a finding that we attribute to how the processor makes use of the richer morphological feature content of morphologically complex anaphors in retrieving an antecedent from memory.

LONG DISTANCE REFLEXIVES AND LOCALITY EFFECTS

In recent years, there have been a number of experimental investigations into the real-time processing behavior of the Mandarin Chinese long-distance reflexive *ziji*. *Ziji* is a morphologically simplex reflexive, literally meaning *self* (Huang et al., 2009). *Ziji* is a *long-distance* reflexive, unlike English reflexives which must be bound within their immediate tensed clause (their *binding domain*; Chomsky, 1981). Long-distance reflexives are so called because their binding domain is larger than their immediate tensed clause, although the exact size of their expanded domain varies across languages (see Büring, 2005). For Mandarin *ziji*, it appears that the binding domain is the entire root clause in which *ziji* is found (Tang, 1989; Xue et al., 1994; Huang and Liu, 2001; Büring, 2005; Huang et al., 2006, 2009). To take one example, in (1) *ziji* may be bound either by the subject of its immediate clause *Lisi*, or by the subject of the higher (root) clause *Zhangsan* (subscripts are used to indicate acceptable coindexation):

(1) 张三 说 李四 弄伤了 自己
Zhangsan$_j$ shuo Lisi$_i$ nongshang-le ziji$_{i/j}$
Zhangsan say Lisi harm-PERF self
"Zhangsan says that Lisi harmed him/herself"

Ziji requires an animate antecedent (Tang, 1989; Xue et al., 1994; Huang and Liu, 2001), and receives an interpretation analogous to English reflexive forms. *Ziji* does not bear any overt morphological features, however, and so may take antecedents regardless of their gender, number, or person features.

Given the possibility of long-distance binding, it is interesting to note that many experimental studies have shown that comprehenders show a locality bias when processing *ziji*, preferring or more easily processing antecedents in their local clause over antecedents found in more distant clauses. For example, Li and Zhou (2010) conducted an ERP experiment in Mandarin, measuring the electrophysiological response to the anaphor "ziji" in examples like (2):

(2) a. 小李 让 小张 不要 伪装 自己
Xiaoli$_i$ rang Xiaozhang$_j$ buyao weizhuang ziji$_{?i/j}$.
Xiaoli ask Xiaozhang not disguise ziji.
"Xiaoli asked Xiaozhang not to disguise himself."

b. 小李 让 小张 不要 牵连 自己
Xiaoli$_i$ rang Xiaozhang$_j$ buyao qianlian ziji$_{i/?j}$.
Xiaoli ask Xiaozhang not embroil ziji.
"Xiaoli asked Xiaozhang not to embroil him."

Li and Zhou observed a larger positivity (P300/P600) at *ziji* when the semantics of the verb created a bias toward a long-distance reading of the reflexive, as in (2b), compared to when the meaning of the verb biased comprehenders toward a local reading of the reflexive, as in (2a).

Cross-modal priming studies point to a similar advantage for local antecedents over long-distance antecedents. Gao and colleagues (Gao et al., 2005; Liu, 2009) presented participants with spoken sentences of the form in (1). Upon reaching the sentence-final *ziji*, participants were presented with a visual probe word. When the probe was presented immediately after the anaphor, participants recognized probes that were semantic associates of a local antecedent more quickly than they did probes associated with long-distance antecedents; this locality advantage disappeared or reversed at slightly longer SOAs (160 and 370 ms).

Using a different experimental paradigm, Chen et al. (2012) showed using self-paced reading that locally bound *ziji* was read more quickly than long-distance bound *ziji*. These authors leveraged the observation that *ziji* requires an animate antecedent to create the pair of experimental sentences in (3) (brackets are used to indicate tensed clause boundaries):

(3) a. 反对派 领袖 表示 这个 声明
Fanduipai-lingxiu$_i$ biaoshi [zhe-ge shengming$_j$
在 抗议 失控 的 时候 告诫了 自己 的
[zai kangyi$_k$ shikong de- shihou] gaojie-le ziji$_{i/*j/*k}$
党员
de dangyuan]
opposition-leader say [this- cl announcement [at protest out.of.control time] warn- PERF ziji de party.member]
"The opposition leader said that this announcement warned his party members when the protest was out of control"

b. 这个 声明 表示 反对派 领袖
Zhe-ge shengming$_i$ biaoshi [fanduipai- lingxiu$_j$
在 抗议 失控 的 时候 告诫了 自己 的
[zai kangyi$_k$ shikong de-shihou] gaojie-le ziji$_{*i/j/*k}$
党员
de dangyuan]
this-CL announcement say [opposition-leader [at protest out.of.control time] warn-PERF ziji de party.member]
"The announcement said that the opposition leader warned his party members when the protest was out of control"

In these examples, *ziji* is the possessor of the direct object NP and it appears immediately after the embedded verb. In (3a), the only animate, c-commanding antecedent is *fanduipai-lingxiu*, "opposition leader." Thus, in this example, *ziji* must take a long-distance antecedent in the immediately higher clause. In contrast, the embedded subject in (3b) is the only animate and c-commanding antecedent, and so *ziji* must take a syntactically local antecedent. In this paradigm, the difference in reading times between *ziji* in (3a) and (3b) is taken to indicate the difficulty of constructing a long-distance *ziji* interpretation in (3a). Chen and colleagues observed a small but reliable RT slow-down in reading times at the region following *ziji* (*de*) in (3a) relative to (3b), suggesting more difficulty in constructing a long-distance than local interpretation of *ziji*. This result was subsequently replicated in an eye-tracking while reading study, using direct object *ziji* in place of possessive *ziji*, and without an adverbial clause intervening between the subject and the verb (Jäger et al., 2015; Experiment 2).

Dillon et al. (2014) asked whether the locality bias associated with *ziji* reflected a difference in processing *speed* for accessing long-distance antecedents, or simply a difference in processing *accuracy* associated with long-distance antecedents. For example, it is possible that the memory trace of a distant antecedent is of relatively poor quality compared to that of a local antecedent, perhaps due to memory decay processes. Such a difference in the representational quality of the antecedent could have given rise to the locality bias effect observed in previous studies, without any difference in processing speed. Alternatively, it may be that comprehenders simply take more time to access a long-distance antecedent, such that local antecedent positions have a temporal advantage compared to more structurally or linearly distant positions. Simple response time measures cannot tease these possibilities apart (see McElree, 2006). In order to ask this question, Dillon and colleagues used the multiple-response speed-accuracy tradeoff (MR-SAT) technique to investigate sentences similar to those in (3). The (MR-)SAT technique involves eliciting behavioral responses at a series of pre-defined response deadlines. This allows the researchers to chart how accuracy on a response measure grows as a function of time, giving a full picture of the time-course of processing. Importantly, the resulting SAT function may be separated into independent measures of processing speed and processing accuracy. Dillon et al.'s results indicated that the difference between local and long-distance binding was reflected in the rate parameter of the SAT function, suggesting that comprehenders took longer to retrieve long-distance antecedents for *ziji* than local antecedents.

Thus, a growing body of evidence suggests that despite the formal possibility of long-distance antecedents for *ziji*, comprehenders experience relatively more processing difficulty when *ziji*'s antecedent is not local. Furthermore, this locality bias seems to reflect a temporal advantage for local antecedents over long-distance antecedents: comprehenders more rapidly access local antecedent positions than long-distance positions.

ZIJI AND *TA-ZIJI*

Although previous research on *ziji* provides much evidence for a locality bias associated with *ziji*, it is not known how general this locality bias is. It is possible that a locality preference is a general property of long-distance reflexive anaphors. This might be expected if *ziji*'s locality bias was simply a reflection of the temporal or linear proximity of a local antecedent. In this case, we might expect all reflexive forms that can find an antecedent outside of their immediate clause to show comparable locality bias. On the other hand, it may be the case that the locality bias is rooted in some other specific property of *ziji*. For example, it may be the case that *ziji*'s lack of overt morphological features causes comprehenders to rely more heavily on positional cues when identifying an antecedent, which could lead to a preference for structurally local antecedents. If this is true, then we might expect the presence of locality bias effects to vary from anaphor to anaphor, depending on the surface form of the anaphor.

Mandarin grammar allows us to ask this question, because reflexive anaphors in Mandarin come in two forms: the morphologically simple, "bare" reflexive *ziji*, and morphologically complex anaphors. An example of a morphologically complex anaphor is *ta-ziji*, which consists of a third singular pronoun along with the bare reflexive (e.g. "he-self"). Other morphologically complex reflexives may be formed by combining other pronouns with *ziji* (e.g., *wo-ziji*, myself; *ni-ziji*, yourself), although here we focus on the third person singular form *ta-ziji*. *Ta-ziji* has a distribution that partially overlaps with *ziji*. For instance, when the antecedent of the anaphor is in the local clause, *ta-ziji* and *ziji* are interchangeable:

(4) 李四 弄伤了 自己 / 他自己
Lisi$_i$ nongshang-le ziji$_i$/ta-ziji$_i$
Lisi harm-PERF self / 3sg-self
"Lisi harmed himself"

The morphological differences between *ziji* and *ta-ziji* could lead to processing differences, because there are reasons to suspect that the addition of an overt pronoun to form a morphologically complex anaphor will yield richer cues for purposes of identifying an antecedent. First, the orthographic representation of the pronoun overtly provides gender and personhood cues: 他 (tā) is used for human male referents, 她 (tā) is used for human female referents, and 它 (tā) is reserved for non-human or gender-neutral referents. These forms are

distinguished orthographically, but they are not distinguished phonologically. In addition, the use of an overt pronoun is statistically more likely for gendered human antecedents than for non-human or gender neutral antecedents. A search of the Google Books corpus for simplified Chinese in the last 50 years reveals that approximately 73–80% of tokens of the third person singular pronoun refer to explicitly gendered human antecedents.

If this is correct, then we might say that *ta-ziji* has more features that can be used as cues to identify an antecedent when processing the reflexive. In particular, the addition of a pronominal form contributes humanness cues (i.e., [+human]) and gender cues. In contrast, the reflexive form *ziji* may only contribute animacy cues, because this is the only restriction that it places on potential antecedents. An interesting question to ask is whether the relatively more specified feature content of *ta-ziji* will lead to diminished locality bias for *ta-ziji* compared to *ziji*. If the locality bias associated with *ziji* reflects solely the influence of standard memory variables, such as decay or interference, then we might not expect the size or magnitude of the effect to vary with the surface form of the anaphor. If, on the other hand, the surface form of the anaphor contributes additional cues to identifying an antecedent, then locality bias might be diminished or eliminated for anaphors whose surface form bears more overt features. Thus, in the present study we aim to provide a head to head comparison of the locality bias associated with morphologically simple and morphologically complex anaphors, in an attempt to determine how generally locality bias is in the processing of anaphors in Mandarin.

Unfortunately, a direct comparison of the locality effects for *ziji* and *ta-ziji* is somewhat complicated by the fact that they do not have identical syntactic distributions. In contrast to *ziji*, the size of *ta-ziji*'s binding domain is a matter of some controversy. Huang et al. (2009) reported that it must be bound within its immediate tensed clause, like English *himself*. However, Pan (1998, 2000) argued that the binding domain of *ta-ziji* is fixed by the closest accessible animate antecedent, such that *ta-zij* can be bound outside of its local clause if the local subject is inanimate. What is clear is that *ta-ziji* places greater restrictions on long-distance antecedents than does *ziji*, and for this reason it is sometimes classified as a purely local reflexive in Mandarin (Huang et al., 2009). Because of the lack of clarity in the binding domains associated with these two reflexives, it is not ideal to compare these reflexives in the same embedding configurations used in previous studies (Chen et al., 2012; Dillon et al., 2014; Jäger et al., 2015).

Instead, we compared the behavior of *ziji* and *ta-ziji* in environments where they do have reliably overlapping distributions. For both *ziji* and *ta-ziji* the c-command relation that regulates binding in English (Chomsky, 1981) appears to be too restrictive. Instead, antecedents that do not strictly c-command these anaphors may be grammatically available, as in (5) (Tang, 1989):

(5) 张三 的 骄傲 害了 自己 / 他自己
Zhangsan$_i$ de jiao'ao$_j$ hai-le ziji$_{i/*j}$/ ta-ziji$_{i/*j}$
Zhangsan de arrogance harm-PERF ziji / 3sg-ziji
"Zhangsan's arrogance harmed him."

In (5), the antecedent *Zhangsan* is embedded inside the subject, and hence does not c-command the anaphor. Nonetheless, in this configuration it is available to bind the reflexive. The structural relationship between Zhangsan and *(ta-)ziji* in (5) is referred to as *subcommand* (Tang, 1989; Huang and Tang, 1991). An NP is said to subcommand the anaphor if it is contained within an NP in subject position that c-commands or subcommands the anaphor (Tang, 1989).

However, it is important to note that subcommanding antecedents are not freely available. Instead, a subcommanding antecedent is only available when no animate c-commanding or subcommanding antecedent is structurally closer to *ziji*. Thus, when the subject head noun is animate, subcommanding antecedents are grammatically blocked as in (6):

(6) 张三 的 小孩 害了 自己 / 他自己
Zhangsan$_i$ de xiaohai$_j$ hai-le ziji$_{*i/j}$/ ta-ziji$_{*i/j}$
Zhangsan de son harm-PERF ziji / 3sg-ziji
"Zhangsan's son harmed himself."

As *ziji* and *ta-ziji* distribute similarly in subcommanding environments, we may compare the processing of *ziji* and *ta-ziji* in configurations like (7):

(7) a. 张 太太 经常 光顾 的 那个
[Zhang taitai$_i$ jingchang guanggu de] na-ge
时装店 上个星期 把 自己 /她自己
shizhuangdian$_j$ shang-ge-xingqi ba ziji$_{i/*j}$ / ta-ziji$_{i/*j}$
不小心 弄伤了
bu xiaoxin nongshang-le.
Mrs. Zhang often visit DE that-CL boutique last-week ba self / 3sg-self not careful harm-PERF.
"The boutique that Mrs. Zhang often visits carelessly harmed her last week."

b. 媒体 报道 的 那个 女裁缝 上个星期
[Meiti$_i$ baodao de] na-ge nü-caifeng$_j$ shang-ge-xingqi
把 自己 / 他自己 不小心 弄伤了
ba ziji$_{*i/j}$ / ta-zij$_{*i/j}$ bu xiaoxin nongshang-le.
Media report-on DE that-CL seamstress last-week ba self / 3sg-self not careful harm-PERF.
"The seamstress that the media reported on carelessly harmed herself last week."

These examples have an object extracted relative clause (e.g., "that the media reported on") modifying a subject noun (e.g., *na-ge nü-caifeng* "that seamstress"). This structure creates two subject positions that could in principle bind an anaphor: the *local* subject position inside the matrix clause, and a *distant* subject position inside the relative clause. Given the licensing constraints on *ziji* and *ta-ziji*, we expect that both the local subject *na-ge nü-caifeng* "that seamstress" in (7b) and the long-distance subject *Zhang taitai* "Mrs Zhang" in (7b) should be grammatically accessible antecedents. However, these two antecedents differ in their structural and linear distance from the reflexive. The subcommanding antecedent *Mrs. Zhang* in (7a) is a long-distance antecedent because it is linearly and structurally more distant from the anaphor than the local antecedent *na-ge nü-caifeng* in

(7b). For this pair of conditions, then, a locality effect should present as increased reading times on the anaphor in (7a) compared to (7b).

We now present two experiments that investigate *ziji* and *ta-ziji* in Mandarin Chinese. Our goal in these experiments was to compare the processing profile of these two anaphors on a number of different dimensions. First, and most importantly, we ask whether both *ziji* and *ta-ziji* show locality effects of comparable magnitude in online sentence comprehension. In addition, we ask whether the processing of both *ziji* and *ta-ziji* is equally affected by the presence of multiple feature-matched antecedents. Previous research suggests that the presence of multiple, feature-matched antecedents may cause processing difficulty (the *multiple match effect* of Badecker and Straub, 2002), although this effect has not be observed in all studies (e.g., Clifton et al., 1999).

EXPERIMENT 1

Participants

Forty-one students from the University of Maryland community participated in the experiment. One participant was removed prior to analysis due to low comprehension question accuracy (see below). All participants were native Mandarin Chinese speakers from mainland China, and all had normal or corrected-to-normal vision. They were paid $10 for their participation in the experiment. Experimental sessions lasted approximately 45 min. Participants gave informed consent under an experimental protocol approved by the University of Maryland Institutional Review Board.

Stimuli

We created stimuli with the sentence structure in (7). We orthogonally manipulated the animacy of the local subject position and the embedded subject position, yielding four experimental conditions. These conditions are summarized in (8).

a. *LOCAL MATCH:*

媒体 报道 的 那个 女裁缝 上个星期 把
Meiti/ baodao de/ na-ge/ nücaifeng/ shang-ge-xingqi/ ba/
自己不小心 弄伤了
ziji/ bu xiaoxin/ nongshang-le.
Media/ report-on DE / that-CL / seamstress/ last-week/ BA/ self/ not careful/ harm-PERF.
"The seamtress that the media reported on carelessly harmed herself last week."

b. *DISTANT MATCH:*

张 太太 经常 光顾 的 那个 时装店
Zhang taitai/ jingchang guanggu de/ na-ge/ shizhuangdian/
上个星期 把 自己不小心 弄伤了
shang-ge-xingqi/ ba/ ziji/ bu xiaoxin/ nongshang-le.
Mrs. Zhang/ often visit DE / that-CL / boutique/ last-week/ BA/ self/ not careful/ harm-PERF.
"The boutique that Mrs. Zhang often visits carelessly harmed her last week."

c. *MULTIPLE MATCH:*

张 太太 经常 光顾 的 那个 女裁缝
Zhang taitai/ jingchang guanggu de/ na-ge/ nücaifeng/
上个星期 把 自己 不小心 弄伤了
shang-ge-xingqi/ ba/ ziji/ bu xiaoxin/ nongshang-le.
Mrs. Zhang/ often visit DE / that- CL/ seamstress/ last-week/ BA/ self/ not careful/ harm-PERF.
"The seamstress that Mrs. Zhang often visits carelessly harmed her/herself last week."

d. *NO MATCH:*

媒体 报道 的 那个 时装店 上个星期
Meiti/ baodao de/ na-ge/ shizhuangdian/ shang-ge-xingqi/
把 自己不小心 弄伤了
ba/ ziji/ bu xiaoxin/ nongshang-le.
Media/ report-on DE/ that-CL/ boutique/ last-week/ BA/ self/ not careful/ harm-PERF.
"The boutique that the media reported on carelessly harmed her last week."

The paradigm employed here thus followed Chen et al. (2012), Dillon et al. (2014), and Jäger et al. (2015) in using animacy to manipulate the binding possibilities for *ziji*. In the *LOCAL MATCH* and *DISTANT MATCH* conditions, the antecedent of *ziji* is the animate subject. In the *NO MATCH* condition, there is no intra-sentential antecedent for *ziji*. In the *MULTIPLE MATCH* condition, the local subject *na-ge nü-caifeng* "that seamstress" is the only grammatically available antecedent of *ziji*. In this condition, the animate local subject blocks access to the distant subject Mrs. Zhang; therefore, the interpretation of *ziji* is not ambiguous in the *MULTIPLE MATCH* condition (see Tang, 1989).

The primary comparison of interest for the present purposes is the difference in reading times between the *LOCAL MATCH* and *DISTANT MATCH* conditions at the anaphor. The *MULTIPLE MATCH* and *NO MATCH* conditions were included for two reasons. First, the factorial manipulation of the animacy of the two subject positions allows us to dissociate effects of interest from simple effects of local or distant subject animacy. Second, the inclusion of the *MULTIPLE MATCH* conditions allows us to estimate any reading time effects associated with multiple feature-matched antecedents (the multiple match effect, Badecker and Straub, 2002). The inclusion of the *NO MATCH* condition serves as a control. This allows us to evaluate whether readers were indeed attempting to find an antecedent for *ziji*; if this is the case, then the failure to find an appropriate antecedent in this condition should lead to longer reading times.

The distant (sub-commanding) antecedent position was always the subject of an object relative clause that modified the main clause subject. Owing to the head-final order of noun phrases in Mandarin Chinese, this embedded subject (distant antecedent) is both structurally and linearly further away from

the anaphor than the main clause subject (local antecedent). The local antecedent always followed the relative clause verb and the relativizing particle *de*. In order to construct plausible and natural sentences, the predicate inside the relative clause was different for animate (8a-c; e.g., "that Mrs. Zhang often visits") and inanimate (8b-d; "that the media reported on") relative clause subjects. The main clause predicate was constant across conditions.

In order to avoid having the critical word (the anaphor) in sentence-final position, the *ba* construction was used, because this construction has an S-ba-O-V word order (in contrast to the canonical SVO word order of Mandarin). A temporal adverbial was placed between the main clause subject (the local antecedent) and the *ba*-marked *ziji* to ensure that they were not adjacent to each other. A manner adverbial was placed between *ziji* and the main clause verb in order to provide an extra spillover region.

Eighteen sets of experimental items were produced, and distributed into four lists in a pseudo-Latin square fashion. They were combined with 77 fillers, including materials from an unrelated experiment, for a total of 95 sentences. The ratio of acceptable-to-unacceptable sentences varied slightly from list to list due to the pseudo-Latin square procedure, but remained between 83 and 85% acceptable. The fillers included 10 sentences that contained *ba* followed by non-anaphoric NPs in order to prevent participants from associating *ba* with *ziji* within the experiment.

Procedure

Sentences were presented using a moving-window self-paced reading paradigm, using the Linger software (Rohde, 2003). Each sentence was presented in black characters on a white screen, and no sentence was more than one line long. All sentences were presented using simplified Chinese characters. The sentences were segmented into 9 regions according to native speaker intuitions about where best to insert boundaries [regions are indicted by slashes in (8)]. This procedure resulted in regions that ranged from one character (e.g., *ba*) to 6 characters (e.g., *yishuticaoguanjun*, "gymnastics champion").

Sentences initially appeared as a series of dashes that obscured the entire sentence. Participants pressed the space bar to present the first region, and each subsequent space bar press masked the current region and triggered presentation of the subsequent region. Reaction times between button presses were recorded. After approximately 50% of the filler sentences, a Yes/No comprehension question was presented in its entirety on the screen, and participants were instructed to press one of two buttons to indicate their response. Feedback was given for incorrect responses. The critical *ziji* sentences never were followed by comprehension questions.

In the analyses below we refer to the region containing *ziji* as the *critical region*, and the region that follows (e.g., *bu xiaoxin*) as the *spillover region*.

Statistical Analysis

We performed a single statistical analysis over the pooled data in Experiments 1 and 2, which we present after Experiment 2. Reaction time data from both experiments were analyzed using linear mixed effects models with three critical experimental contrasts. Taking the *LOCAL MATCH* condition as the baseline, we defined the *Locality contrast* as the difference between the *DISTANT MATCH* condition and the *LOCAL MATCH* condition. As in previous studies (Chen et al., 2012; Dillon et al., 2014; Jäger et al., 2015), this contrast is interpreted as the penalty associated with long-distance binding of the anaphor. We further defined the *Multiple Match contrast* as the difference between the *MULTIPLE MATCH* condition and the *LOCAL MATCH* condition; this contrast is interpreted as the penalty associated with having multiple NPs that matched the features of the anaphor. Lastly, we defined the *No Antecedent contrast* as the difference between the *LOCAL MATCH* condition and the *NO MATCH* conditions. Each of these contrasts was coded with treatment coding, treating *LOCAL MATCH* as the baseline. These experimental contrasts were shared across Experiments 1 and 2. In addition to these fixed effects, we further included *Experiment* as a fixed effect with treatment coding, treating Experiment 1 as the baseline. Lastly, to test for differences in our experimental contrasts across experiments, we included terms for the interaction of *Experiment* with each experimental contrast.

Because our linear mixed effects models assume a normally-distributed response, we applied the Box-Cox procedure to reaction times at the regions we analyzed to determine a transformation that would yield a normally distributed response variable (Box and Cox, 1964). This procedure suggested a transformation in-between a negative reciprocal transform and a logarithmic transformation. Exploratory data analyses revealed that the qualitative pattern of results did not change under different transformations, and so we present the results of linear mixed effects models fit to logarithmically transformed reading time data. We adopted a "maximal" random effects structure, including random intercepts and random slopes for all fixed effect parameters within both subject and item grouping factors where possible (Barr et al., 2013). If the full model failed to converge, we removed random correlations but retained random slopes for all fixed effects.

Because of the pseudo-Latin square procedure, the number of sentences within each condition was not balanced within subjects. To test for any effects this imbalance may have had, we performed log-likelihood ratio tests of models with and without a fixed effect for experimental list. If log-likelihood tests indicated an effect of *list*, we performed further model comparisons to determine if the effect of *list* interacted with our experimental fixed effects.

In constructing the materials, we did not attempt to control the length or frequency of the subject noun phrases within items. However, as it has been shown that antecedent frequency is inversely correlated with reading times on anaphoric expressions (Van Gompel and Majid, 2004), we included antecedent frequency and antecedent length for both embedded and matrix subject positions as fixed effect control predictors in all mixed effects models. Antecedent frequency was estimated using the SUBTLEX Chinese corpus (Cai and Brysbaert, 2010). Many of our antecedent phrases were noun-noun compounds that were unattested in the corpus (e.g., *laladuiyuan*, "cheerleading squad member"). If the entire compound phrase was unattested, we

TABLE 1 | Mean acceptability ratings in Experiment 1.

LOCAL MATCH	*DISTANT MATCH*	*MULTIPLE MATCH*	*NO MATCH*
5.2 (0.3)	4.4 (0.3)	4.8 (0.3)	3.5 (0.3)

Parentheses represent standard error by participants, corrected for between-participant variance (Bakeman and McArthur, 1996).

used frequency of the head noun. Length was entered into the model as the number of characters of the head noun in the subject phrase. Both antecedent frequency and length were centered before being entered into the model.

Analysis was performed for three regions of the experimental sentences: the pre-critical region *ba*, the critical region *ziji*, and the spillover region [e.g., *bu xiaoxin* in (8)].

Results

Offline Judgments

Prior to Experiment 1, we gathered offline acceptability judgments of all experimental materials. All experimental stimuli, including fillers, were entered into the online experimental platform IbexFarm (Drummond, 2011). Twenty-two native Mandarin speakers were recruited from Beijing Normal University. They were directed to a web address that hosted the offline naturalness judgment questionnaire and they were asked to rate each experimental stimulus on a scale from 1 (not natural) to 7 (very natural).

The results of this offline judgment study are presented in **Table 1**. These data were analyzed using linear mixed effects modeling, with fixed effects for matrix subject animacy, distant subject animacy, and their interaction. This analysis revealed a main effect of local NP animacy (Est $= -1.09 \pm 0.25$, $t = -4.3$), and an interaction of local and distant NP animacy (Est $= -1.29 \pm 0.38$, $t = -3.45$). There were lower acceptability ratings for both conditions with a local inanimate subject (*DISTANT MATCH* and *NO MATCH*). However, a *post-hoc* comparison between these two conditions revealed that average ratings were significantly lower in the *NO MATCH* condition than in the *DISTANT MATCH* condition ($\bar{x} = -0.9$, 95%CI: $[-1.4, -0.4]$).

Comprehension

One participant was removed from further analysis due to low accuracy (less than 70% accurate). After this exclusion, accuracy on the comprehension questions in Experiment 1 averaged 87% across participants, indicating that the participant attended to the stimuli. Across participants, accuracy ranged from 73 to 98%.

Reading Times

Raw mean reading times in Experiment 1 are presented in **Table 2** and in **Figure 1**.

Discussion

The results of the offline judgment experiment revealed that raters assigned lower ratings to sentences where there was not a local antecedent for *ziji*. The lowest ratings were given to the *NO MATCH* condition, presumably reflecting the unacceptability that results from the lack of an intra-sentential antecedent. Interestingly, the *DISTANT MATCH* condition was rated lower than the *LOCAL MATCH* and *MULTIPLE MATCH* conditions. This penalty is consistent with the presence of a locality effect. This conclusion is supported by independent evidence that the acceptability of grammatical sentences is reliably modulated by the length of a binding dependency (Sprouse et al., 2011). However, it is also possible that this reflects relative unacceptability that results from having an inanimate matrix subject. Critically for the present purposes, the *DISTANT MATCH* condition was rated as more acceptable than the *NO MATCH* condition, consistent with the claim that the distant subject is grammatically accessible as an antecedent for *ziji* (Tang, 1989; Huang and Tang, 1991).

The results of the self-paced reading experiment suggest a locality effect in the reading times, with the *DISTANT MATCH* condition being read 41 ms more slowly than the *LOCAL MATCH* condition at the critical region, and 215 ms more slowly in the spillover region. If reliable, this finding would extend the locality bias effect observed in previous experiments to the subcommanding configuration tested here (Chen et al., 2012; Dillon et al., 2014; Jäger et al., 2015). The data also suggest numerically smaller effects of the *Multiple Match contrast* and the *No Antecedent contrast*: in the spillover region, reading times were 81 ms longer in the *MULTIPLE MATCH* condition than in the *LOCAL MATCH* condition, and 84 ms longer in the *NO MATCH* condition than in the *LOCAL MATCH* condition.

Before further interpreting the data in Experiment 1, we present the results of Experiment 2.

EXPERIMENT 2

Experiment 2 was identical to Experiment 1 in all major respects, except that Experiment 2 investigates the processing of the complex anaphor *ta-ziji*.

Participants

Seventy students from the University of Maryland community participated in the experiment. All participants were native Mandarin Chinese speakers from mainland China, and all had normal or corrected-to-normal vision. They were paid $10 per hour for their participation in the experiment. Participants gave informed consent under an experimental protocol approved by the University of Maryland Institutional Review Board.

Stimuli

The materials were largely identical to those from Experiment 1. Two important changes were made to these materials. First, all instances of *ziji* were replaced with *ta-ziji*. The materials were additionally modified so that within an experimental item set, the animate nouns in each position were of the same gender. This was done to ensure that both NPs in the *MULTIPLE MATCH* condition matched the features of the reflexive. This change was necessary because *ta* orthographically marks gender. Half of the revised materials had male nouns, and the other half had female nouns.

All other aspects of the stimuli, including the fillers and comprehension questions, were identical to Experiment 1.

TABLE 2 | Mean reading times per region in Experiment 1.

	1	2	3	4	5	6	7	8	9
LOCAL MATCH	762 (45)	731 (35)	638 (40)	849 (80)	703 (35)	564 (30)	448 (28)	522 (26)	615 (35)
DISTANT MATCH	728 (74)	644 (24)	599 (27)	914 (90)	715 (29)	525 (28)	489 (25)	737 (35)	847 (81)
MULTIPLE MATCH	681 (38)	706 (66)	546 (24)	780 (58)	767 (37)	642 (81)	467 (25)	603 (35)	695 (57)
NO MATCH	825 (71)	803 (44)	614 (53)	669 (44)	784 (43)	592 (31)	541 (60)	606 (30)	766 (51)

Parentheses represent standard error by participants, corrected for between-participant variance (Bakeman and McArthur, 1996). Region labels are as follows: 1:Zhang taitai 2:jingchang guanggu de 3:na-ge 4:shizhuangdian 5:shang-ge-xingqi 6:ba 7:ziji 8:bu xiaoxin 9:nongshang-le.

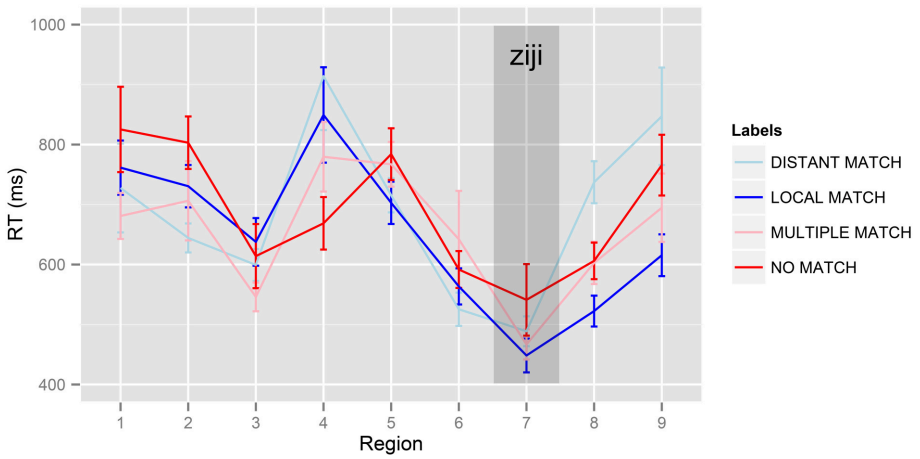


FIGURE 1 | Mean reading times per region in Experiment 1. Error bars represent standard error by participants, corrected for between-participant variance (Bakeman and McArthur, 1996). 1:*Zhang taitai* 2:*jingchang guanggu de* 3:*na-ge* 4:*shizhuangdian* 5:*shang-ge-xingqi* 6:*ba* 7:*ta-ziji* 8:*bu xiaoxin* 9:*nongshang-le.*

TABLE 3 | Mean acceptability ratings in Experiment 2.

LOCAL MATCH	*DISTANT MATCH*	*MULTIPLE MATCH*	*NO MATCH*
5.1 (0.3)	5 (0.3)	4.1 (0.3)	4 (0.3)

Parentheses represent standard error by participants, corrected for between-participant variance (Bakeman and McArthur, 1996).

Procedure

The experimental procedure was identical to Experiment 1.

Offline Judgments

As in Experiment 1, we gathered offline naturalness judgments of all experimental materials prior to running Experiment 2. Collection of judgments and recruitment of participants proceeded in the same fashion as the offline pre-test for Experiment 1. Twenty-six native Mandarin speakers were recruited from Beijing Normal University.

The results of the offline judgment study are presented in **Table 3**. Linear mixed effects modeling revealed only an interaction of local and distant NP animacy (Est $= -1.88 \pm 0.38$, $t = -4.15$). This interaction was driven by low ratings in the *NO MATCH* and *MULTIPLE MATCH* conditions. There was no appreciable difference between the ratings of the *LOCAL MATCH* and *DISTANT MATCH* conditions.

Comprehension

As in Experiment 1, one participant was removed from further analysis due to low accuracy (less than 70% accurate). Accuracy on the comprehension questions averaged 84% across participants, indicating that the participant attended to the stimuli. Across participants, accuracy ranged from 71 to 100%.

Reading Times

Raw mean reading times are presented in **Table 4** and **Figure 2**. Visual inspection of the means suggests a weaker locality effect in Experiment 2 than in Experiment 1: the difference between the *LOCAL MATCH* and *DISTANT MATCH* conditions was 57 ms at the critical region, and 53 ms at the spillover region (compared to 215 ms in Experiment 1). The reading times suggest a numerically an effect of the *No Antecedent contrast* (109 ms at the critical region, 148 ms in the spillover region), and a small effect for the *Multiple Match contrast* (10 ms at the critical region, 50 ms in the spillover region).

The results of the statistical modeling of the reaction times at the pre-critical, critical, and spillover regions are presented in **Tables 5**–**7**. Analysis revealed no significant effects of counterbalancing list, and so we report models that do not include list as a fixed effect predictor.

At the pre-critical region, *ba*, we did not observe any statistically significant effects. This pattern suggests any early differences in the materials between conditions—such as the

TABLE 4 | Mean reading times per region in Experiment 2.

	1	2	3	4	5	6	7	8	9
LOCAL MATCH	832 (29)	872 (33)	643 (25)	818 (32)	787 (32)	584 (23)	590 (22)	572 (22)	670 (43)
DISTANT MATCH	778 (27)	761 (22)	698 (46)	785 (37)	759 (33)	614 (27)	632 (21)	619 (18)	754 (42)
MULTIPLE MATCH	765 (30)	793 (30)	634 (20)	824 (33)	809 (28)	637 (28)	591 (15)	613 (22)	695 (37)
NO MATCH	842 (32)	904 (54)	701 (31)	765 (36)	773 (27)	599 (17)	682 (26)	716 (29)	750 (28)

Parentheses represent standard error by participants, corrected for between-participant variance (Bakeman and McArthur, 1996). 1:Zhang taitai 2:jingchang guanggu de 3:na-ge 4:shizhuangdian 5:shang-ge-xingqi 6:ba 7:ta-ziji 8:bu xiaoxin 9:nongshang-le.

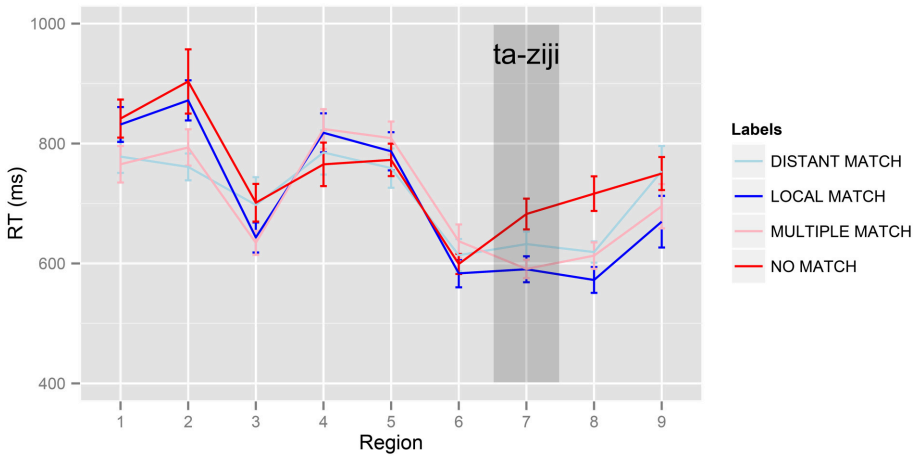


FIGURE 2 | Mean reading times per region in Experiment 2. Error bars represent standard error by participants, corrected for between-participant variance (Bakeman and McArthur, 1996). 1:*Zhang taitai*2:*jingchang guanggu de* 3:*na-ge* 4:*shizhuangdian* 5:*shang-ge-xingqi* 6:*ba* 7:*ta-ziji* 8:*bu xiaoxin* 9:*nongshang-le.*

TABLE 5 | Experimental fixed effects estimates from linear mixed effects modeling of pre-critical region across Experiments 1 and 2.

	Estimate	*t*
Experiment	0.09 (0.06)	1.45
LOCALITY	−0.02 (0.05)	−0.40
NO MATCH	0.04 (0.04)	0.90
MULTIPLE MATCH	0 (0.05)	0.10
Experiment: *LOCALITY*	0.03 (0.06)	0.49
Experiment: *NO MATCH*	−0.01 (0.06)	−0.17
Experiment: *MULTIPLE MATCH*	0.06 (0.06)	1.00

TABLE 6 | Experimental fixed effects estimates from linear mixed effects modeling of critical region across Experiments 1 and 2.

	Estimate	*t*
Experiment	0.24 (0.06)	4.30
LOCALITY	0.06 (0.05)	1.26
NO MATCH	0.05 (0.04)	1.02
MULTIPLE MATCH	0 (0.04)	0.12
Experiment: *LOCALITY*	−0.02 (0.05)	−0.38
Experiment: *NO MATCH*	0.05 (0.05)	0.91
Experiment: *MULTIPLE MATCH*	0.02 (0.05)	0.38

animacy of the subject, or the different relative clauses used in different conditions—had returned to a neutral baseline prior to the critical region. We examined this observation further by performing an additional analysis of the region that immediately preceded the pre-critical region (e.g., *shangge xingqi*, "last week"). As in the pre-critical region, we failed to observe any statistically significant effects, providing further evidence that pre-critical differences in the materials across conditions did not have durable or long-lasting effects on reading times preceding the critical region.

In the critical region, we observed only a fixed effect of *Experiment*. Reading times in the anaphor region were significantly longer in Experiment 2 compared to Experiment 1, presumably reflecting the fact that *ta-ziji* is longer than *ziji*.

In the spillover region, we observed a statistically significant effect of the *Locality contrast*, and a statistically significant effect of the *No Antecedent contrast*. We did not observe any significant effects of antecedent frequency or length. Critically, we observed an interaction of *Experiment* with the *Locality contrast*. The direction of this coefficient indicates that the locality contrast was significantly smaller in Experiment 2 than it was in Experiment 1. To further investigate the interaction of *Experiment* with the *Locality contrast*, we fit a second model in which the critical *Locality contrast* was nested within individual levels of

TABLE 7 | Experimental fixed effects estimates from linear mixed effects modeling of spillover region across Experiments 1 and 2.

	Estimate	*t*
Experiment	0.13 (0.06)	2.03
LOCALITY	0.29 (0.05)	6.31
NO MATCH	0.14 (0.04)	3.05
MULTIPLE MATCH	0.05 (0.05)	1.08
Experiment: *LOCALITY*	−0.22 (0.06)	−3.94
Experiment: *NO MATCH*	0.03 (0.06)	0.52
Experiment: *MULTIPLE MATCH*	0.01 (0.05)	0.14

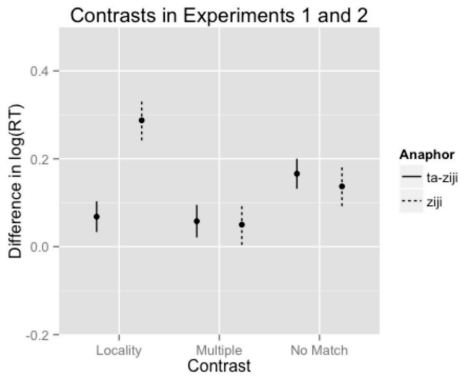


FIGURE 3 | Fixed effects estimates for Experimental contrasts in the spill-over region in both Experiment 1 (*ziji*) and Experiment 2 (*ta-ziji*). Error bars indicate the standard error associated with the fixed effect estimate.

Experiment. This model revealed that there was a significant *Locality contrast* for Experiment 1 (0.29 (0.05), $t = 6.31$). In Experiment 2, the estimated size of this effect was much smaller than in Experiment 1, and it was only marginal for Experiment 2 (0.07 (0.04), $t = 1.95$). The magnitude of the *No Antecedent contrast* was comparable between Experiments 1 and 2, and it reached statistical significance in both Experiments [Experiment 1: 0.14 (0.04), $t = 3.05$; Experiment 2: 0.17 (0.03), $t = 4.85$]. The *Multiple Match contrast* did not reach significance in either Experiment, although the magnitude of the observed effect and its sign were comparable across experiments [Experiment 1: 0.05 (0.05), $t = 1.08$; Experiment 2: 0.06 (0.04), $t = 1.56$]. The estimates of the fixed effects contrasts by *Experiment* yielded by this model are presented in **Figure 3**.

Discussion

The offline judgments for sentences containing *ta-ziji* revealed that the *DISTANT MATCH* and *LOCAL MATCH* conditions were considered equally acceptable, and that both were considered more acceptable than the *NO MATCH* condition. This pattern confirms that the distant subject position is accessible as an antecedent for *ta-ziji* in our materials. Furthermore, this pattern gives no indication of a locality bias in the offline judgments for *ta-ziji*. This contrasts sharply with the clear offline locality bias observed for *ziji* in Experiment 1.

Note furthermore that our *DISTANT MATCH* and *LOCAL MATCH* conditions differed in whether the main clause subject was animate. As the *DISTANT MATCH* condition was rated equally highly as the *LOCAL MATCH* condition in this experiment, one may infer that the inanimate main clause subjects in the *DISTANT MATCH* condition did not impact the naturalness of the sentences. This further suggests that the difference we observed between *DISTANT MATCH* and *LOCAL MATCH* in the judgments and reading times in Experiment 1 reflect aspects of the processing of *ziji*, rather than unacceptability that results from the presence of inanimate main clause subjects in the *DISTANT MATCH* condition.

Turning to the reading times, statistical modeling of the results yields several important insights. First, although both the *Locality contrast* was significant at the spillover region for both *ziji* and *ta-ziji*, there was a significant interaction of *Locality* and *Experiment*: the magnitude of the locality effect was much smaller for *ta-ziji* than for *ziji*. Although the locality effect was several times smaller for *ta-ziji* than for *ziji*, *post-hoc* analysis revealed that the *Locality contrast* was significant for both anaphors.

However, apart from this crucial difference, the processing of both anaphors was qualitatively similar. Our analysis revealed a significant *No Antecedent contrast* that did not differ in magnitude across studies. This indicates that comprehenders did indeed try to assign a referent to the anaphor upon encountering it, and moreover, it suggests that comprehenders experienced a similar amount of processing difficulty when there was no sentence-internal antecedent for both *ziji* and *ta-ziji*. Likewise, the magnitude of the *Multiple Match contrast* was similar across the two experiments, although it failed to reach statistical significance either in the omnibus analysis, or in the *post-hoc* analyses by experiment.

GENERAL DISCUSSION

In two self-paced reading experiments, we investigated the processing of two reflexive anaphors in Mandarin: the bare monomorphemic reflexive *ziji*, and the morphologically complex reflexive *ta-ziji*. In both offline acceptability rating and online reading time results, we observed that *ziji* was associated with a robust locality bias. Non-local interpretation of *ziji* was associated with lower naturalness ratings and longer reading times. In contrast, we observed a significantly smaller locality effect for *ta-ziji* in reading times, and no locality effect in offline acceptability judgments. Interestingly, this was the only difference we observed between *ziji* and *ta-ziji*. For both anaphors, we observed reliable reading time slowdowns when there was no licit antecedent in the sentence, and the size of this *no match* penalty did not reliably differ between anaphors in reading time measures. Likewise, for both anaphors we observed a trend toward a multiple match penalty. This effect did not reach statistical significance, although the consistency of the effect in sign and magnitude across experiments raises the possibility that the failure to observe this effect reflects a lack of statistical power. We take up each of these effects in turn.

Feature Richness and Antecedent Search

Our findings suggest that the locality bias that is associated with *ziji* does not generalize to other Mandarin reflexives that can take antecedents outside of their immediate tensed clause. Specifically, the morphologically complex anaphor *ta-ziji* shows a much diminished locality bias in online processing measures. One plausible hypothesis about this difference is that the overt morphological feature content on *ta-ziji* leads to faster or more reliable access to structurally distant antecedents. In contrast, *ziji* has fewer overt morphological cues to its antecedent, and so comprehenders may need to rely more heavily on positional cues to isolate its antecedent in memory, leading to relatively more pronounced locality bias.

This hypothesis is plausible given existing theories of how comprehenders access information in working memory during sentence comprehension. For a wide range of linguistic dependencies, there is evidence that the processor makes use of a content-addressable retrieval mechanism to form syntactic and referential dependencies between temporally distant phrases (McElree, 2000, 2006, 2014; McElree et al., 2003; Lewis and Vasishth, 2005; Lewis et al., 2006; McElree, 2006; Van Dyke and McElree, 2006, 2011; Foraker and McElree, 2011). A content-addressable retrieval mechanism accesses a representation in memory using the inherent features of the representation as cues to guide memory retrieval process. For example, a pronoun like *him* may be said to retrieve its antecedent by using gender features as cues to locate an antecedent in memory (e.g., Foraker and McElree, 2007). These cues are said to provide *direct access* to the desired representation, obviating any need to search through irrelevant representations at retrieval. This mechanism has the benefit of granting extremely rapid access to information in memory, making this an attractive mechanism for memory access in the human sentence processor (Lewis et al., 2006). In general, models that posit a content-addressable retrieval mechanism predict that an increase in structural or linear distance between the retrieval site (e.g., the anaphor) and the target of retrieval (e.g., its antecedent) may lead to reduced retrieval accuracy, but that structural or linear distance *per se* should not result in longer retrieval times. On some theoretical proposals, the speed of retrieval may be modulated by variables such as retrieval interference and temporal decay (Lewis and Vasishth, 2005; Lewis et al., 2006); on others, these variables primarily impact the probability of successfully recovering a target representation (McElree, 2006). Although cue-based models have the advantage of offering rapid access to representations in memory when they are required for sentence comprehension, they encounter difficulty if multiple representations in memory match the retrieval cues used at retrieval. If this occurs, it may be more difficult to isolate the *target* representation in memory, a phenomenon known as *retrieval interference*. Retrieval interference has been shown to be a primary cause of difficulty in sentence comprehension (Lewis, 1996; Van Dyke and Lewis, 2003; Lewis and Vasishth, 2005; Lewis et al., 2006; Van Dyke and McElree, 2006; Van Dyke, 2007; see also Gordon et al., 2001, 2002, 2004; for a recent review, see Van Dyke and Johns, 2012).

Several distinct hypotheses that draw upon this basic framework have been proposed to explain the locality bias associated with *ziji*. Chen et al. (2012) and Jäger et al. (2015) offered an account that draws on the ACT-R model of Lewis and Vasishth (2005). This model explains the locality bias as the result of decay and interference reducing the activation level of the distant subject, which in turn leads to longer antecedent retrieval times when the antecedent is distant from the anaphor. An alternative explanation is offered by Dillon et al. (2014), who proposed that the locality effect arises because comprehenders tend to initially retrieve the local subject as an antecedent for *ziji*. On this view, comprehenders must reject the local subject as a plausible antecedent and execute additional memory retrievals to access a distant antecedent for *ziji*. On this view, more retrieval operations are necessary to access distant antecedents, and so it is predicted that processing times should increase whenever *ziji* needs to take an antecedent other than the most local one.

One finding that distinguishes these two accounts is the MR-SAT study reported by Dillon et al. (2014). Dillon and colleagues observed that *ziji* with distant antecedents led to significantly slower processing rates in the speed-accuracy tradeoff function than *ziji* with local antecedents. Standard memory variables such as temporal decay and interference alone have not been shown to modulate processing speed in the speed-accuracy tradeoff functions associated with linguistic processing (McElree, 2000; McElree et al., 2003; Foraker and McElree, 2007, 2011; Martin and McElree, 2008, 2009, 2011; Van Dyke and McElree, 2011). However, processing speed as measured in the speed-accuracy tradeoff function has been shown to be slowed down by increasing the number of required retrieval operations, and in situations where syntactic reanalysis is required (McElree et al., 2003; Bornkessel et al., 2004; Foraker and McElree, 2007). Thus, the MR-SAT data lend support to the view that the locality effect arises because comprehenders are tempted to initially retrieve and evaluate the local subject as an antecedent when processing *ziji*.

This account also offers some insight into how the overt feature content of *ta-ziji* may allow comprehenders to overcome locality bias in comprehension. Many different implementations of cue-based retrieval mechanisms predict that the more highly specified a retrieval probe is in terms of the cues used, the less likely it is that partially matching (*distractor*) items in memory will cause retrieval interference, and compete with other representations at retrieval (Lewis and Vasishth, 2005; Van Dyke and McElree, 2006; Van Dyke, 2007). These models hold that retrieval probes that contain a greater number of retrieval cues will see a corresponding increase in the probability of recovering a target item in memory, because more numerous and specific retrieval cues will in general decrease the probability of retrieving a distractor that only matches a subset of the retrieval cues.

To illustrate how these models would account for the difference in the magnitude of the locality effect for *ziji* and *ta-ziji*, consider again the critical configuration in (9):

(9) 张 太太 经常 光顾 的那个时装店
Zhang taitai$_i$ jingchang guanggu de na-ge shizhuangdian$_j$
上个星期 把自己 不小心 弄伤了
shang-ge-xingqi ba ziji$_{i/*j}$ bu xiaoxin nongshang-le.
Mrs. Zhang often visit DE that-CL boutique last-week BA self not careful harm-PERF.
"The boutique that Mrs. Zhang often visits carelessly harmed her last week."

Upon reaching the anaphor, comprehenders will, by hypothesis, recruit a mixture of syntactic cues (e.g., cues to subjecthood, such as syntactic case) and semantic or morphological cues (e.g., animacy in the case of *ziji*; animacy, humanness, and perhaps gender in the case of *ta-ziji*). Although these cues form a perfect match to the target antecedent, the inappropriate local subject *shizhuangdian* (boutique) matches only the syntactic cues. Thus, it is possible for the local antecedent to be mis-retrieved some proportion of the time, because it partially matches the syntactic cue content of the retrieval probe. Although we might reasonably expect the semantically appropriate long-distance antecedent to outcompete the local subject at retrieval in many cases, the competition contributed by the local subject may be exacerbated by its recency or its structural proximity to the anaphor (Dillon et al., 2014). If the retrieval probe contains relatively few semantic cues, the likelihood of mis-retrieving the local subject may be relatively high. On the account offered by Dillon et al. (2014), this is precisely what happens when comprehenders process *ziji*: although *ziji* contains animacy cues, these are not enough to overcome interference from the local subject, and so the local subject is retrieved some proportion of the time. When this occurs, comprehenders must attempt additional retrievals in order to arrive at an acceptable interpretation of the anaphor.

In the case of *ta-ziji*, the addition of humanness and gender features into the retrieval probe ensures that the local subject *na-ge shizhuangdian* "that boutique" matches fewer retrieval cues in the probe. This decreases the probability of retrieving the partially matching local subject, resulting in a greater proportion of trials when comprehenders are able to access the desired antecedent without sampling multiple antecedent representations from memory. Thus, if the locality bias reflects a tendency to mis-retrieve and consider the local subject, rather than decay of the distant antecedent *per se*, then the locality bias is predicted to be smaller for *ta-ziji* than for *ziji*. Put simply, *ta-ziji*'s additional feature content decreases the attractiveness of the local subject as a distractor and makes it more likely that comprehenders will successfully retrieve the long-distance antecedent on their first attempt.

Although, we have offered an explanation of our results in terms of the likelihood that the local subject will be (mis-)retrieved when processing the anaphor, it remains to be seen whether accounts that explain the locality bias effect as decreased activation of the distant antecedent can account for the difference between *ziji* and *ta-ziji*. The contrast between these anaphors rules out the simple hypothesis that the locality bias associated with *ziji* is due to recency or temporal decay alone. This is because this hypothesis would predict an equal locality effect for both anaphors. However, it may be possible to capture the present finding in more sophisticated models where the activation of an item in memory is partially a function of the retrieval cues used to access memory, such as the ACT-R model of Lewis and Vasishth (2005). It is difficult to evaluate the predictions of these models without the aid of an implemented computational model. The predictions of this account vary substantially with specific modeling assumptions that one makes. For example, if one assumes that the distant antecedent in examples like (9) is a perfect match to the retrieval probe of *ziji* and *ta-ziji* alike, then the activation of the distant antecedent should not be modulated by the number of retrieval cues in the retrieval probe[1] . This is because the total activation boost that a retrieval probe gives to an item in memory is constant in ACT-R. Adding more retrieval cues to the probe therefore does not increase the amount of activation afforded to a perfectly matching item in memory; it instead diminishes the amount of activation boost that is contributed by any one cue on its own. Under these conditions, availability of the distant antecedent is not predicted to differ between *ziji* and *ta-ziji*, all else being equal. However, if one relaxes these assumptions, it may be possible to capture this result. Thus, although we cannot claim that the present results are incompatible with the explanation of the locality bias effect offered by Chen et al. (2012) and Jäger et al. (2015), more research and modeling work is necessary to determine the specific circumstances under which these models can capture the contrast between *ziji* and *ta-ziji*.

Ziji vs. *Ta-ziji*

The explanation we offer for our findings posits that overt morphological features provide retrieval cues for recovering an antecedent for an anaphoric expression. However, the precise relationship between overt morphological form and the cues used to retrieve an antecedent remains unclear. Indeed, the problem of specifying the nature of the retrieval cues that support language processing is a key theoretical issue for cue-based approaches, and remains an area where much further work is needed (Van Dyke and McElree, 2011; Dillon et al., 2013). Previous research suggests that the link between overt morphological feature content and retrieval cues may be rather indirect, such that there is not a one-to-one mapping between overt morphological features and retrieval cues. For example, Dillon et al. (2013) presented a series of studies that investigated the processing of English reflexive *himself*. On the basis of a comparison between computational retrieval models and online reading time data, they suggest that *himself* does not use gender and number features as cues to retrieve its antecedent, instead relying on a mixture of structural and locality cues (Dillon et al., 2013; Dillon, 2014; but see Jäger et al., 2015, for a critical view of this conclusion and an alternative analysis of these findings). Put differently, Dillon et al. (2013) proposed that the morphologically complex English reflexive *himself* deploys a cue set that is fundamentally similar to the cue set proposed for Mandarin *ziji*, despite the fact

[1]We are grateful to Lena Jäger for this observation.

that *himself* is morphologically more similar to Mandarin *ta-ziji* than it is to *ziji*. This contrast suggests that the simple hypothesis that overt morphological features are recruited as retrieval cues may not be correct, but we presently lack a theory of how morphological and semantic features of anaphoric expressions are used during antecedent retrieval.

Resolving this tension is beyond the scope of this paper, but there are several possibilities that suggest themselves. It may be that the direct link between morphology and retrieval cues that we offer as an explanation for the present findings is misguided, and some other difference between *ziji* and *ta-ziji* is responsible for the difference in their processing behavior. One plausible alternative explanation of our result could leverage the observation that *ta-ziji* may be readily interpreted as a contrastive reflexive, analogous to *he himself* in English (Pan, 1998; what Baker, 1995 calls an *intensive pronoun*). Baker (1995) suggests that the interpretive constraints on these intensive pronouns are best understood in terms of discourse prominence and contrastiveness of potential antecedents, rather than their syntactic positions. It is possible that the diminished locality effect for *ta-ziji* reflects a preference to construe *ta-ziji* as an intensive pronoun in our experiment. This could have caused readers to weight discourse cues more heavily than syntactic cues when retrieving an antecedent for *ta-ziji*, leading to a smaller locality effect. While we find this an interesting possibility, we cannot confidently endorse it on the basis of the present data because it is at present unclear whether readers understood *ta-ziji* as an anaphor or an intensive pronoun in our experiment. Another possibility is that the difference in locality bias reflects a difference in the frequency with which each anaphor takes antecedents beyond its local clause. Although we cannot rule out this possibility, we find it unlikely: the number of syntactic environments where *ta-ziji* can find an antecedent outside of its tensed clause is much smaller than the number of environments where *ziji* can, making it unlikely that *ziji* more often takes a local antecedent than *ta-ziji* in a Mandarin speaker's language experience. Nonetheless, corpus work would be necessary to secure this conclusion, and at present we must regard it as a possible, but unlikely, explanation of the present finding.

Multiple Match Effects

A further prediction of cue-based retrieval models is that the presence of multiple antecedents that match the retrieval cues of the anaphor should create retrieval interference, which should in turn create processing difficulty. For example, consider our *MULTIPLE MATCH* condition:

(10) 张 太太 经常 光顾 的 那个 女裁缝
Zhang taitai$_i$ jingchang guanggu de na-ge nücaifeng$_j$
上个星期 把自己 不小心 弄伤了
shang-ge-xingqi ba ziji$_{*i/j}$ buxiaoxin nongshang-le.
Mrs. Zhang often visit DE that-CL seamstress last-week ba self not careful harm-PERF.
"The seamstress that Mrs. Zhang often visits carelessly harmed her last week."

This sentence is not ambiguous, as the animate head noun *nücaifeng* ("seamstress") blocks access to the embedded subject *Zhang taitai* ("Mrs. Zhang"; see Tang, 1989). In other words, *Zhang taitai* is a grammatically inaccessible distractor in this example, even though this syntactic position would have been grammatically accessible if the relative clause's head noun were inanimate. Nonetheless, because the distant subject *Zhang taitai* matches the retrieval cues of the anaphor, it is predicted to create retrieval interference. Because there are multiple antecedents that match the animacy cues associated with *ziji*, it should be more difficult for comprehenders to isolate the correct antecedent *nücaifeng* in memory. In terms of our experimental manipulations, cue-based parsing models broadly predict that the *MULTIPLE MATCH* condition should be more difficult than the *LOCAL MATCH* condition at the reflexive, because the *MULTIPLE MATCH* condition contains more representations that match the retrieval cues, contributing to retrieval interference in the *MULTIPLE MATCH* condition that should inhibit access to the target antecedent. Moreover, it is predicted that the size of the multiple match effect should be greater for *ta-ziji* than for *ziji*, because the distractor matches a greater proportion of the retrieval cues for the complex anaphor.

Our experiments failed to provide clear evidence to support or disconfirm these predictions. In both Experiments 1 and 2, we observed numerically longer reading times on the *MULTIPLE MATCH* condition than in the *LOCAL MATCH* condition, but this contrast did not reach statistical significance in either experiment alone, and we failed to observe any trend toward a larger multiple match effect for *ta-ziji*. The interpretation of the present findings, and their relationship with previous findings, must therefore be treated with caution. We note that the size and magnitude of the multiple match effect was consistent, and in the predicted direction, in both experiments. This pattern suggests that our failure to find an effect may reflect a lack of statistical power.

Although a multiple match penalty has been observed in previous reading time studies (Badecker and Straub, 1994, 2002; Kennison et al., 2003; Felser et al., 2009; Chen et al., 2012; Jäger et al., 2015; see also Rigalleau et al., 2004; Stewart et al., 2007), the empirical generalization has remained unclear, especially for reflexive pronouns. Although some studies have presented evidence that feature-matched, but structurally inaccessible antecedents create processing difficulty for reflexive pronouns (Badecker and Straub, 2002; Felser et al., 2009; Chen et al., 2012; Jäger et al., 2015), many more studies have failed to find reliable evidence for such an effect in reading time measures, or found it only in limited contexts (Clifton et al., 1999; Sturt, 2003; Xiang et al., 2009; Cunnings and Felser, 2013; Dillon et al., 2013; Cunnings and Sturt, 2014; Kush and Phillips, 2014; Jäger et al., 2015). It is notable that Chen and colleagues reported that reading times on *ziji* were longer in the presence of a grammatically inaccessible, but semantically appropriate antecedent (c.f., Jäger et al., 2015). On balance, however, the repeated failures to find multiple match effects suggest that grammatically illicit antecedents do not create substantial interference effects, and so our failure to find

any multiple match effects in the present study may not be surprising. Dillon (2014) and Sturt (2013) offer reviews of the empirical landscape, and suggest that grammatical constraints act as strong filters on antecedent retrieval, allowing for very little (if any) retrieval interference from grammatically illicit antecedents.

A similar pattern emerges for studies that have focused on the processing of direct object pronouns in English. Chow et al. (2014) reported five experiments that sought to find multiple match effects with direct object pronouns in English, and failed to find any evidence of such an effect (including a near direct replication of Badecker and Straub, 2002). On the basis of this finding, Chow and colleagues argued that, in line with the processing of reflexives, structural constraints acted immediately to help rule out grammatically inaccessible antecedents for object pronouns as well (see also Clifton et al., 1997; Lee and Williams, 2008; Patterson et al., 2014).

On the basis of the non-significant multiple match effects in the present studies, very little can be concluded about whether the inaccessible antecedent in our *MULTIPLE MATCH* conditions created retrieval interference. However, inconsistent with claims that grammatical constraints rule out inaccessible antecedents, we did find clear trends in the predicted direction in both experiments. We thus regard it as an open empirical question whether an animate distant subject interferes with the retrieval of the correct local subject when processing *ziji* and *ta-ziji*.

CONCLUSION

In our Experiment 1, we observed that the morphologically simple long-distance reflexive *ziji* showed a robust locality bias in reading time measures. Experiment 2 revealed that the morphologically complex, local reflexive *ta-ziji* showed a much reduced locality bias in processing. We proposed that this contrast was due to the number of morphological and semantic features each anaphor uses during the process of retrieving an antecedent. Morphologically simple anaphors like *ziji*, which have relatively fewer retrieval cues, are more likely to access non-target antecedents at retrieval. This requires comprehenders to sample multiple antecedents in order to achieve an interpretation for the anaphor, leading to locality effects. In contrast, the relatively more specified *ta-ziji* has more cues for antecedent retrieval, which makes it less susceptible to interference from non-target representations. For this reason, complex anaphors like *ta-ziji* show diminished locality effects in comprehension.

ACKNOWLEDGMENTS

We would like to thank Lyn Frazier, Lena Jäger, and Dave Kush for comments on and discussion of an earlier version of this work, and Colin Phillips for extended discussion of the experiments presented here. This work was supported by a University of Maryland Fellowship to BD and WC. Experiments 1 and 2 were initially presented as part of Dillon (2011).

REFERENCES

Badecker, W., and Straub, K. (1994). "Evidence that binding principles participate in a constraint satisfaction process," in *Poster Presented at the 7th CUNY Sentence Processing Conference* (New York, NY).

Badecker, W., and Straub, K. (2002). The processing role of structural constraints on interpretation of pronouns and anaphors. *J. Exp. Psychol. Learn. Mem. Cogn.* 28, 748. doi: 10.1037/0278-7393.28.4.748

Bakeman, R., and McArthur, D. (1996). Picturing repeated measures: comments on Loftus, Morrison, and others. *Behav. Res. Methods Instrum. Comput.* 28, 584–589. doi: 10.3758/BF03200546

Baker, C. L. (1995). Contrast, discourse prominence, and intensification, with special reference to locally free reflexives in British English. *Language* 71, 63–101. doi: 10.2307/415963

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: keep it maximal. *J. Mem. Lang.* 68, 255–278. doi: 10.1016/j.jml.2012.11.001

Bornkessel, I., McElree, B., Schlesewsky, M., and Friederici, A. D. (2004). Multi-dimensional contributions to garden path strength: dissociating phrase structure from case marking. *J. Mem. Lang.* 51, 495–522. doi: 10.1016/j.jml.2004.06.011

Box, G. E., and Cox, D. R. (1964). An analysis of transformations. *J. R. Stat. Soc. B* 26, 211–252.

Büring, D. (2005). *Binding Theory*. Cambridge: Cambridge University Press.

Cai, Q., and Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PLoS ONE* 5:e10729. doi: 10.1371/journal.pone.0010729

Chen, Z., Jäger, L., and Vasishth, S. (2012). How structure-sensitive is the parser? Evidence from Mandarin Chinese. *Emp. Approaches Linguist. Theory*111:43. doi: 10.1515/9781614510888.43

Chomsky, N. (1981). *Lectures on Government and Binding*. Dordrecht: Foris.

Chow, W. Y., Lewis, S., and Phillips, C. (2014). Immediate sensitivity to structural constraints in pronoun resolution. *Front. Psychol.* 5:630. doi: 10.3389/fpsyg.2014.00630

Clifton, C., Frazier, L., and Deevy, P. (1999). Feature manipulation in sentence comprehension. *Riv. Lingüíst.* 11, 11–40.

Clifton, C., Kennison, S. M., and Albrecht, J. E. (1997). Reading the wordsher, his, him: implications for parsing principles based on frequency and on structure. *J. Mem. Lang.* 36, 276–292. doi: 10.1006/jmla.1996.2499

Cunnings, I., and Felser, C. (2013). The role of working memory in the processing of reflexives. *Lang. Cogn. Process.* 28, 188–219. doi: 10.1080/01690965.2010.548391

Cunnings, I., and Sturt, P. (2014). Coargumenthood and the processing of reflexives. *J. Mem. Lang.* 75, 117–139. doi: 10.1016/j.jml.2014.05.006

Dillon, B. (2011). *Structured Access in Sentence Comprehension*. Unpublished Ph.D. dissertation, University of Maryland.

Dillon, B. (2014). Syntactic memory in the comprehension of reflexive dependencies: an overview. *Lang. Linguist. Compass* 8, 171–187. doi: 10.1111/lnc3.12075

Dillon, B., Chow, W. Y., Wagers, M., Guo, T., Liu, F., and Phillips, C. (2014). The structure-sensitivity of memory access: evidence from Mandarin Chinese. *Front. Psychol.* 5:1025. doi: 10.3389/fpsyg.2014.01025

Dillon, B., Mishler, A., Sloggett, S., and Phillips, C. (2013). Contrasting intrusion profiles for agreement and anaphora: experimental and modeling evidence. *J. Mem. Lang.* 69, 85–103. doi: 10.1016/j.jml.2013.04.003

Drummond, A. (2011). *IbexFarm (Version 0.2.7) [Software]*. Available online at: http://spellout.net/ibexfarm/

Elbourne, P. D. (2008). The interpretation of pronouns. *Lang. Linguist. Compass* 2, 119–150. doi: 10.1111/j.1749-818X.2007.00041.x

Felser, C., Sato, M., and Bertenshaw, N. (2009). The on-line application of binding Principle A in English as a second language. *Bilingualism Lang. Cogn.* 12, 485–502. doi: 10.1017/S1366728909990228

Foraker, S., and McElree, B. (2007). The role of prominence in pronoun resolution: availability versus accessibility. *J. Mem. Lang.* 56, 357–383. doi: 10.1016/j.jml.2006.07.004

Foraker, S., and McElree, B. (2011). Comprehension of linguistic dependencies: speed−accuracy tradeoff evidence for direct−access retrieval from memory. *Lang. Linguist. Compass* 5, 764–783. doi: 10.1111/j.1749-818X.2011.00313.x

Gao, L., Liu, Z., and Huang, Y. (2005). Who is *ziji*? An experimental research on Binding Principle. *Linguist. Sci.* 2, 39–50.

Gernsbacher, M. A. (1989). Mechanisms that improve referential access. *Cognition* 32, 99–156. doi: 10.1016/0010-0277(89)90001-2

Gordon, P. C., Hendrick, R., and Johnson, M. (2001). Memory interference during sentence processing. *Psychol. Sci.* 13, 425–430. doi: 10.1111/1467-9280.00475

Gordon, P. C., Hendrick, R., and Johnson, M. (2004). Effects of noun phrase type on sentence complexity. *J. Mem. Lang.* 51, 97–114. doi: 10.1016/j.jml.2004.02.003

Gordon, P. C., Hendrick, R., and Levine, W. H. (2002). Memory-load interference in syntactic processing. *Psychol. Sci.* 13, 425–430. doi: 10.1111/1467-9280.00475

Greene, S. B., McKoon, G., and Ratcliff, R. (1992). Pronoun resolution and discourse models. *J. Exp. Psychol. Learn. Mem. Cogn.* 18:266. doi: 10.1037/0278-7393.18.2.266

Heim, I. (1982). *The Semantics of Definite and Indefinite Noun Phrases*. Ph.D. dissertation, University of Massachusetts, Amherst.

Huang, C.-T. J., Cole, P., and Hermon, G. (2006). "Long-distance reflexives: an East Asian perspective," in *The Blackwell Companion to Syntax*, Volu. III, eds M. Everaert, H. van Riemsdijk, R. Goedemans, and B Hollebrandse (Location: Blackwell), 21–84.

Huang, C.-T. J., and Liu, L. (2001). "Logophoricity, attitudes, and *ziji* at the interface," in in Long *Distance Reflexives, Syntax and Semantics* 33, eds P. Cole, C.-T. J. Huang, and G. Hermon (New York, NY: Academic Press), 141–195.

Huang, C.-T. J., Li, Y.-H. A., and Li, Y. (2009). *The Syntax of Chinese*. Cambridge, UK: Cambridge University Press.

Huang, C. T. J., and Tang, C. C. J. (1991). "The local nature of the long-distance reflexive in Chinese," in *Long Distance Anaphora*, eds J. Koster and E. Reuland (Cambridge: Cambridge University Press), 263–283. doi: 10.1017/CBO9780511627835.014

Jäger, L. A., Engelmann, F., and Vasishth, S. (2015). Retrieval interference in reflexive processing: experimental evidence from Mandarin, and computational modeling. *Front. Psychol.* 6:617. doi: 10.3389/fpsyg.2015.00617

Kennison, S. M. (2003). Comprehending the pronouns *her*, *him*, and *his*: implications for theories of referential processing. *J. Mem. Lang.* 49, 335–352. doi: 10.1016/S0749-596X(03)00071-8

Kintsch, W. (1975). *Memory Representations of Text. Information Processing and Cognition*. Hillsdale, NJ: Erlbaum.

Kush, D. (2013). *Respecting Relations: Memory Access and Antecedent Retrieval in Incremental Sentence Processing*. Unpublished doctoral dissertation, University of Maryland, College Park.

Kush, D., and Phillips, C. (2014). Local anaphor licensing in an SOV language: implications for retrieval strategies. *Front. Psychol.* 5:1252. doi: 10.3389/fpsyg.2014.01252

Lee, M. W., and Williams, J. N. (2008). *The Role of Grammatical Constraints in Intra-Sentential Pronoun Resolution*. London: London Metropolitan University/Cambridge University manuscript.

Lewis, R., and Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cogn. Sci.* 29, 375–419. doi: 10.1207/s15516709cog0000_25

Lewis, R., Vasishth, S., and Van Dyke, J. (2006). Computational principles of working memory in sentence comprehension. *Trends Cogn. Sci.* 10, 447–454. doi: 10.1016/j.tics.2006.08.007

Lewis, R. L. (1996). Interference in short-term memory: the magical number two (or three) in sentence processing. *J. Psycholinguist. Res.* 25, 93–115. doi: 10.1007/BF01708421

Li, X., and Zhou, X. (2010). Who is *ziji*? ERP responses to the Chinese reflexive pronouns during sentence comprehension. *Brain Res.* 1331, 96–104. doi: 10.1016/j.brainres.2010.03.050

Liu, Z. (2009). The cognitive process of Chinese reflexive processing. *J. Chinese Linguist.* 37, 1–27.

Martin, A. E., and McElree, B. (2008). A content-addressable pointer mechanism underlies comprehension of verb-phrase ellipsis. *J. Mem. Lang.* 58, 879–906. doi: 10.1016/j.jml.2007.06.010

Martin, A. E., and McElree, B. (2009). Memory operations that support language comprehension: evidence from verb-phrase ellipsis. *J. Exp. Psychol. Learn. Mem. Cogn.* 35, 1231–1239. doi: 10.1037/a0016271

Martin, A. E., and McElree, B. (2011). Direct-access retrieval during sentence comprehension: evidence from Sluicing. *J. Mem. Lang.* 64, 327–343. doi: 10.1016/j.jml.2010.12.006

McElree, B. (2000). Sentence comprehension is mediated by content-addressable memory structures. *J. Psycholinguist. Res.* 29, 111–123. doi: 10.1023/A:1005184709695

McElree, B. (2006). Accessing recent events. *Psychol. Learn. Mot.* 46, 155–200. doi: 10.1016/S0079-7421(06)46005-9

McElree, B. (2014). "Memory processes underlying real-time language comprehension," in *Remembering: Attributions, Processes, and Control in Human Memory*, eds D. S. Lindsay, C. M. Kelley, A. P. Yonelinas and H. L. Roediger (New York, NY: Psychology Press), 133–151.

McElree, B., Foraker, S., and Dyer, L. (2003). Memory structures that subserve sentence comprehension. *J. Mem. Lang.* 48, 67–91. doi: 10.1016/S0749-596X(02)00515-6

Myers, J. L., and O'Brien, E. J. (1998). Accessing the discourse representation during reading. *Discourse Process.* 26, 131–157. doi: 10.1080/01638539809545042

Pan, H. (1998). Closeness, prominence, and binding theory. *Nat. Lang. Linguist. Theory* 16, 771–815. doi: 10.1023/A:1006056414208

Pan, H. (2000). Why the blocking effect. *Syntax Semant.* 33, 279–316. doi: 10.1108/S0092-4563(2000)0000033011

Patterson, C., Trompelt, H., and Felser, C. (2014). The online application of binding condition B in native and non-native pronoun resolution. *Front. Psychol.* 5:147. doi: 10.3389/fpsyg.2014.00147

Rigalleau, F., Caplan, D., and Baudiffier, V. (2004). New arguments in favour of an automatic gender pronominal process. *Q. J. Exp. Psychol. A*, 57, 893–933. doi: 10.1080/02724980343000549

Rohde, D. (2003). *Linger (Version 2.94) [Software]*. Available online at: http://tedlab.mit.edu/~dr/Linger/

Sprouse, J., Fukuda, S., Ono, H., and Kluender, R. (2011). Reverse island effects and the backward search for a licensor in multiple *wh*−questions. *Syntax* 14, 179–203. doi: 10.1111/j.1467-9612.2011.00153.x

Stewart, A. J., Holler, J., and Kidd, E. (2007). Shallow processing of ambiguous pronouns: evidence for delay. *Q. J. Exp. Psychol.* 60, 1680–1696. doi: 10.1080/17470210601160807

Sturt, P. (2003). The time-course of the application of binding constraints in reference resolution. *J. Mem. Lang.* 48, 542–562. doi: 10.1016/S0749-596X(02)00536-3

Sturt, P. (2013). "Constraints on referential processing," in *Sentence Processing*, ed R. van Gompel (London: Routledge), 136–159.

Tang, C. C. J. (1989). Chinese reflexives. *Nat. Lang. Linguist. Theory* 7, 93–121. doi: 10.1007/BF00141348

Van Dyke, J. A. (2007). Interference effects from grammatically unavailable constituents during sentence processing. *J. Exp. Psychol. Learn. Mem. Cogn.* 33, 407–430. doi: 10.1037/0278-7393.33.2.407

Van Dyke, J. A., and Johns, C. L. (2012). Memory interference as a determinant of language comprehension. *Lang. Linguist. Compass* 6, 193–211. doi: 10.1002/lnc3.330

Van Dyke, J. A., and Lewis, R. L. (2003). Distinguishing effects of structure and decay on attachment and repair: a cue-based parsing account of recovery from misanalyzed ambiguities. *J. Mem. Lang.* 49, 285–316. doi: 10.1016/S0749-596X(03)00081-0

Van Dyke, J. A., and McElree, B. (2006). Retrieval interference in sentence processing. *J. Mem. Lang.* 55, 157–166. doi: 10.1016/j.jml.2006.03.007

Van Dyke, J. A., and McElree, B. (2011). Cue-dependent interference in comprehension. *J. Mem. Lang.* 65, 247–263. doi: 10.1016/j.jml.2011.05.002

Van Gompel, R. P., and Majid, A. (2004). Antecedent frequency effects during the processing of pronouns. *Cognition* 90, 255–264. doi: 10.1016/S0010-0277(03)00161-6

Xiang, M., Dillon, B., and Phillips, C. (2009). Illusory licensing effects across dependency types: ERP evidence. *Brain Lang.* 108, 40–55. doi: 10.1016/j.bandl.2008.10.002

Xue, P., Pollard, C., and Sag, I. (1994). "A new perspective on Chinese ziji," in *Proceedings of the 13th Annual Meeting of the West Coast Conference on Formal Linguistics* (Stanford, CA: CSLI), 432–447.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2016 Dillon, Chow and Xiang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 14 April 2016
doi: 10.3389/fpsyg.2016.00284

Processing the Chinese Reflexive "ziji": Effects of Featural Constraints on Anaphor Resolution

Xiao He[1] *and Elsi Kaiser*[2]*

[1] *ProSearch Strategies, Inc., Los Angeles, CA, USA,* [2] *Department of Linguistics, University of Southern California, Los Angeles, CA, USA*

We present three self-paced reading experiments that investigate the reflexive *ziji* "self" in Chinese—in particular, we tested whether and how person-feature-based blocking guides comprehenders' real-time processing and final interpretation of *ziji*. Prior work claims that in Chinese sentences like "John thought that {I/you/Bill} did not like ZIJI," (i) the reflexive ziji can refer to the matrix subject John if the intervening subject is also a third person entity (e.g., Bill), but that (ii) an intervening first or second person pronoun blocks reference to the matrix subject, causing *ziji* to refer to the first or second person pronoun. However, native speakers' judgments regarding the accessibility of long-distance antecedents are rather unstable, and researchers also disagree on what the exact configurations are that allow blocking. In addition, many open questions persist regarding the real-time processing of reflexives more generally, in particular regarding the accessibility (or lack thereof) of structurally unlicensed antecedents. We conducted three self-paced reading studies where we recorded people's word-by-word reading times and also asked questions that probed their off-line interpretation of the reflexive *ziji*. People's answers to the off-line questions show that blocking is not absolute: Comprehenders do allow significant numbers of non-local choices in both the first and the second person blocking conditions, albeit in small numbers. At the same time, the reading time data, particularly those from Experiments 2 and 3, show that comprehenders use person feature cues to quickly filter out inaccessible long-distance referents. The difference between on-line and off-line patterns points to the possibility that the interpretation of *ziji* unfolds over time: it seems that initially, during real-time processing, person-feature cues weigh more heavily and constrain what antecedent candidates get considered, but that at some later point, other kinds of information are also integrated and perhaps outweigh the person-feature constraint, resulting in consideration of referents that were initially "blocked" due to the person-feature constraint. In sum, in addition to the structural constraints identified in prior work, person-featural cues also play a key role in regulating the on-line processing of reflexives in Chinese.

Keywords: sentence processing, reflexive pronouns, Chinese, self-paced reading, blocking effects, binding theory

OPEN ACCESS

Edited by:
Claudia Felser,
University of Potsdam, Germany

Reviewed by:
Petra B. Schumacher,
University of Cologne, Germany
Zhong Chen,
Rochester Institute of Technology, USA

***Correspondence:**
Elsi Kaiser
emkaiser@usc.edu

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: *07 October 2015*
Accepted: *14 February 2016*
Published: *14 April 2016*

Citation:
He X and Kaiser E (2016) Processing the Chinese Reflexive "ziji": Effects of Featural Constraints on Anaphor Resolution. Front. Psychol. 7:284. doi: 10.3389/fpsyg.2016.00284

INTRODUCTION

In real-time language comprehension, a major challenge faced by comprehenders is the need to resolve dependency relationships, where the interpretation of one linguistic element depends on another. One such case is the interpretation of reflexive pronouns (e.g., *himself/herself*), which is traditionally argued to depend on a set of structural constraints. In English, for example, reflexives are constrained by a set of rules termed Binding Principle A (Chomsky, 1981, 1986). According to this principle, a reflexive [*himself* in (1)] can only refer to a referent within the local clause (*Bill*) and not a referent outside the local clause (*John*).

(1) John_1 said that Bill_2 disliked $\text{himself}_{*1/2}$.

While structural constraints seem to adequately capture the patterning of reflexives in many contexts, their influence on the real-time processing of reflexives is the subject of an ongoing debate. Specifically, researchers disagree on whether or not structural information has an immediate effect on what referents are considered potential antecedents. Early work by Nicol and Swinney (1989) and Sturt (2003), among others, showed that comprehenders' consideration of potential antecedent candidates is immediately determined by structural constraints. More recent evidence in the same direction comes from Xiang et al. (2009), Dillon et al. (2013) and others. Hence, according to these studies, structurally-incompatible/inaccessible referents [such as *John* in (1)] do not cause interference during the processing of reflexives.

However, other studies found that comprehenders do not fully abide by structural rules, at least in the early stage of processing (e.g., Badecker and Straub, 2002; Runner et al., 2006; Kaiser et al., 2009; Clackson and Heyer, 2014). These findings suggest that initial consideration of possible antecedents can be influenced by featural properties of potential referents (which can act as retrieval cues), such as person, animacy, and number, and that comprehenders at least temporarily consider feature-compatible but structurally inappropriate referents before they eventually reach the correct interpretation. Thus, opinions diverge regarding the role of structural constraints in the real-time interpretation of reflexives.

This situation is complicated by the fact that reflexives in non-English languages are not necessarily constrained by the same principles that govern English reflexives (e.g., Kuno, 1972; Sells, 1987; Iida and Sells, 1988; Jayaseelan, 1999; Sohng, 2004, for crosslinguistic examples in Japanese, Malayalam, Korean, etc.). In particular, the phenomenon of long-distance reflexivization, where reflexives are bound by antecedents outside the local domain, has attracted considerable attention and poses challenges for the traditional definition of Binding Principle A[1]. Long-distance reflexives exist in many languages, including Chinese, Japanese, Korean and Icelandic. For example, the Chinese reflexive *ziji* can be long-distance bound in configurations such as (2a):

(2a) John_1 juede Bill_2 bu xihuan $\text{ziji}_{1/2}$.
John$_1$ thought Bill$_2$ NEG like SELF$_{1/2}$
'John$_1$ thought that Bill$_2$ did not like him$_1$/himself$_2$' [*ziji* = ambiguous]

Here, the reflexive *ziji* can refer to either the local subject Bill or the long-distance matrix subject John[2]. It has often been noted that, cross-linguistically, long-distance reflexives are subject to various language-specific constraints, such as the kinds of clause types that allow long-distance binding (e.g., infinitivals, subjunctive clauses, indicative clauses), the animacy of the antecedent, the type of verb in the matrix clause, and the person features of the referents in the sentences (see Huang, 2000, for an overview).

In this paper, we present three self-paced reading experiments on the processing of the Chinese long-distance reflexive *ziji* "self," in order to enrich our understanding of the real-time processing of reflexives from a cross-linguistic perspective. Looking at *ziji* allows us to see what happens in a language where the accessibility of potential antecedents is governed by referents' person features: More specifically, while *ziji* could potentially refer to any subject-position referent (local or non-local), *the person feature* of intervening referents plays a key role in determining the accessibility of a long-distance referent.

Long-Distance Reflexives in Chinese: Blocking Effects

Here, we take a closer look at the Blocking Effects that have been claimed to guide the interpretation of long-distance reflexives in Chinese. Let us consider (2b). If the local subject is the first person pronoun *wo* "I" or the second person pronoun *ni* "you," the widespread claim in the theoretical literature is that *ziji* is bound by this local subject (*wo*/*ni* "I/you") and "blocked" from reaching the matrix/non-local subject ("John" in 2b) (see Xu, 1993; Pan, 1997, 2001; Huang and Liu, 2001). In contrast, if the local subject is a third person referent ("Bill" in 2a), *ziji* can refer to either the local subject or the matrix subject. Hence, there is an *asymmetric* Blocking Effect: An intervening first or second person pronoun blocks long-distance binding whereas a third person referent does not.

(2b) John_1 juede wo_2/ni_2 bu xihuan ziji_2.
John thought I/you NEG like SELF
'John thought that I/you did not like myself/yourself'

Various theoretical analyses have been proposed for this Blocking Effect. One widely used syntactic strategy is to argue that apparent long-distance binding effects can be derived from local dependencies. For example, Tang (1989) (see also Cole et al., 1990; Cole and Sung, 1994; Cole and Wang, 1996, etc.) analyzed long-distance binding as involving a series of movements at the level of logical form such that each movement satisfies the requirement of local binding. Under this view, long-distance

[1]It is important to note that alternative approaches to Chomsky's Binding principles have also been proposed. For example, some researchers argue for predicate-based theories, e.g., Reinhart and Reuland (1993).

[2]Example (2a), like our experiments, uses a proper name in the subject position of the embedded clause. This is because a pronoun in that position would allow for coreference between the embedded subject and the matrix subject (e.g., $John_i$ thought that he_i did not like SELF). This would make it impossible to tell whether people interpret *ziji* as referring to the embedded subject or the matrix subject.

binding in Chinese underlyingly satisfies Chomsky's Binding Principle A.

In addition to syntactic accounts, semantic accounts have been proposed. The most prominent of these attributes blocking to a perspectival conflict, and is based on the "direct discourse complementation" analysis of Kuno (1972). According to Kuno, when a third-person pronoun in an embedded clause refers to the matrix subject who is the speaker/thinker of the embedded clause, the embedded clause can underlyingly be a direct speech event so that the third person pronoun in the surface form is directly derived from an underlying first person pronoun. A sentence like "John$_i$ said he$_i$ hated pancakes" is derived from the underlying form "John said, 'I hate pancakes'."

Building on this, Huang et al. (1984) argued that when *ziji* is in an indirect/reported speech event inside an embedded clause and used as a long-distance reflexive, the embedded clause is derived from a direct speech event. Consider (2a). When *ziji* refers to the matrix subject *John*, the underlying form of the sentence is represented of (2a′) where *ziji* is replaced with the first person *wo* in the direct quote. Here, the first person pronoun *wo* refers to the matrix subject *John*. Hence, a long-distance co-referential interpretation is established. On the other hand, if *ziji* refers to the local subject *Bill*, (2a) is derived from the underlying form in (2a″).

(2a′) John$_1$ juede Bill$_2$ bu xihuan wo$_1$.
John$_1$ thought Bill$_2$ NEG like I$_1$
'John$_1$ thought: "Bill$_2$ doesn't like me." '

(2a″) John$_1$ juede Bill$_2$ bu xihuan ziji$_2$.
John$_1$ thought Bill$_2$ NEG like SELF$_2$
'John$_1$ thought: "Bill$_2$ doesn't like himself$_2$." '

Huang et al. (1984) argued that this approach explains the Blocking Effect. Consider (2b). Based on Huang et al., if *ziji* in (2b) is long-distance bound, the sentence is represented as (2b′), with the original *ziji* in (2b) being represented as the second occurrence of *wo* "I" in (2b′). This second instance of *wo* is intended to refer to the matrix subject *John*, but such a co-referential relationship is not allowed because it results in a conflict in perspectives: The two occurrences of *wo* in (2b′) refer to different referents: the first one refers to the external speaker of the sentence, and the second one to the matrix subject John. Hence, there is a conflict in perspectives. According to Huang et al. (1984), this perspectival conflict is the cause of blocking. In contrast, if *ziji* is locally bound, the underlying form is as in (2b″), and there is no perspectival conflict.

(2b′) ∗ John$_1$ juede wo$_2$ bu xihuan wo$_2$.
John thought I NEG like I
*'John thought: "I(=external speaker) don't like me(=John)".'

(2b″) John$_1$ juede wo$_2$ bu xihuan ziji$_2$.
John thought I NEG like SELF
'John thought: "I do not like myself."'

What about second person blocking? Huang et al. (1984) capture this in a similar way. In (2b), if the embedded subject is *ni* "you" and if *ziji* refers to the matrix subject (*ziji* = John), the underlying direct speech representation of (2b) would be (2c′) where *ziji* is replaced by first person *wo*. Note, however, that inside the direct speech, the second person pronoun *ni* refers to an addressee from the perspective of the *speaker of the entire sentence*, and not from the perspective of the matrix subject John—although, inside the direct speech, the first person pronoun *wo* is anchored to the matrix subject John (me=John). Hence, within the direct quote, we have two different perspectives, which cause a perspectival conflict. According to Huang et al., this again blocks long-distance binding.

(2c′) * John$_1$ juede ni$_2$ bu xihuan wo$_1$.
John thought you NEG like I
*'John thought: "you(=addressee) don't like me(=John)".'

(2c″) John$_1$ juede wo$_2$ bu xihuan ziji$_2$.
John thought I NEG like SELF
'John thought: "I do not like myself".'

Building upon Huang et al. (1984), Huang and Liu (2001) analyzed *ziji* by using a combination of structural and semantic principles. They argued that (i) locally bound *ziji* is a true reflexive and is governed by Binding Principle A and that (ii) long-distance bound *ziji* is a logophor and not subject to Binding Theory.

In sum, these kinds of approaches attribute the Blocking Effects of intervening first and second person pronouns to a perspective conflict that stems from the embedded clause being underlyingly represented as direct speech. Both first and second person pronouns cause a perspective clash when realized in the embedded subject position—unlike third person referents—and thus first and second person referents trigger a Blocking Effect[3].

However, this characterization of blocking is not universally agreed upon. Native speakers' judgments vary regarding the ability of intervening *third* person referents to block long-distance binding. For example, Tang (1989) and Pollard and Xue (1998) treat blocking as a symmetric process whereby a *difference* in person feature between a local referent and a long-distance referent suffices to induce blocking (regardless of the person feature of the intervening referent). Based on this view, the matrix subject *wo* ("I") in (3) cannot antecede the reflexive *ziji*, even though the intervening referent is third person. In other words, some claim that Blocking Effects arise not just with first and second person pronouns, but instead occur in any context where the person features of the local and the long-distance referent are different.

(3) wo$_1$ juede Bill$_2$ bu xihuan ziji$_{?1/2}$.
I thought Bill NEG like SELF
'I thought that Bill did not like SELF.'

In light of the divergent native speaker judgments, it would seem that a psycholinguistic approach could help clarify the situation.

[3]It is also worth noting that a situation where first and second person behave differently from third person may also be related to their different positions on the extended animacy hierarchy ((Dixon, 1979), see also (Croft, 2003)), which ranks first and second above third person.

However, while there exists a large body of experimental work on English reflexives, experimental research on *ziji* has only recently become more frequent. The work that has been done has led to mixed results. For example, Dillon et al. (2009) showed that non-c-commanding subjects did not cause immediate interference in the real-time processing of *ziji*. Based on this finding, Dillon et al. argued that comprehenders only search for structurally compatible referents—c-commanding subject-position referents. Hence, structural constraints play an immediate role. In contrast, Chen and Vasishth (2011) and Jäger et al. (2015b), using self-paced reading and eye-tracking respectively, found that intervening feature-compatible (animate) non-c-commanding subjects caused reliable interference. (For recent work on interference effects and a cue-based retrieval mechanism in German and Swedish, see Jäger et al., 2015a). Based on these findings, they argued that consideration of potential antecedent candidates also relies on featural cues (e.g., animacy). In related work that also points to featural effects, Schumacher et al. (2011) conducted an ERP experiment which found that self-directed verbs exhibit different ERP responses with first- and third-person interveners than with second person interveners.

Aims of the Present Work

The three self-paced reading experiments presented here aim to broaden our understanding of *ziji* by looking at whether and how person-feature-based blocking guides comprehenders' real-time processing and final interpretation of *ziji*.

We have three main aims: First, we want to test to what extent first person and second person interveners block access to long-distance subjects. Even before we turn to the debate regarding third person interveners, it is worth emphasizing that although it is often claimed that long-distance antecedents are not possible in the presence of an intervening first or second person pronoun, judgments seem to actually be rather murky. For example, in our experience, explicitly eliciting judgments from native speakers yields a mixed set of responses. Indeed, when we probed this in an off-line pilot study with 30 Mandarin speakers, we found that people would accept the supposedly impossible long-distance antecedent for *ziji* in a first-person blocking condition [like (2b)] 36.2% of the time. This seems like a rather high number for an interpretation that is supposed to be unavailable/ungrammatical. In order to be able to make progress on this issue, we feel that an experimental investigation of large groups of native speakers is an important step.

Second, given the debate on whether blocking is asymmetric, the present experiments are intended to test whether intervening third person referents block long-distance antecedents like their first and second person counterparts. Thus, in addition to our observations and off-line data which suggest that blocking by first and second person interveners may not be as absolute as some claim, there also exists a fundamental debate—both theoretical and empirical—about what exactly can act as a blocker.

Lastly, while previous experimental work on *ziji* focused primarily on the effect of structural constraints on non-c-commanding subjects, the current experiments examine the real-time effect of a different kind of constraint, namely person-feature cues. (Related work by Schumacher et al., 2011 on person features is discussed in more detail below.) We look at whether in real-time, person-feature cues can immediately reduce interference from blocked/inaccessible long-distance c-commanding subjects.

The three experiments presented in this paper investigate both the on-line processing and the final off-line interpretation of the reflexive *ziji* in the presence of potential first person, second person and third person interveners. Experiment 1 focuses on first person and third person interveners, whereas Experiment 2 tests second person and third person interveners. Furthermore, by changing the type of verb used in the matrix clause in Experiment 3, we test what happens when it is no longer possible to interpret the embedded clause as a direct speech act produced by the matrix subject.

EXPERIMENT 1: FIRST PERSON BLOCKING

Experiment 1 is a self-paced reading study that investigates the effects of intervening first-person pronouns on the real-time processing and off-line interpretation of the reflexive *ziji*. Specifically, we look at whether the presence of an intervening first person pronoun can fully block access to the long-distance matrix subject, as predicted by Blocking. If an intervening first person pronoun acts as an absolute blocker, the reflexive *ziji* should not trigger any consideration of the matrix subject—i.e., we should not see any sign of interference from the matrix subject, either in participants' on-line reading times or off-line interpretations. In addition, given the debate about the (a)symmetry of blocking, we also test whether a *difference* in person feature between a local referent and a long-distance referent suffices to induce blocking. In other words, can an intervening third person referent block access to a matrix subject with a different person feature, such as a first person subject? If yes, this would be evidence in favor of symmetric analyses of blocking and against asymmetric analyses.

Methods

Participants

Twenty adult native speakers of Mainland Mandarin Chinese (graduate students at the University of Southern California at time of testing) took part in Experiment 1 in exchange for USD 10. All of them had normal or corrected-to-normal vision and reported no known learning disabilities or hearing impairments. All studies reported in this paper were reviewed and approved by the University of Southern California University Park Institutional Review Board, which is fully accredited by the Association for the Accreditation of Human Research Protection Programs (AAHRPP). Due to the nature of the experiments, the Institutional Review Board determined that written consent was not needed.

Materials

We used a 2 × 2 design by manipulating the form of the matrix subject and the embedded subject (first person pronoun vs. third person pronoun). Sample sentences of the four conditions are in (4).

(4) Sample sentences for the four conditions
1st-1st 我告诉别人我觉得自己明年可以考进好大学。

wo gaosu bieren **wo** juede ***ziji*** mingnian keyi kaojin
I tell others **I** think **SELF** next year able get-in
hao daxue.
good college.
"I told others that I thought SELF could get into a good college next year."

1st-3rd 我告诉别人李四觉得自己明年可以考进好大学。

wo gaosu bieren **Lisi** juede ***ziji*** mingnian keyi kaojin
I tell others **Lisi** think **SELF** next year able get-in
hao daxue.
good college.
"I told others that Lisi thought SELF could get into a good college next year."

3rd-1st 张三告诉别人我觉得自己明年可以考进好大学。

Zhangsan gaosu bieren **wo** juede ***ziji*** mingnian keyi
Zhangsan tell others **I** think **SELF** next year able
kaojin hao daxue.
get-in good college.
"Zhangsan told others that I thought SELF could get into a good college next year."

3rd-3rd 张三告诉别人李四觉得自己明年可以考进好大学。

Zhangsan gaosu bieren **Lisi** juede ***ziji*** mingnian keyi
Zhangsan tell others **Lisi** think **SELF** next year able
kaojin hao daxue.
get-in good college.
"Zhangsan told others that Lisi thought SELF could get into a good college next year."

We created 32 target items, all of which contained 11 words[4]. (See Supplementary Materials for a full list of targets used in the experiments reported in this paper). The first and the fourth words were the matrix subject and the embedded subject, separated by a verb (*gaosu* "tell") and an object (*bieren* "others") both of which remained the same across all target items. The embedded subject was followed by a verb and then the reflexive *ziji*. In this study as well as the other studies reported in this paper, the verbs (and other lexical items) used in the embedded clauses were designed to be semantically neutral, i.e., to allow *ziji* to be interpreted as referring to either the matrix subject (e.g., Zhangsan) or the embedded subject (e.g., Lisi). [For work on the effects of self- vs. other-directed verbs on *ziji*, see Schumacher et al. (2011), He (2014) and others.]

Following the critical word *ziji* were five words (spillover region). This spillover region is important, because it is well known that in self-paced reading studies, effects many not be detectable until one, two or even three words after the critical word (e.g., Badecker and Straub, 2002, and many others). Our target items used *ziji* in the subject position of an embedded clause, because this allowed us to have a spillover region without a clause boundary *inside* the spillover region. (Clause and sentence boundaries are known to result in "wrap-up" slowdowns, e.g., Warren et al., 2009, which could potentially mask other effects. Indeed, we find signs of wrap-up slowdowns on the last word in our items, but this final word is not relevant for our analyses).

We employed a Latin Square design, resulting in four lists. Each participant saw 32 targets (8 per condition) and 72 fillers, described below. Each target item appeared once in each list but in a different condition in each list. (All experiments reported here used a Latin Square design.)

In addition to the 32 targets, 72 filler items were created. None of the fillers contained *ziji*. In this experiment, as in the other two experiments reported in this paper, the filler items were similar in length to the targets, and also contained multiple clauses (e.g., "Little An suggested that I go to a very renowned seafood restaurant by the seaside" and "Little Zhang heard from others that Little Liu's brother made Little Xiao very depressed").

All targets and fillers were followed by a forced-choice question with two possible answer choices shown on the screen. Target questions probed participants' interpretations of *ziji*, as shown in (5). Because antecedent choice questions could not be used in the 1st-1st condition, we included a referent unmentioned in the sentence as one of the two antecedent choices (6). The forced-choice questions after fillers asked about referents mentioned in the filler items (e.g., "Who recommended a seafood restaurant?" (Little An/I), "Who was very depressed?" Little Xiao / Little Zhang). Positions (left vs. right) of the answer choices for the forced-choice questions were counterbalanced.

(5) Sample comprehension question:
Sentence: Zhangsan gaosu bieren **Lisi** juede ***ziji*** mingnian keyi kaojin hao daxue.
'Zhangsan told others that Lisi thought SELF could get into a good college next year.'

Comprehension question: Shui mingnian keyi kaojin hao daxue?
'Who can get into a good college next year?'
(A) Zhangsan (B) Lisi

(6) Sample comprehension question for the 1st–1st condition:
Sentence: wo gaosu bieren **wo** juede ***ziji*** mingnian keyi kaojin hao daxue.
'I told others that I thought SELF could get into a good college next year.'

Comprehension question: Shui mingnian keyi kaojin hao daxue?
'Who can get into a good college next year?'
(A) Wangwu (B) I

Procedure

We used a moving-window word-by-word self-paced reading paradigm (Just et al., 1982; see also Badecker and Straub, 2002). Participants were tested individually on a laptop computer, using the *Linger* software (D. Rohde, MIT; Rohde, 2010). They first read the instructions and then proceeded to the practice items. The experimental trials started after the practice items. Participants

[4]The 11 words are as shown in the pinyin transliteration and the English word-by-word glosses in (4). Some of the words consist of more than one character in Chinese.

read sentences word-by-word by pressing the spacebar. When a sentence was finished, a comprehension question with two answer choices appeared at the center of the screen. Participants responded by pressing the F key for the answer on the left side or the H key for the answer on the right side.

Predictions

Antecedent Choices

If the blocking effect of the first person pronoun is absolute, long-distance antecedents should be available in the 3rd-3rd Condition but crucially not in the 3rd-1st Condition due to the first person intervener. We should also keep in mind that researchers disagree about whether intervening third person referents can induce blocking: While some argue that only first and second person interveners lead to blocking, others claim that any person-feature mismatch between long-distance and local referents leads to blocking. Hence, antecedent choices data from the 1st-3rd Condition allow us to obtain a clearer picture of the status of third person interveners.

Reading Times

Reading time slowdowns are taken to indicate competition or interference (e.g., Badecker and Straub, 2002). We follow Badecker and Straub (2002) in assuming that if a reflexive has two "candidate antecedents," then additional processing is required to select a unique antecedent, and this increase in processing load is reflected in slower reading times. In other words, competition/interference results in a reading time slowdown, relative to a situation where only one antecedent is being considered. Thus, the 1st-1st Condition should be read rapidly as it has only one referent, the first person pronoun. The 3rd-3rd Condition, on the other hand, should exhibit slowdowns at the reflexive and/or beyond, due to the third person matrix subject competing with the third person embedded subject.

What about the 3rd-1st Condition? If the first person intervener immediately excludes the long-distance matrix subject from the set of possible antecedents, the matrix subject should not cause interference, and this condition should not be read more slowly than the 1st-1st Condition. Alternatively, if the first person pronoun is not an absolute blocker, interference reflected in reading time slowdowns should arise. Predictions for the 1st-3rd Condition are similar to the 3rd-1st Condition. If this condition exhibits blocking as some have argued, no interference should be expected from the first person matrix subject. Otherwise, this condition should also exhibit interference from the matrix subject as reflected in significant reading time slowdowns.

Data Analysis

We used participants' accuracy on the unambiguous filler comprehension questions to check whether they were attending to the task. Since all participants correctly answered at least 90% of the questions, all participants' data were included in subsequent trimming and analyses.

Reading times smaller than 100 ms or above 4000 ms were excluded first. Then, data points were log-transformed to reduce the non-normality of residuals. Afterwards, reading times more than 2.5 standard deviations away from the mean by word and by condition were removed, resulting in the exclusion of approximately 2.7% of data points. Statistical analyses were conducted in R (Baayen et al., 2008; R Core Team, 2015, see also Baayen, 2008). Data for each of the first 10 word positions in the target items were analyzed using linear-mixed effects regression implemented in the R package *lme4* (Bates et al., 2014). Main effects and interaction effects were computed with the R package *car* (Fox and Weisberg, 2011).

Unless otherwise mentioned, logistic mixed-effects regression implemented in *lme4* was used to analyze antecedent choices data due to their binary nature (see Jaeger, 2008, for discussion). In the analyses of antecedent choices, we excluded the data from the 1st-1st Condition because the sentences in this condition only contained the first person pronoun *wo* "I," and one of the two options for the comprehension questions in this condition was a referent unmentioned in the sentence [see (6)]. Participants chose the unmentioned referent on 1.25% of trials in this condition, presumably by mistake.

To specify the random effects in each mixed-effects model, we started with fully crossed and fully specified random effects, testing whether the model could converge. If the model did not converge, we then reduced the random effects until the model reached convergence (see Jaeger at http://hlplab.wordpress.com). We then used likelihood ratio tests to test each random effect and removed those that did not contribute significantly to the model.

Results

Antecedent Choices

As **Figure 1** shows, there was an overall preference for the local (embedded) subject in all conditions (1st-3rd: 95.92%; 3rd-1st: 73.12%; 3rd-3rd: 85.67%). This locality bias is expected based on earlier work (e.g., Chen et al., 2012; Dillon et al., 2014; Jäger et al., 2015b). Furthermore, we see a striking pattern in the 3rd-1st Condition: Although blocking predicts the 3rd-1st Condition to have the lowest rate of non-local choices, in this condition participants opted for the matrix subject and violated blocking on 26.88% of trials. The 3rd-3rd Condition, which—prior research agrees—permits non-local choices, actually had fewer non-local choices than the 3rd-1st Condition. Lastly, the 1st-3rd Condition numerically exhibited the fewest non-local antecedent choices (4.08%).

Antecedent choices were compared using logistic mixed-effects regressions. Participants chose the matrix subject significantly more in the 3rd-3rd and the 3rd-1st Conditions than in the 1st-3rd Condition (**Table 1**). Although the 3rd-1st Condition numerically produced more matrix subject choices than the 3rd-3rd Condition, this difference was not significant (**Table 1**). The higher-than-expected rate of matrix subject choices in the 3rd-1st Condition goes against the prediction of blocking. The low rate of matrix subject choices in the 1st-3rd Condition goes against the claim that third person interveners do not cause blocking.

Lastly, we conducted Bonferroni-corrected one-sample t-tests to check whether the number of non-local, matrix subject choices in each condition was significantly above zero. (Here and elsewhere, we multiplied the *p*-values by the number of comparisons, instead of dividing the alpha level by the number of

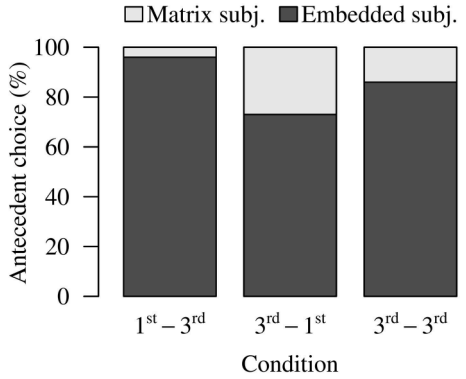


FIGURE 1 | Experiment 1 antecedent choice data.

TABLE 1 | Experiment 1: Comparing the numbers of non-local antecedent choices in the 1st-3rd, the 3rd-1st, and the 3rd-3rd Conditions ("*": $p < 0.05$; ".": $p < 0.1$).

Contrast	β	z	Pr(>|z|)
3rd-1st vs. 1st-3rd	3.2087	2.732	<0.01*
3rd-3rd vs. 1st-3rd	3.1997	3.160	<0.005*
3rd-1st vs. 3rd-3rd	0.0090	0.016	1.0000

comparisons. These two options are mathematically equivalent.) The results showed that the amounts of non-local choices were significantly above zero in the 3rd-1st Condition [$t_{1(19)} = 3.849$, $p < 0.010$; $t_{2(31)} = 6.294$, $p < 0.0001$] and the 3rd-3rd Condition [$t_{1(19)} = 5.511$, $p < 0.001$; $t_{2(31)} = 4.776$, $p < 0.0001$]. For the 1st-3rd Condition, only the by-item test reached significance [$t_{1(19)} = 1.926$, $p = 0.104$; $t_{2(31)} = 2.239$, $p = 0.4870$]. Hence, the 3rd-1st Condition and the 3rd-3rd Condition and to a lesser extent the 1st-3rd Condition allow some amount of non-local choices.

Reading Times

Reading time patterns are shown in **Figure 2**, and results of omnibus tests in **Table 2**. In the five word positions prior to the critical word *ziji*, significant effects of MATRIX SUBJECT were observed, suggesting that conditions with third person matrix subjects were read more slowly than those with first person matrix subjects. At the embedded subject and the following verb, significant effects of EMBEDDED SUBJECT emerged, indicating that conditions with third person embedded subjects were read more slowly. Existing work suggests that third person names are generally read more slowly than first and second person pronouns (Warren and Gibson, 2002), so these patterns are expected but are not central to the aims of this experiment, namely the processing of *ziji*.

Starting from *ziji* (Word 6) and onward, a significant effect of MATRIX SUBJECT was observed at Word 7 ("next year"), but this effect was qualified by a MATRIX SUBJECT × EMBEDDED SUBJECT interaction. At *ziji* and several spillover words that followed, significant interaction effects were observed. To assess these interactions more closely, we compared the three two-referent conditions with the single-referent 1st-1st Condition. At word 7, all three double-referent conditions show significant slowdowns compared to the single-referent (1st-1st) condition (see **Table 3**). In fact, the 3rd-1st condition shows slowdowns relative to 1st-1st on *ziji* (word 6), word 7 as well as word 10. The 1st-3rd condition shows slowdowns relative to 1st-1st on *ziji* (word 6), word 7, and word 9.

Discussion

One of the goals of Experiment 1 was to look at whether the intervening first person pronoun constrains comprehenders' off-line interpretations of *ziji*. The **antecedent choices** in this experiment showed a higher-than-expected rate of matrix subject choices, indicating that blocking is not an absolute principle and that comprehenders sometimes do interpret *ziji* as referring to long-distance antecedents, even (or especially) in the presence of first person interveners. The current experiment also aimed to examine comprehenders' judgments of the 1st-3rd configuration. Researchers diverge regarding whether third person interveners can block access to person-feature mismatching long-distance referents (e.g., the 1st person matrix subject). Our data indicate the intervening third person referent in the 1st-3rd condition can "block" access to the long-distance subject in comprehenders' off-line judgments, in sense that we find less than 5% matrix-subject choices. This finding suggests that it is not accurate to analyze first person interveners as being "better" blockers than third person interveners, and supports previous research that treated third person referents as possible blockers as well (Tang, 1989; Pollard and Xue, 1998).

For the **reading time data**, the single-referent 1st-1st Condition and the double-referent 3rd-3rd Condition patterned as expected. The former was read fast, and the latter, in comparison, was read more slowly due to the two competing antecedents. This finding confirms the prediction that multiple accessible referents can cause competition, reflected in reading time slowdowns.

For the 3rd-1st Condition, if the intervening first person pronoun immediately constrains participants' consideration of antecedent candidates, the matrix subject should not be accessible and thus should not cause an interference (slowdown) effect. However, as we have seen, this condition gave rise to reading time slowdowns at *ziji* and in the spillover region, suggesting that the first person pronoun does not block the accessibility of the matrix subject in real-time and that the presence of two competing referents leads to an interference effect. This finding goes against theoretical claims which regard blocking as a categorical principle. However, it is in line with our off-line data which show that participants violated blocking on an unexpected high rate of trials in this condition.

Additionally, the 1st-3rd Condition, which produced the *fewest* non-local choices and hence exhibited a more stable blocking effect in the offline data, also showed reading time slowdowns. These slowdowns indicate that the "blocked" inaccessible matrix subject in sentences with this kind of configuration can still interfere with the local subject. The results from this condition are in line with existing work (e.g., Kaiser et al., 2009) that suggests comprehenders' off-line judgments do not always coincide with their real-time processing pattern. In

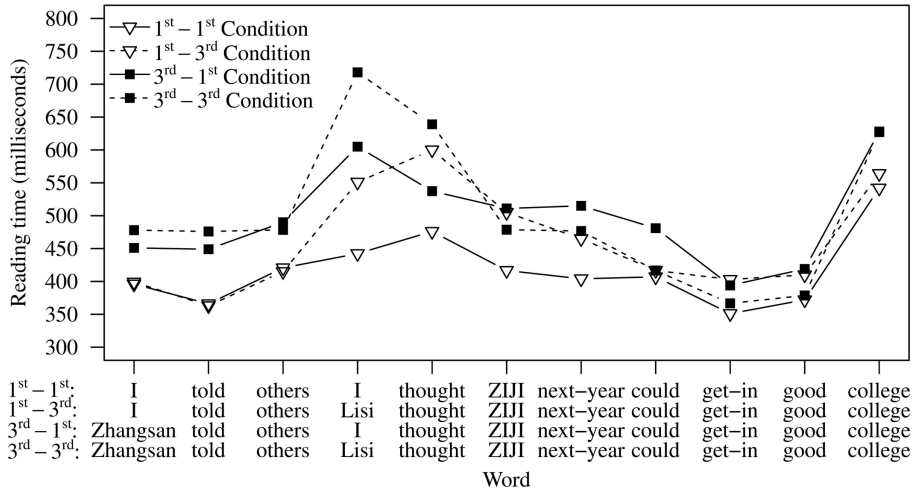


FIGURE 2 | Average reading times per word in Experiment 1.

TABLE 2 | Experiment 1: Reading time results ("*": $p < 0.05$; ".": $p < 0.1$).

	Words	Main effects						Interaction		
		Matrix subject			Embedded subject			Matrix * Embedded		
		χ^2	*df*	*p*	χ^2	*df*	*p*	χ^2	*df*	*p*
1	I/Zhangsan	14.88	1	<0.001*	0.010	1	0.91	0.138	1	0.71
2	Told	86.53	1	<0.001*	0.000	1	1.00	0.366	1	0.55
3	Others	31.29	1	<0.001*	0.028	1	0.87	0.272	1	0.60
4	I/Lisi	36.34	1	<0.001*	18.40	1	<0.001*	1.928	1	0.17
5	thought	6.84	1	<0.010*	43.16	1	<0.001*	0.004	1	0.95
6	**ZIJI**	0.56	1	0.450	1.571	1	0.210	7.379	1	<0.01*
7	next year	6.37	1	<0.050*	3.178	1	0.075·	6.416	1	<0.05*
8	Could	0.053	1	0.820	0.014	1	0.906	2.836	1	0.09·
9	get in	0.065	1	0.799	1.152	1	0.283	7.656	1	<0.01*
10	Good	0.801	1	0.371	0.625	1	0.803	7.282	1	<0.01*

TABLE 3 | Experiment 1: Planned comparisons ("*": $p < 0.05$; ".": $p < 0.1$).

Words		Contrasts	Results			
			β	*SE*	*t*	Pr(>|t|)
6	ZIJI	1st-3rd vs. 1st-1st	0.1067	0.0381	2.804	0.0142*
		3rd-1st vs. 1st-1st	0.0936	0.0382	2.451	0.0384*
7	Next year	1st-3rd vs. 1st-1st	0.1305	0.0428	3.052	0.0064*
		3rd-1st vs. 1st-1st	0.1535	0.4293	3.576	0.0011*
		3rd-3rd vs. 1st-1st	0.1308	0.0428	3.057	0.0061*
9	Get in	1st-3rd vs. 1st-1st	0.0872	0.0321	2.714	0.0184*
10	Good	3rd-1st vs. 1st-1st	0.0721	0.0284	2.539	0.0303*

our case, even though the off-line judgments suggest a stable blocking effect, in real-time, comprehenders can still briefly consider those "blocked" referents.

EXPERIMENT 2: SECOND PERSON PRONOUNS

In Experiment 1, we found a higher-than-expected rate of matrix-subject choices in the first person blocking condition, suggesting that first person blocking is not absolute and that comprehenders sometimes interpret *ziji* as referring the long-distance, matrix subject despite the presence of intervening first person pronouns. Additionally, the results also suggest that in terms of comprehenders' off-line judgments, the intervening third person subject can also serve as a blocker if the long-distance subject has a different person feature, such as first person in Experiment 1. To further examine the interpretation of *ziji* and the blocking effect, Experiment 2 looked at second person blocking. Based on existing theoretical work, we do not expect the first person pronoun and the second person pronoun to differ in their effectiveness as blockers. However, ERP work on Chinese by Schumacher et al. (2011) found that blocking configurations with first- vs. second-person pronouns triggered

different brain responses. This brings up the question of whether first- and second-person pronouns could actually differ in their effectiveness as blockers. This idea receives preliminary (but indirect) support from Brunyé et al. (2009) work on English, which suggests that first- and second-person differ in their ability to induce perspective-taking (see also Ditman et al., 2010; Brunyé et al., 2011). This is especially interesting in light of claims by Huang and Liu (2001) and others that the Blocking effect in Chinese results from a perspective-taking process. In sum, there is (i) a need to better understand the strength of the Blocking effect, given the controversial judgments in this domain, and (ii) a need to better understand whether first- and second-person pronouns differ in their Blocking behavior. Answers to these questions can enrich our understanding of how reflexives are processed.

Methods and Data Analysis

Twenty-eight adult native speakers of Mainland Mandarin Chinese (graduate students at the University of Southern California at time of testing) participated in exchange for USD 10. None of them took part in the previous experiment. All had normal or corrected-to-normal vision and reported no known learning disabilities or hearing impairments. The experimental design, materials, and procedure were identical to those used in Experiment 1, except that all the first person pronouns in the experimental items were replaced by second person pronouns (*ni* "you"). Like Experiment 1, this study also used a Latin Square design.

All participants were highly accurate on the comprehension questions for filler items (90% and above); thus, all participants' data were included in subsequent analyses. The trimming criteria were identical to Experiment 1, resulting in the exclusion of approximately 3.3% of data points. The same statistical methods used in the previous experiment were used here.

Results

Antecedent Choices

In line with the pattern observed in Experiment 1, there was an overall preference for local antecedent choices (**Figure 3**). In the 2nd-3rd, the 3rd-2nd, and the 3rd-3rd conditions, participants chose the embedded subject on 96.88%, 90.18%, and 87.05% of the trials, respectively (As in Experiment 1, we excluded the 2nd-2nd condition from this analysis because this condition only contained one referent, the second person pronoun *ni* "you"). We can also see that the 2nd-3rd Condition and the 3rd-3rd Condition in this experiment were numerically comparable to their counterparts in Experiment 1. However, long-distance choices were relatively rare in the 3rd-2nd Condition (9.82%), compared to the relatively high rate of long-distance choices in the 3rd-1st condition in Experiment 1 (26.88%). We conducted logistic mixed-effects regression to compare these three conditions. The results (**Table 4**) showed that the 3rd-3rd Condition produced significantly more long-distance choices than the 2nd-3rd Condition and the 3rd-2nd Condition. The 2nd-3rd Condition and the 3rd-2nd Condition did not differ significantly from each other.

Bonferroni-corrected by-subject and by-item one-sample t-tests were used to check whether the average number of non-local choices in each condition was significantly above zero. The results show that the number of non-local choices was significantly above zero in all three conditions [2nd-3rd: $t_{1(26)} = 3.017$, $p < 0.010$; $t_{2(31)} = 3.950$, $p < 0.0010$; 3rd-2nd: $t_{1(26)} = 4.837$, $p < 0.001$; $t_{2(31)} = 6.428$, $p < 0.0001$; 3rd-3rd: $t_{1(26)} = 4.416$, $p < 0.001$; $t_{2(31)} = 2.234$, $p = 0.0492$]. Hence, all three conditions allow non-local interpretations of *ziji* to a certain extent.

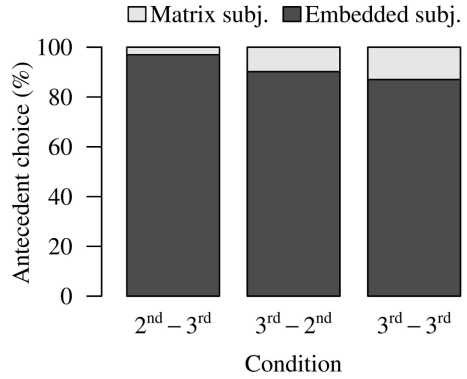


FIGURE 3 | Experiment 2 antecedent choice data.

TABLE 4 | Experiment 2: Comparing the numbers of non-local antecedent choices in the 2nd-3rd, the 3rd-2nd, and the 3rd-3rd Conditions ("*": $p < 0.05$; ".": $p < 0.1$).

Contrast	β	Z	Pr(>|z|)
3rd-2nd vs. 2nd-3rd	0.1675	0.198	0.8433
3rd-3rd vs. 2nd-3rd	1.4281	3.199	<0.0100*
3rd-3rd vs. 3rd-2nd	1.8285	2.788	<0.0100*

Comparing Antecedent Choices in Experiments 1 and 2

In Experiment 1, with first person interveners, participants chose long-distance interpretations of *ziji* in the presence of first person blocking (3rd-1st Condition) on a considerable subset of trials (26.88%). In contrast, in Experiment 2, the 3rd-2nd Condition showed a numerically lower rate of matrix-subject choices (9.82%). Logistic mixed-effects regression was used to directly compare the antecedent choices in Experiment 1 (first person intervener) and Experiment 2 (second-person intervener). We found that the 3rd-1st Condition in Experiment 1 had significantly more matrix subject choices than the 3rd-2nd Condition in Experiment 2 ($\beta = 2.0946$, $z = -2.444$, $p < 0.05$). This difference suggests that relative to the first person pronoun, the second person pronoun constrains comprehenders' final interpretations of *ziji* more consistently. No significant differences were observed between the 1st-3rd Condition and the 2nd-3rd Condition ($\beta = 0.3152$, $z = 0.572$, $p = 1.000$) or between the 3rd-3rd Condition from Experiment 2 and the 3rd-3rd Condition from Experiment 2 ($\beta = 0.1637$, $z = 0.509$, $p = 1.000$).

Reading Times

The reading time patterns for Experiment 2 are shown in **Figure 4**, and the results obtained from omnibus statistical

tests are presented in **Table 5**. The five words preceding *ziji* show a pattern similar to Experiment 1. Significant main effects of MATRIX SUBJECT and EMBEDDED SUBJECT were observed, indicating that third person matrix and embedded subjects elicited longer reading times than their second person counterparts (**Table 6**). A significant interaction was also found at the embedded subject. A closer look at this interaction effect revealed that the 2nd-3rd Condition and the 3rd-3rd Condition were significantly slower than the single-referent 2nd-2nd Condition (2nd-3rd: $\beta = 0.116$, $z = 2.301$, $p < 0.05$; 3rd-3rd: $\beta = 0.377$, $z = 6.90$, $p < 0.001$) and that the slowdown in the 3rd-2nd Condition was marginally significant ($\beta = 0.0703$, $z = 1.716$, $p = 0.087$). As we already mentioned when discussing Experiment 1, which shows a very similar pattern at this point, these results are in line with existing work showing that third person names are generally read more slowly than reduced nouns such as first and second person pronouns (Warren and Gibson, 2002).

At the reflexive *ziji*, a significant main effect of EMBEDDED SUBJECT emerged, suggesting that the two conditions with third person embedded subjects (1st-3rd and 3rd-3rd Conditions) were read more slowly. The word immediately following *ziji* showed a significant interaction effect. Planned comparisons showed that the 2nd-3rd Condition and 3rd-2nd Condition were marginally slower than the 2nd-2nd Condition (2nd-3rd: $\beta = 0.0691$, $z = 2.080$, $p = 0.096$; 3rd-2nd: $\beta = 0.0742$, $z = 2.23$, $p = 0.0672$).

Discussion

Building upon Experiment 1, Experiment 2 aimed to examine whether and how the intervening second person pronoun constrains comprehenders' interpretation of *ziji* both in real-time and off-line. The antecedent choice data provide additional insights into comprehenders' interpretations of the reflexive *ziji*. The results show that the second person blocking condition (3rd-2nd) exhibited a low rate of blocking violations. Direct

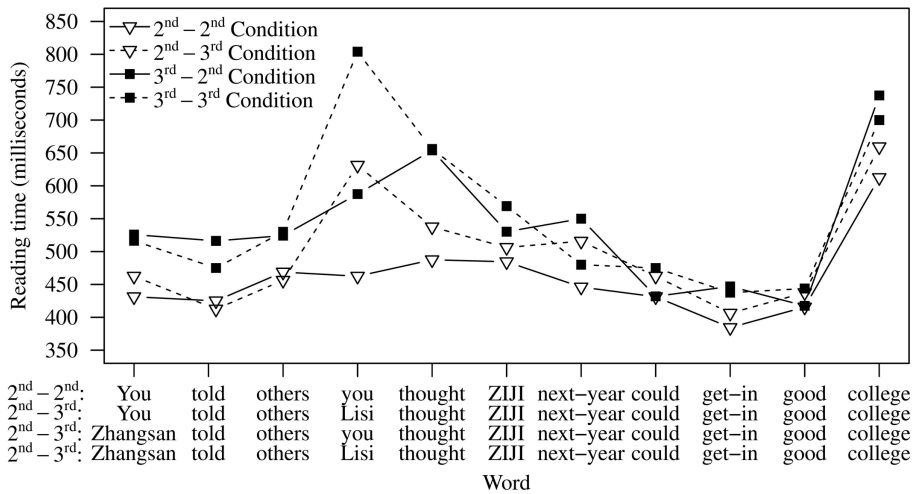


FIGURE 4 | Experiment 2 average reading time data.

TABLE 5 | Experiment 2: Reading time results ("*": $p < 0.05$; ".": $p < 0.1$).

	Words	Main effects						Interaction		
		Matrix subject			Embedded subject			Matrix * Embedded		
		χ^2	*df*	*p*	χ^2	*df*	*p*	χ^2	*df*	*p*
1	You/Zhangsan	19.38	1	<0.001*	1.122	1	0.270	0.496	1	0.481
2	Told	70.76	1	<0.001*	0.245	1	0.621	0.860	1	0.354
3	Others	20.52	1	<0.001*	0.004	1	0.951	0.626	1	0.429
4	you/Lisi	26.18	1	<0.001*	24.74	1	<0.001*	10.43	1	<0.001*
5	thought	20.72	1	<0.010*	10.82	1	<0.010*	0.054	1	0.817
6	**ZIJI**	0.938	1	0.333	8.932	1	<0.010*	0.635	1	0.425
7	next year	0.660	1	0.417	0.363	1	0.547	5.483	1	<0.050*
8	Could	0.023	1	0.374	0.791	1	0.374	0.100	1	0.752
9	get in	1.351	1	0.574	0.316	1	0.574	1.957	1	0.162
10	Good	2.722	1	0.099.	0.851	1	0.447	0.851	1	0.356

TABLE 6 | Target stimuli from Experiments 1 and 2.

1	2	3	4	5	6	7–11
Matrix subject	**Verb**	**Object**	**Embedded subject**	**Verb**	**Reflexive**	**Spillover**
wo/Zhangsan "I/Zhangsan" / ni/Zhangsan "you/Zhangsan"	gaosu "told"	bieren "others"	I/Lisi / you/Lisi	juede "thought"	ziji	mingnian keyi kaojin hao daxue "next-year could get-in good college"

Exp 1: "{I/Zhangsan} told others that {I/Lisi} thought ZIJI could get into a good college next year."
Exp 2: "{You/Zhangsan} told others that {you/Lisi} thought ZIJI could get into a good college next year."

comparisons of antecedent choice patterns between Experiments 1 and 2 confirm that the second person pronoun is indeed a more consistently effective blocker than the first person pronoun. The low rate of long-distance interpretations in the 2nd-3rd Condition, on the other hand, provides additional support for the claim that intervening third person referents can also cause blocking if the long-distance referent has a different person feature.

Using the reading time data, we aimed to examine whether the intervening second person can immediately constrain participants' consideration of potential antecedent candidates in real-time. If the effect of the second person intervener is similar to that of the first person intervener observed in Experiment 1, then an interference effect from the matrix subject should arise in the 3rd-2nd Condition. The results in Experiment 2 showed that the 3rd-2nd Condition was marginally slower than the single-referent 2nd-2nd Condition at the word immediately following *ziji*, but not at any of the subsequent spillover words. This contrasts with Experiment 1, which found significant slowdowns with first person interveners. This suggests that the second person can immediately determine what antecedent candidates get considered, i.e., that the matrix subject can be immediately excluded from consideration.

As a whole, the results from Experiment 2 show hints of the second person pronoun being a stronger blocker than the first person pronoun, given (i) the significantly *fewer* long-distance choices in the 3rd-2nd Condition compared to the 3rd-1st Condition in Experiment 1 and (ii) the absence of competition (i.e., absence of reading-time slowdowns) in the 3rd-2nd Condition.

In related ERP work, Schumacher et al. (2011) found differences between first person and second person interveners with self-directed verbs in constructions like "Wangwu asked me/you to examine myself/yourself." They found a more pronounced early positivity with self-directed verbs with second person interveners than first or third person interveners. Self-directed verbs like "examine" tend to have objects that corefer with their subject/agents (Xi examined Xi). Schumacher et al. (2011) also note that sentences with the second person pronoun report a directive/imperative speech act whereas sentences with the first person pronoun report an assertive speech act. They suggest that, due to the imperative interpretation, the second person is higher on the person hierarchy than the first or the third person. Schumacher et al. (2011) also tested other-directed verbs and found no clear differences between first and second person interveners. Their results for self-directed verbs constitute the first published discussion of differences between first and second person interveners (to the best of our knowledge). However, our stimuli are different in a number of ways. Our target sentences used largely neutral verbs in the embedded clause (to allow *ziji* to refer to either the local or the matrix subject), and the matrix sentence used the verb "told others" (e.g., *Zhangsan told others that I/you/Lisi thought SELF could get into a good college next year.*) Thus, the addressee of "told" in our sentences is "the others," and as a result an imperative interpretation is not possible, in contrast to Schumacher et al. (2011), who derive the different behavior of first and second person interveners from a hierarchy ranking related to the directive/imperative vs. assertive distinction.

As we will see below in Experiment 3, the apparent difference in the blocking strength of first and second person pronouns may in fact be a side effect having to do with participants' (mis)representing the embedded clauses in the test sentences as direct/quoted speech, rather than an intrinsic difference in the blocking behavior of these two forms.

EXPERIMENT 3

Experiments 1 and 2 looked at the effects of first and second person blocking on the real-time processing and off-line interpretation of *ziji*. We saw that the first person pronoun did not seem to show a persistent blocking effect either off-line or in real-time. The second person pronoun, however, exhibited a more reliable blocking effect, significantly reducing interference from the matrix subject. This stronger blocking effect of second person interveners is not predicted by the majority of the existing literature on *ziji*—but see Schumacher et al.'s (2011)—and seems to point to a systematic difference in the blocking strength of the two pronouns.

However, let us take a careful look at the stimuli in Experiments 1 and 2, to see if there could be another reason for the asymmetry we observed. Target items had the sentence

structure shown in **Table 6**: In both experiments, the main-clause verb (Word 2: *gaosu* "tell") was a speech verb. Thus, the embedded clause (Words 4–11) following *gaosu bieren* "tell others" was indirect/reported speech.

For example, in (7), the embedded clause *wo juede ziji...* "I thought SELF..." is a reported speech event: Here, the person who uttered this sentence (the speaker) was reporting what Zhangsan said about the speaker's thoughts. Thus, the embedded subject *wo* "I" refers to the speaker of the entire sentence and not the matrix subject *Zhangsan*.

(7) Zhangsan gaosu bieren [wo juede *ziji* mingnian keyi
Zhangsan tell others [I thought SELF next-year can
kaojin hao daxue].
get-in good college]
"*Zhangsan told others that* [$I_{speaker}$ *thought SELF could get into a good college next year*]."

(8) Zhangsan gaosu bieren: "wo juede *ziji* mingnian keyi
Zhangsan tell others: "I thought SELF next-year can
kaojin hao daxue."
get-in good college"
Zhangsan told others: "$I_{Zhangsan}$ *thought SELF could get into a good college next year."*

However, we suggest that encountering a sentence like (7) may also activate, in people's minds, something similar to (8), which is direct/quoted speech. Crucially, if the embedded clause is direct/quoted speech spoken by Zhangsan, then *wo* "I" refers to Zhangsan and not the speaker of the sentence. (This idea differs from the earlier "transformation-based" approach of Kuno (1972) and Huang et al. (1984). We suggest that a sentence like (7) is, in some sense, ambiguous between reported speech and direct speech—or at least ambiguous enough that it at least partially activates a direct speech representation in participants' minds.) Let us now consider why we think that an example like (7) might partially activate a direct/quoted speech representation like (8).

First, Chinese lacks (overt) complementizers and hence a clause following a speech verb is (in terms of its lexical items) ambiguous between direct/quoted speech and indirect/reported speech. This is unlike English: Compare "*John said (that) I am tired*" with "*John said, 'I am tired'*." English does not use complementizers before direct speech, but optionally uses them before indirect speech. This probabilistic cue is entirely missing in Chinese. This ambiguity in Chinese may result in a sentence like (7) activating a direct speech representation (perhaps in parallel with an indirect speech representation or perhaps stochastically).

Second, the word-by-word self-paced reading paradigm may create the impression of potential "pauses" between words, which may make direct speech interpretations more likely. Given that the start of a direct/quoted speech event in spoken speech is typically characterized by a longer pause (Klewitz and Couper-Kuhlen, 1999), it could be that the boundaries between words created by the self-paced reading method led participants to activate a direct speech representation of the embedded clause. For example, it could be that the break between *bieren* "others" and *wo* "I" in (7) led participants to mentally represent (7) as (8) on some trials. (Like English, Chinese normally uses quotation marks to denote directed/quoted speech, but such cues—or the absence thereof—may be less salient in self-paced reading than normal reading which allows preview and regressions.)

In sum, we suggest that in Experiments 1 and 2, participants may have been partially activating direct speech representations, alongside indirect/reported speech representations. If participants are activating direct speech representations in addition to reported speech, this would lead precisely to the asymmetry between first and second person interveners that we found (i.e., first person pronouns seemingly acting as weaker blockers than second person pronouns):

In Experiment 1, with first person pronouns, under a direct speech representation, on blocking trials (3rd-1st Condition), the first person pronoun *wo* refers to the matrix subject [e.g., *wo* "I" = Zhangsan, as shown in (8)]. Then, if the reflexive *ziji* refers to *wo*, it also refers to the matrix subject [e.g., *Zhangsan* in (8)]. This might explain the apparent violations of blocking that occur on almost 30% of trials in the 3rd-1st condition of Experiment 1: The reflexive *ziji* only *seems* to skip the local subject in favor of the matrix subject: Actually, under a direct speech interpretation, *ziji* is bound by/refers to the local subject and thus also refers to the matrix subject, since local subject is coreferential with the matrix subject. So, according to this line of reasoning, the apparent long-distance interpretation is an *illusion* made possible by a direct speech interpretation, and *ziji* underlyingly refers to the local/embedded subject. Thus, if participants are partially activating direct speech representations alongside reported speech representations, on some proportion of the trials, the activation of the direct speech representation will presumably be sufficiently high to result in selection of the matrix subject.

In Experiment 2, with second person pronouns, the situation is different. In the blocking condition (3rd-2nd), whether or not comprehenders represent the embedded clause as direct speech, the second person pronoun *ni* "you" cannot refer to the matrix subject and can only refer to the addressee in both cases [see (9) and its direct speech counterpart (10)]. Thus, regardless of whether the embedded clause is interpreted as direct or indirect speech, the reflexive *ziji* in sentences with the second person pronoun *ni* "you" cannot use the same means to get to the matrix subject as in sentences with "I." Therefore, the "escape hatch" that is possible with first person embedded subjects in direct speech is not possible with second person embedded subjects in direct speech.

(9) Zhangsan gaosu bieren [ni juede *ziji* mingnian keyi
Zhangsan tell others [you thought SELF next-year can
kaojin hao daxue].
get-in good college]
"Zhangsan told others that [$you_{addressee}$ *thought SELF could get into a good college next year]."*

(10) Zhangsan gaosu bieren: "ni juede *ziji* mingnian keyi
Zhangsan tell others: "ni thought SELF next-year can
kaojin hao daxue."
get-in good college"
"Zhangsan told others: "You$_{addressee}$ thought SELF could get into a good college next year."

This difference between first and second person pronouns fits with our results in Experiments 1 and 2, i.e., the finding that first person pronouns apparently fail to fully block reference to the matrix subject whereas second person pronouns are significantly stronger blockers. In sum, then, this line of reasoning explains the seemingly weaker blocking ability of first person pronouns as "illusory." Under the direct-speech idea, the apparent weakness of first person pronouns as blockers stems from the fact that, with a speech verb in the matrix clause, first person pronouns can be coreferential with the matrix subject whereas second person pronouns cannot.

Experiment 3 aimed to investigate the validity of this idea. Instead of using speech verbs such as *gaosu* "to tell," we used the serial verb structure *ting bieren shuo* "hear others say." The use of the perception verb *ting* "hear" should eliminate the possibility that the embedded clause can be represented as the quoted direct speech of the matrix subject. Thus, Experiment 3 will allow us to see whether (i) the first person pronoun really is a weaker blocker than the second person pronoun, or whether (ii) the weakness of the first person pronoun as a blocker is actually due to direct speech representations.

Methods

Participants

Forty-two adult native speakers of Mainland Mandarin Chinese from the Hunan Normal University in China participated in this experiment in exchange for 60 RMB (equivalent to 10 USD). All had normal or corrected-to-normal vision and reported no known learning disabilities or hearing impairments.

Design and Stimuli

A 2 by 3 design was used. The first factor was PRONOUN TYPE, with two levels—first person and second person. The second factor was REFERENT COMBINATION, with three levels—pronoun-pronoun (or PRO-PRO) vs. pronoun-name (or PRO-NAME) vs. name-pronoun (NAME-PRO). This created a total of 6 conditions. The target sentence structure and examples of the 6 conditions are in **Table 7**.

A total of 42 target items were created, each with 13 words (**Table 7**). The first and the fifth words were the matrix subject and the embedded subject, respectively. The two subjects were separated by a serial verb—*ting biren shuo* "hear others say"[5]—that was constant across all target items. The reflexive was the eighth word and was in the possessive NP position (e.g., *ziji de chengji* "SELF's grade"). Finally, *ziji* was followed by five spillover words (Words 9–13). (The grammatical role of *ziji* in Experiment 3 is different from Experiments 1 and 2. This was necessitated by the change of verb from "tell others" to "hear others say," because we wanted to ensure that the sentences were felicitous and that *ziji* could, in principle, be interpreted as referring to either the matrix or the embedded subject.)

Crucially, the use of the perception verb *ting* "hear" should eliminate the possibility of interpreting the embedded clause (Words 5-13) as the quoted direct speech of the matrix subject. In (11) for example, the embedded clause *wo keyi...* "I could..." cannot be the quoted speech of the matrix subject *Zhangsan*, and can only be what *Zhangsan* heard. Hence, the first person pronoun *wo* "I" refers to the person who uttered the entire sentence and cannot refer to the matrix subject (unlike Experiment 1).

(11) **Zhangsan** ting bieren shuo [wo keyi ba ziji de
Zhangsan heard others say [**I** could BA **ZIJI** DE
chengji gei bieren kan.
grade let others see].
'Zhangsan heard others say [I could let others see SELF's grade].'

Like Experiments 1 and 2, this study also used a Latin Square design. Experiment 3 had six lists, due to its 2 by 3 design. Each participant saw 42 targets (seven per condition), and 68 fillers. The fillers were similar to those in Experiments 1 and 2 (see Section Materials). Similar to the previous two experiments, all items were followed by a forced-choice question. Left-vs.-right positions of the answer choices were counterbalanced.

Procedure

The experimental procedure in this experiment was identical to those in Experiments 1 and 2.

Predictions

Antecedent Choices

If the intervening first person pronoun indeed has a weaker blocking effect than the second person pronoun, we should observe a relatively high rate of blocking violations—i.e., matrix subject choices—in the conditions with first person interveners (3rd-1st) when compared to the conditions with second person interveners (3rd-2nd). Alternatively, if the weakness of the first person pronoun as a blocker (as in Experiment 1) is an illusion due to the "escape hatch" provided by direct speech representations which are ruled out in Experiment 3, then we expect the rate of blocking violations in conditions with first and second person interveners to now be comparable (a low number of blocking violations in both conditions).

Reading Times

If the first person pronoun has a weaker blocking effect than the second person pronoun, then in conditions with first person interveners, we should see reading time slowdowns from *ziji* and onwards as a result of competitions between the matrix subject and the embedded subject. In particular, the reading times in the first person blocking condition should be slower compared to those in the second person blocking condition, if the first person pronoun is weaker blocker (i.e., allows more competition from the matrix subject) than the second person pronoun. On

[5]In Chinese, there is no construction equivalent to the English structure *hear + embedded clause* (e.g., I heard Peter went to Costa Rica). The closest structures are *hear-say + embedded clause* and *hear-others-say + embedded clause*.

TABLE 7 | Sentence structure for target items and sample target sentences in Experiment 3.

Condition	1	2	3	4	5	6	7	8	9–13
1st-1st	**wo "**I"				**wo "**I"				
1st-3rd	**wo "**I"				**Lisi**				
3rd-1st	**Zhangsan**	ting "hear"	bieren "others"	shuo "say"	**wo "**I"	keyi "can"	BA	**ZIJI** "SELF"	de chengji rang bieren kan "DE grade let others see"
2nd-2nd	**ni "**you"				**ni "**you"				
2nd-3rd	**ni "**you"				**Lisi**				
3rd-2nd	**Zhangsan**				**ni "**you"				

1st-1st:* "I** *heard others say* **I** *could give* **ZIJI's score to others to look at."*
1st-3rd:* "I** *heard others say* **Lisi** *could give* **ZIJI's score to others to look at."*
3rd-1st:* "Zhangsan** *heard others say* **I** *could give* **ZIJI's score to others to look at."*
2nd-2nd:* "You** *heard others say* **you** *could give* **ZIJI's score to others to look at."*
2nd-3rd:* "You** *heard others say* **Lisi** *could give* **ZIJI's score to others to look at."*
3rd-2nd:* "Zhangsan** *heard others say* **you** *could give* **ZIJI's score to others to look at."*

the other hand, if the blocking violations in Experiment 1 (first person interveners) were actually "illusions" due to direct speech, then in Experiment 3, we should not observe competitions between the blocked matrix subject and the local subject.

Data Analysis

All participants were highly accurate on the comprehension questions for filler items (90% and above); thus, all data were included in subsequent analyses. The trimming criteria were identical to those used in Experiments 1 and 2 and resulted in the exclusion of 2.59% of data points. The same statistical methods were used to analyze data in the present experiment. The reading time data for the first 12 words of the target items were analyzed.

Results and Discussion

In this section, we present the results of Experiment 3 and discuss them briefly. We postpone an in-depth discussion of Experiment 3 until the General Discussion, because the full import of this third study is best appreciated when it is compared to the results of Experiments 1 and 2.

Antecedent Choices

As shown in **Figure 5**, matrix subject choices were numerically relatively rare (1st-3rd: 1.70%; 3rd-1st: 6.46%; 2nd-3rd: 4.08%; 3rd-2nd: 13.95%). However, the first person blocking condition (3rd-1st) and the second person blocking condition (3rd-2nd) had somewhat higher percentages of matrix subject choices than the other conditions. Surprisingly, the 3rd-2nd Condition actually had numerically the *highest* rate of matrix subject choices. (As in the preceding studies, we excluded the data from the two single-referent conditions, 2nd-2nd and 1st-1st, from analysis the antecedent-choice analyses.)

Logistic mixed-effects regression was used to test the effects of PRONOUN TYPE and REFERENT COMBINATION on antecedent choices. The results showed significant main effects of pronoun type [$\chi^2 = 13.069$, $df = 1$, $p < 0.001$] and referent combination [$\chi^2 = 21.071$, $df = 1$, $p < 0.001$] but no interaction [$\chi^2 = 0.0349$, $df = 1$, $p = 0.852$]. Hence, conditions with second person pronouns were more likely to elicit matrix subject choices than conditions with first person pronouns, and the two blocking configurations were more likely to produce matrix subject choices than the other two conditions.

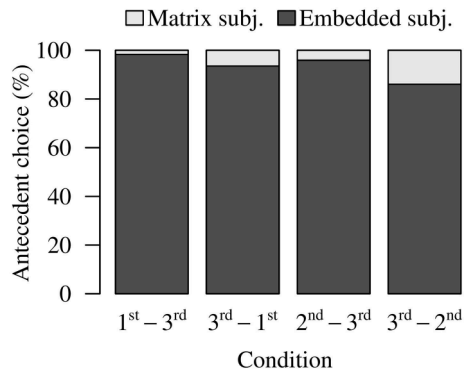


FIGURE 5 | Experiment 3 antecedent choice data.

Although the omnibus test reported above did not yield a significant interaction, a set of four planned comparisons was carried out. The results (**Table 8**) showed that the 3rd-1st Condition elicited more matrix subject choices than the 1st-3rd Condition, and that the 3rd-2nd Condition elicited more matrix subject choices than the 2nd-3rd Condition. In addition, the 3rd-1st Condition actually had *fewer* matrix subject choices than the 3rd-2nd Condition. This result is different from what we saw in Experiments 1 and 2 where the opposite pattern was observed. That is, the 3rd-1st Condition in Experiment 1 led to significantly more matrix subject choices than the 3rd-2nd Condition in Experiment 2. The finding that in Experiment 3 (when direct speech interpretations are blocked), the 3rd-1st Condition resulted in *fewer* matrix subject choices than the 3rd-2nd Condition clearly argues against the idea that the first person is an inherently weaker blocker than the second person.

As with the first two experiments, we used Bonferroni-corrected one-sample *t*-tests to check whether the number of non-local, matrix subject choices in each condition was significantly above zero. The results showed that the numbers of matrix subject choices in the two blocking conditions were significantly above zero, and were marginally above zero in the 1st-3rd condition and significantly above zero in the 2nd-3rd condition [1st-3rd: $t_{1(41)} = 2.354$, $p = 0.094$; $t_{2(41)} = 3.950$,

TABLE 8 | Experiment 3: Planned comparisons for antecedent choice data ("*": $p < 0.05$; ".": $p < 0.1$).

Contrast	β	z	Pr(>|z|)
3rd-1st vs. 1st-3rd	2.0751	2.975	0.0107*
3rd-2st vs. 2nd-3rd	2.4489	4.618	<0.001*
3rd-1st vs. 3rd-2nd	−1.2112	−3.604	0.0012*
1st-3rd vs. 2nd-3rd	−0.8374	−1.051	0.6672

$p = 0.094$; 3rd-1st: $t_{1(41)} = 2.953$, $p = 0.021$; $t_{2(41)} = 6.428$, $p < 0.001$; 2nd-3rd: $t_{1(41)} = 3.106$, $p = 0.014$; $t_{2(41)} = 3.950$, $p = 0.007$; 3rd-2nd: $t_{1(41)} = 3.683$, $p = 0.003$; $t_{2(41)} = 6.428$, $p < 0.001$].

In sum, we find that the 3rd-1st and the 3rd-2nd Conditions, which are often regarded as the prototypical blocking conditions, allow rates of blocking violations (matrix subject choices) that are significantly higher than 0, and in fact higher than the 1st-3rd and 2nd-3rd Conditions respectively. This indicates that blocking is not a strict, categorical phenomenon. Furthermore, we find no evidence that first person pronouns are weaker blockers than second person pronouns. In fact, in this study, the 3rd-2nd Condition results in more matrix subject choices than the 3rd-1st condition. This suggests that the high rate of matrix subject choices in Experiment 1 may indeed have been due to participants activating direct speech representations, which function as an "escape hatch" to allow *ziji* to refer to the matrix subject in the presence of an intervening first-person subject.

Reading Times

The reading time patterns are presented in **Figure 6** (conditions with first person pronouns) and **Figure 7** (conditions with second person pronouns). Linear mixed-effects regression was used to analyze log-transformed reading time data.

At the 7 words prior to *ziji*, a persistent effect of REFERENT COMBINATION was observed (**Table 9**), reflecting an increased processing effort involved in reading third person names (compared to first and second person pronouns). This pattern is similar to what we observed in Experiments 1 and 2 and is also expected based on existing research (Warren and Gibson, 2002). From *ziji* and onwards, marginally significant effects of REFERENT COMBINATION were observed at Words 9 and 10. However, planned comparisons did not yield any significant differences at these positions. A significant REFERENT COMBINATION × PRONOUN TYPE interaction was found at word 12, but planned comparisons did not reveal any significant contrasts at this position.

In sum, we find no clear evidence for reading-time slowdowns (i.e., competition between multiple referents) after *ziji*. This suggests that in this experiment, when direct speech representations are not possible, participants are not considering the matrix subject as a potential referent—or not sufficiently for it to result in a reading-time slowdown—for the reflexive *ziji*. In the General Discussion, we take a closer look at how these results relate to the outcomes of Experiments 1 and 2, and what the implications of these comparisons are.

GENERAL DISCUSSION

The experiments reported here tested whether and how person-feature-based blocking guides comprehenders' real-time processing and final interpretation of the Chinese reflexive *ziji* "self." Our work was motivated by three main aims. First, we wanted to test experimentally to what extent native speakers' judgments fit with the view often presented in theoretical work that first person and second person interveners block access to long-distance subjects. Second, there is debate in existing work concerning the configurations that can result in blocking—in particular, whether blocking only occurs with intervening first and second person pronouns, or whether it can also occur with third person pronouns as long as there exists a mismatch in the featural make-up of the matrix subject and the embedded subject. We tested whether intervening third person referents block long-distance antecedents like their first and second person counterparts. Third, we complement prior on-line work by testing whether person-feature cues can immediately reduce interference from blocked / inaccessible long-distance c-commanding subjects.

Absence of Absolute Blocking Effects, and Potential Asymmetries between First and Second Person Pronouns

Regarding the first and third questions mentioned above, Experiment 1 found that first person interveners in the purported blocking condition (3rd-1st) resulted in a higher-than-expected rate of matrix subject choices, as well as reading-time slowdowns. This suggests that when comprehenders encounter sentences with third person matrix subjects, first person embedded subjects, and a reflexive *ziji* in the embedded clause (3rd-1st), both the embedded and matrix subject compete as potential antecedents for the reflexive. This argues against claims that blocking is categorical, since under that view, we should see no matrix subject choices and no slowdowns. Interestingly, Experiment 2 found that second person interveners in the purported blocking condition (3rd-2nd) exhibited a low rate of matrix subject choices and only short-lived, marginal reading-time slowdowns.

In light of these results, one might be tempted to conclude that second person pronouns are stronger blockers. Such a conclusion might in fact be expected, in light of earlier claims that blocking in Chinese is related to perspective taking (Huang and Liu, 2001). Let us combine this idea with other work on perspective in cognitive psychology which found that in English (at least in some contexts), the second person induces stronger perspective-taking than the first person (Brunyé et al., 2009; Ditman et al., 2010, see also Brunyé et al., 2011). If this stronger perspective-taking effect with the second person pronoun also holds in Chinese and if blocking is indeed related to perspective taking, then we may expect to see a stronger blocking effect with the second person pronoun. However, as we will see below, Experiment 3 shows that this conclusion is too hasty, because the use of a verb of saying as the embedding verb in Experiments 1 and 2 allows first-person pronouns an "escape hatch" that seems

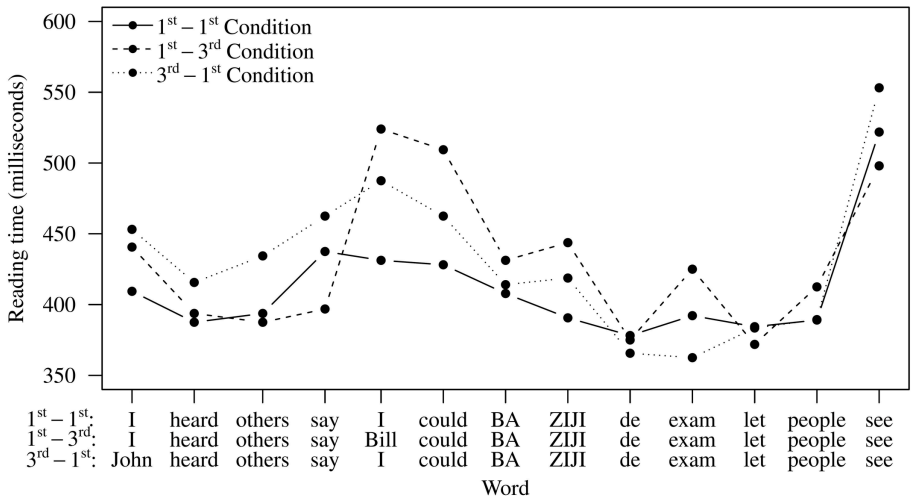


FIGURE 6 | Reading times for conditions with the first person pronoun in Experiment 3.

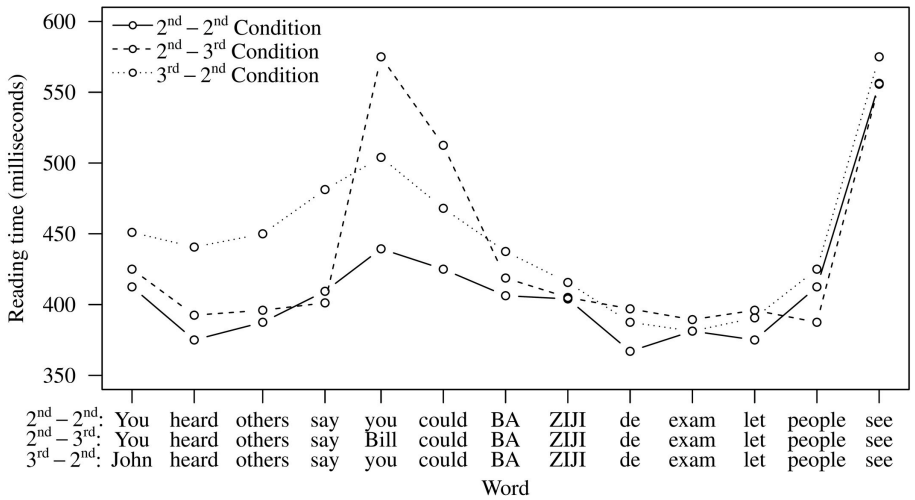


FIGURE 7 | Reading times for conditions with the second person pronoun in Experiment 3.

to be boosting the rate of matrix subject choices without violating blocking.

Evidence for Blocking in Asymmetrical Environments, Even Without First or Second Person Blockers

Regarding the second question above, namely whether blocking (even if it is not absolute) only occurs when the embedded subject (the blocker) is first or second person or whether blocking phenomena can also occur in configurations where the two subjects have different person features (e.g., 1st-3rd or 2nd-3rd, so the blocker is third person), our results argue for the second view. In all three experiments, an intervening third person referent in the 1st-3rd and 2nd-3rd conditions can block access to the long-distance subject in off-line judgments (i.e., we find relatively lower rates of matrix subject interpretations in those conditions than in 3rd-3rd conditions).

However, the reading time patterns in Experiments 1 and 2 are less clear: In Experiment 1, in the 1st-3rd Condition still showed reading time slowdowns (relative to the baseline 1st-1st condition), which could be taken as an indication that the "blocked" inaccessible matrix subject can still interfere with the local subject. In Experiment 2, the 2nd-3rd Condition showed only marginal reading time slowdowns relative to the 2nd-2nd Condition. Thus, even though the off-line judgments suggest a stable blocking effect, in real-time comprehenders may still briefly consider the "blocked" referents. In Experiment 3, the intervening third person referent seems to exclude the first and second person matrix subject from the initial set of

TABLE 9 | Experiment 2: Reading time results ("*": $p < 0.05$; ".": $p < 0.1$).

	Words	Main effect						Interaction		
		Referent combination			Pronoun type					
		χ^2	*df*	*p*	χ^2	*df*	*p*	χ^2	*df*	*p*
1	I/You/Zhangsan	13.71	2	<0.010*	1.136	1	0.287	1.419	2	0.492
2	Heard	81.29	2	<0.001*	1.908	1	0.167	1.844	2	0.398
3	Others	45.42	2	<0.001*	0.367	1	0.545	0.043	2	0.979
4	Say	36.12	2	<0.001*	0.010	1	0.921	2.916	2	0.233
5	I/you/Lisi	30.43	2	<0.001*	3.275	1	0.070.	1.300	2	0.522
6	Could	30.63	2	<0.001*	0.605	1	0.825	0.605	2	0.739
7	BA	16.31	2	<0.001*	0.775	1	0.379	3.772	2	0.152
8	**ZIJI**	3.801	2	0.150	0.000	1	0.995	3.602	2	0.165
9	DE	4.951	2	0.084.	2.951	1	0.086	1.276	2	0.528
10	Exam	5.937	2	0.051.	0.006	1	0.940	1.685	2	0.431
11	Let	0.189	2	0.910	0.761	1	0.383	3.633	2	0.393
12	People	1.248	2	0.546	3.917	1	0.048	8.371	2	0.015*

antecedent candidates, as we find no significant reading time slowdowns. Given that Experiments 1 and 2 allow direct speech interpretations, as discussed above, we assume that the results of Experiment 3 are more reliable in this regard.

As a whole, we interpret these results as supporting theoretical claims that blocking is symmetric and third person interveners can also serve as blockers (Tang, 1989; Pollard and Xue, 1998). However, the finding that in Experiment 3, the third person intervener actually exhibited a stronger blocking effect than first or second person interveners, hints that maybe blocking is not fully symmetric. Perhaps first and second person blocking involves a different mechanism than third person based blocking. We leave this as a question for future research.

Taking a Closer Look at Whether First Person Pronouns are Weaker Blockers

Experiment 3 was designed to test (i) whether the first person pronoun and the second person pronoun differ in their effectiveness as blockers, or (ii) whether the high rate of violations of first person blocking in Experiment 1 could be due to comprehenders treating the embedded clauses as quoted direct speech. As described above, a direct speech interpretation would allow "apparent" blocking violations even when the reflexive is actually bound by the local subject. In Experiment 3, we ruled out potential direct speech interpretations by using a verb of hearing (*heard from others* rather than *told others*).

We did not find any evidence in Experiment 3 that first person pronouns are worse blockers than second person pronouns. Interestingly, although both first and second person blocking conditions triggered matrix-subject choices at above-zero rates, the second person blocking condition actually allowed a *higher* rate of matrix subject choices than the first person blocking condition—contrary to the results of Experiments 1 and 2. Thus, after eliminating the possibility of participants representing the embedded clause as direct speech, the first person pronoun seems to create a more stable configuration than the second person pronoun in determining comprehenders' judgments of *ziji*. This finding was unexpected, and merits further investigation.

As a whole, the antecedent choice data in Experiment 3 provide additional evidence for our conclusion that blocking is not a strict, categorical phenomenon. In fact, the first and second person blocking configurations produced *more* matrix subject choices than the two conditions with third person interveners (1st-3rd and 2nd-3rd). However, the reading time data, particularly those from Experiments 2 and 3, show that comprehenders seem to use person feature cues quickly during real-time processing to filter out inaccessible long-distance referents. For example, in Experiment 3, the 3rd-1st Condition had a reading time pattern comparable to that of the 1st-1st Condition at *ziji* and onwards—in other words, we see no signs of a slowdown in the 3rd-1st condition, suggesting that the matrix subject is not competing as a potential antecedent when direct-speech interpretations are ruled out.

Thus, there seems to be a mismatch between comprehenders' on-line performance and off-line antecedent choices: Participants' final responses suggest that, although the local subject is the preferred antecedent for *ziji*, participants still interpret *ziji* as referring to the non-local subject at rates significantly higher than 0. However, at the same time, reading times suggest that when participants process *ziji,* they do not experience slowdowns/competition effects, i.e., reading time patterns suggest that only one antecedent is being considered at the point where *ziji* is processed. This difference between on-line and off-line patterns points to the possibility that the interpretation of *ziji* unfolds over time: it seems that initially, during real-time processing, person-feature cues weigh more heavily and constrain what antecedent candidates get considered. However, participants' off-line interpretations suggest that at some later point, other kinds of information are also integrated and perhaps outweigh the person-feature constraint, resulting in consideration of referents that were initially "blocked" due to the person-feature constraint.

Implications for Models of Reference Resolution

Our results highlight the role that person features play in guiding the interpretation of reflexives. This contrasts with most existing psycholinguistic models of reflexive processing, which have tended to focus on structural information. For example, Dillon et al. (2009) hypothesize that because *ziji* can be potentially bound by all c-commanding subjects in a discourse, comprehenders should use the c-commanding subjecthood information to search for potential antecedent candidates. Our experiments shed new light on the types of information that guide the interpretation of *ziji*. The finding that comprehenders quickly use person feature cues to guide the search for potential antecedents in real-time suggests that structural information is not the only type of constraint that regulates the real-time processing of *ziji*. In addition, the results from our experiments also suggest that the real-time interpretation of *ziji* can be subtly influenced by comprehenders' mental representations of written texts (i.e., direct vs. indirect speech representations of embedded clauses). These findings are in line with work by Patil et al. (2011), Chen and Vasishth (2011), and Jäger et al. (2015b), who showed that non-structural information also affects the real-time processing of referential forms. The results from Experiments 1–3 also lend support to studies such as Kaiser et al. (2009) that show that comprehenders' antecedent choices do not necessarily follow structural constraints strictly.

AUTHOR CONTRIBUTIONS

XH and EK conceptualized and designed the experiments. XH collected the data and conducted the statistical analyses. Both XH and EK interpreted the data and wrote the manuscript.

ACKNOWLEDGMENTS

We would like to thank the audiences at the 35th Penn Linguistics Colloquium, the 4th Conference on Quantitative Investigations in Theoretic Linguistics and the 23rd North American Conference on Chinese Linguistics, where earlier versions of some of this work were presented. Preliminary analyses of some of the data presented here appear in the University of Pennsylvania Working Papers in Linguistics Vol. 18, and in the dissertation of the first author (He, 2014, University of Southern California). Thanks are also due to the Frontiers reviewers, as well as to Maria Luisa Zubizarreta, Roumyana Pancheva, and Rand Wilcox for useful comments and feedback on many aspects of this work.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpsyg.2016.00284

REFERENCES

Baayen, R. H. (2008). *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R*. Cambridge: Cambridge University Press.

Baayen, R. H., Davidson, D. J., and Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *J. Mem. Lang.* 59, 390–412. doi: 10.1016/j.jml.2007.12.005

Badecker, W., and Straub, K. (2002). The processing role of structural constraints on the interpretation of pronouns and anaphors. *J. Exp. Psychol. Learn. Mem. Cogn.* 28, 748–769. doi: 10.1037/0278-7393.28.4.748

Bates, D., Maechler, M., and Bolker, B. (2014). *lme4: Linear Mixed-Effects Models Using Eigen and S4*. Available online at: http://CRAN.R-project.org/package=lme4

Brunyé, T. T., Ditman, T., Mahoney, C. R., Augustyn, J. S., and Taylor, H. A. (2009). When you and I share perspectives: pronouns modulate perspective taking during narrative. *Psychol. Sci.* 20, 27–32. doi: 10.1111/j.1467-9280.2008.02249.x

Brunyé, T. T., Ditman, T., Mahoney, C. R., and Taylor, H. A. (2011). Better you than I: perspectives and emotion simulation during narrative comprehension. *J. Cogn. Psychol.* 23, 659–666. doi: 10.1080/20445911.2011.559160

Chen, Z., Jäger, L., and Vasishth, S. (2012). "How structure-sensitive is the parser? Evidence from Mandarin Chinese," in *Empirical Approaches to Linguistic Theory: Studies of Meaning and Structure (Studies in Generative Grammar)*, eds B. Stolterfoht and S. Featherston (Berlin: Mouton de Gruyter), 43–62.

Chomsky, N. (1986). *Knowledge of Language: Its Nature, Origin, and Use*. New York, NY: Praeger.

Chen, Z., and Vasishth, S. (2011). "Does the parser exclusively use structure-sensitive search in reflexives? Evidence from Mandarin Chinese," in *Talk Presented at the 23nd Annual Meeting of the CUNY Conference on Human Sentence Processing* (Stanford, CA).

Chomsky, N. (1981). *Lectures on Government and Binding*. Dordrecht: Foris.

Clackson, K., and Heyer, V. (2014). Reflexive anaphor resolution in spoken language comprehension: structural constraints and beyond. *Front. Psychol.* 5:904. doi: 10.3389/fpsyg.2014.00904

Cole, P., Hermon, G., and Sung, L. (1990). Principles and parameters of long-distance reflexives. *Linguist. Inq.* 21, 1–22.

Cole, P., and Sung, L. M. (1994). Head movement and long-distance reflexives. *Linguist. Inq.* 25, 355–406.

Cole, P., and Wang, C. (1996). Antecedents and blockers of long-distance reflexives: the case of Chinese *ziji*. *Linguist. Inq.* 27, 357–390.

Croft, W. (2003). *Typology and Universals*. Cambridge: Cambridge University Press.

Dillon, B., Chow, W.-Y., Wagers, M., Guo, T., Liu, F., and Phillips, C. (2014). The structure-sensitivity of memory access: evidence from Mandarin Chinese. *Front. Psychol.* 5:1025. doi: 10.3389/fpsyg.2014.01025

Dillon, B., Mishler, A., Sloggett, S., and Phillips, C. (2013). Contrasting intrusion profiles for agreement and anaphora: experimental and modeling evidence. *J. Mem. Lang.* 69, 85–103. doi: 10.1016/j.jml.2013.04.003

Dillon, B., Xiang, M., Chow, W.-Y., and Phillips, C. (2009). "The processing of long-distance and local anaphora in Mandarin Chinese," in *Poster Given at the 22nd Annual CUNY Human Sentence Processing Conference 2009* (Davis, CA).

Ditman, T., Brunyé, T. T., Mahoney, C. R., and Taylor, H. A. (2010). Simulating an enactment effect: pronouns guide action simulation during narrative comprehension. *Cognition* 115, 172–178. doi: 10.1016/j.cognition.2009.10.014

Dixon, R. M. W. (1979). Ergativity. *Language* 55, 59–38. doi: 10.2307/412519

Fox, J., and Weisberg, S. (2011). *An R Companion to Applied Regression, 2nd Edn*. Thousand Oaks CA: Sage. Available online at: http://socserv.socsci.mcmaster.ca/jfox/Books/Companion

He, X. (2014). *What 'You' and 'I' Can Say about Reference Resolution and Non-structural Constraints*. Unpublished Ph.D. dissertation, University of Southern California.

Huang, C.-T. J., Huang, Y.-H., Teng, T.-H., and Tiedeman, R. (1984). "Reflexives in Chinese and the teaching of Chinese," in *Proceedings of the First World Conference on Chinese Language* (Taipei), 205–215.

Huang, C.-T. J., and Liu, C.-S. L. (2001). "Logophoricity, attitudes and ziji at the interface," in *Syntax and Semantics*, Vol. 33, eds P. Cole, and C.-T. J. Gabriella Hermon (Bingley: Emerald), 141–195.

Huang, Y. (2000). *Anaphora: A Cross-Linguistic Study*. Oxford; New York, NY: Oxford University Press.

Iida, M., and Sells, P. (1988). "Discourse factors in the binding of zibun," in *Papers from the Second International Workshop on Japanese Syntax*, ed W. J. Poser. (Stanford, CA), 23–46.

Jaeger, T. F. (2008). Categorical data analysis: away from ANOVAs (transformation or not) and towards logit mixed models. *J. Mem. Lang.* 59, 434–446. doi: 10.1016/j.jml.2007.11.007

Jäger, L. A., Benz, L., Roeser, J., Dillon, B. W., and Vasishth, S. (2015a). Teasing apart retrieval and encoding interference in the processing of anaphors. *Front. Psychol.* 6:506. doi: 10.3389/fpsyg.2015.00506

Jäger, L. A., Engelmann, F., and Vasishth, S. (2015b). Retrieval interference in reflexive processing: experimental evidence from Mandarin, and computational modeling. *Front. Psychol.* 6:617. doi: 10.3389/fpsyg.2015.00617

Jayaseelan, K. A. (1999). "Lexical anaphora in Malayalam," in *Lexical Pronouns and Anaphors in Some South Asian Languages: A Principled Typology*, eds B. Lust, K. Wali, J. Gair, and K. V. Subbarao (Berlin: Mouton de Gruyter), 113–168.

Just, M. A., Carpenter, P. A., and Wooley, J. D. (1982). Paradigms and processes in reading comprehension. *J. Exp. Psychol. Gen.* 111, 228–238. doi: 10.1037/0096-3445.111.2.228

Kaiser, E., Runner, J. T., Sussman, R. S., and Tanenbaus, M. K. (2009). Structural and semantic constraints on the resolution of pronouns and reflexives. *Cognition* 112, 55–80. doi: 10.1016/j.cognition.2009.03.010

Klewitz, G., and Couper-Kuhlen, E. (1999). Quote - unquote? The role of prosody in the contextualization of reported speech sequences. *Pragmatics* 9, 459–485. doi: 10.1075/prag.9.4.03kle

Kuno, S. (1972). Pronominalization, reflexivization, and direct discourse. *Linguist. Inq.* 3, 161–195.

Nicol, J., and Swinney, D. (1989). The role of structure in coreference assignment during sentence comprehension. *J. Psycholinguist. Res.* 18, 5–19. doi: 10.1007/BF01069043

Pan, H. (1997). *Constraints on Reflexivization in Mandarin Chinese*. New York, NY: Garland.

Pan, H. (2001). "Why the blocking effect?" in *Syntax and Semantics*, Vol. 33, eds P. Cole, and C.-T. J. Gabriella Hermon (Bingley: Emerald), 279–316.

Patil, U., Vasishth, S., and Lewis, R. (2011). "Early retrieval interference in syntax-guided antecedent search," in *Poster at the 23nd Annual Meeting of the CUNY Conference on Human Sentence Processing* (Stanford, CA).

Pollard, C., and Xue, P. (1998). Chinese reflexive ziji: syntactic reflexive vs. nonsyntactic reflexive. *J. East Asian Ling.* 7, 287–318. doi: 10.1023/A:1008388024532

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available online at: http://www.R-project.org/

Reinhart, T., and Reuland, E. (1993). Reflexivity. *Linguist. Inq.* 24, 657–720.

Rohde, D. (2010). *Linger*. Available online at: http://tedlab.mit.edu/~dr/Linger/

Runner, J. T., Sussman, R. S., and Tanenhaus, M. K. (2006). Assigning reference to reflexives and pronouns in picture noun phrase. *Cogn. Sci.* 30, 1–49. doi: 10.1207/s15516709cog0000_58

Schumacher, P. B., Bisang, W., and Sun, L. (2011). Perspective in the processing of the Chinese reflexive *ziji*: ERP evidence. *Lect. Notes Comput. Sci.* 7009, 119–131. doi: 10.1007/978-3-642-25917-3_11

Sells, P. (1987). Aspects of logophoricity. *Linguist. Inq.* 18, 445–479.

Sohng, H.-K. (2004). A minimalist analysis of X 0 reflexivization in Chinese and Korean. *Stud. Gen. Grammar* 14, 375–396.

Sturt, P. (2003). Time-course of the application of binding constraints in reference resolution. *J. Mem. Lang.* 48, 542–562. doi: 10.1016/S0749-596X(02)00536-3

Tang, C.-C. J. (1989). Chinese reflexives. *Nat. Lang. Linguist. Theory* 7, 93–121.

Warren, T., and Gibson, E. (2002). The influence of referential processing on sentence complexity. *Cognition* 85, 79–112. doi: 10.1016/S0010-0277(02)00087-2

Warren, T., White, S. J., and Reichle, E. D. (2009). Investigating the causes of wrap-up effects: evidence from eye movements and E-Z reader. *Cognition* 111, 132–137. doi: 10.1016/j.cognition.2008.12.011

Xiang, M., Dillon, B., and Phillips, C. (2009). Illusory licensing effects across dependency types: ERP evidence. *J. Brain Lang.* 108, 40–55. doi: 10.1016/j.bandl.2008.10.002

Xu, L. (1993). The long-distance binding of ziji. *J. Chinese Linguist.* 21, 123–141.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2016 He and Kaiser. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 09 October 2015
doi: 10.3389/fpsyg.2015.01504

Wh-filler-gap dependency formation guides reflexive antecedent search

Michael Frazier, Lauren Ackerman, Peter Baumann, David Potter and Masaya Yoshida*

Department of Linguistics, Northwestern University, Evanston, IL, USA

Prior studies on online sentence processing have shown that the parser can resolve non-local dependencies rapidly and accurately. This study investigates the interaction between the processing of two such non-local dependencies: *wh*-filler-gap dependencies (WhFGD) and reflexive-antecedent dependencies. We show that reflexive-antecedent dependency resolution is sensitive to the presence of a WhFGD, and argue that the filler-gap dependency established by WhFGD resolution is selected online as the antecedent of a reflexive dependency. We investigate the processing of constructions like (1), where two NPs might be possible antecedents for the reflexive, namely *which cowgirl* and *Mary*. Even though *Mary* is linearly closer to the reflexive, the only grammatically licit antecedent for the reflexive is the more distant *wh*-NP, *which cowgirl*.

(1). Which cowgirl did Mary expect to have injured herself due to negligence?

Four eye-tracking text-reading experiments were conducted on examples like (1), differing in whether the embedded clause was non-finite (1 and 3) or finite (2 and 4), and in whether the tail of the *wh*-dependency intervened between the reflexive and its closest overt antecedent (1 and 2) or the *wh*-dependency was associated with a position earlier in the sentence (3 and 4). The results of Experiments 1 and 2 indicate the parser accesses the result of WhFGD formation during reflexive antecedent search. The resolution of a wh-dependency alters the representation that reflexive antecedent search operates over, allowing the grammatical but linearly distant antecedent to be accessed rapidly. In the absence of a long-distance WhFGD (Experiments 3 and 4), *wh*-NPs were not found to impact reading times of the reflexive, indicating that the parser's ability to select distant *wh*-NPs as reflexive antecedents crucially involves syntactic structure.

Keywords: reflexive antecedent search, filler-gap dependency resolution, structure-sensitivity, gender mismatch effect, eye-tracking, text-reading

OPEN ACCESS

Edited by:
Matthew Wagers, University of California, Santa Cruz, USA

Reviewed by:
Ian Cunnings, University of Reading, UK
Matthew Alan Tucker, New York University Abu Dhabi, United Arab Emirates

****Correspondence:***
Michael Frazier, Department of Linguistics, Northwestern University, 2016 Sheridan Rd, Evanston, IL 60208, USA
fraze@u.northwestern.edu

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: *03 July 2015*
Accepted: *17 September 2015*
Published: *09 October 2015*

Citation:
Frazier M, Ackerman L, Baumann P, Potter D and Yoshida M (2015) Wh-filler-gap dependency formation guides reflexive antecedent search. Front. Psychol. 6:1504. doi: 10.3389/fpsyg.2015.01504

1. Introduction

In order to interpret sentences of natural language, the human parser must establish non-local dependencies between elements received in the input. Two such kinds of dependencies are wh-dependencies and reflexive-antecedent dependencies. The former is the dependency between a *wh*-word such as *who* or *which* and the empty argument position (e.g., subject, direct object, indirect object) where it is interpreted, which we refer to throughout as a "*Wh*-filler-gap dependency" or "WhFGD." The latter is the dependency between a reflexive pronoun such as *himself* or *herself* and the antecedent noun phrase on which it is referentially dependent, which we refer to as a "reflexive dependency" or "RD."

WhFGDs and RDs differ from one another in a number of ways. While in a WhFGD the presence of a *wh*-word at the left edge of a clause can provide evidence for the existence of an empty NP position later on, such as the empty direct object position in a sentence like *What did Mary eat?*, a RD cannot be recognized until later. This is because in a RD, it is typically the later-occurring element in the dependency that contains bottom-up evidence of the need for a reflexive-antecedent relation. That is, in a sentence like *Mary saw herself*, there is no indication that *Mary* will need to be associated with a reflexive later in the sentence until the reflexive *herself* is actually encountered. Evidence from many psycholinguistic studies, discussed below, indicates that both of these dependency resolution processes occur quite rapidly in online reading. If the presence of a WhFGD affects the online operation of a subsequent RD resolution, this would constitute evidence that the antecedent retrieval process is sensitive to syntactic structure, namely to the presence and location of the WhFGD.

In this paper, we investigate the interaction between the processes of the parser that establish these two kinds of dependencies in on-line sentence comprehension. In particular, we examine whether resolving a WhFGD establishes a new candidate antecedent in the representation that is searched during the resolution of a RD. Converging evidence from the psycholinguistic sentence processing literature indicates that the process of WhFGD resolution is an active process. In particular, upon encountering a *wh*-word, the parser does not wait to receive bottom-up information determining the location of the tail of the WhFGD, but actively posits or hypothesizes the dependency tail whenever it detects an incoming position at which resolving the dependency would be grammatically licit (Stowe, 1986; Traxler and Pickering, 1996; Phillips, 2006). Likewise, reflexive resolution is known to be rapid and grammatically sensitive (Nicol and Swinney, 1989; Sturt, 2003; Jäger et al., 2015): upon encountering the reflexive, the parser tries to link the reflexive to grammatically licit antecedents in the early stages of online processing.

Considering RDs like 1, a reflexive normally co-refers with its closest potential antecedent. In (1), *himself* is understood to co-refer with *the man*, not with *Jane*[1]. In (2), however, the *wh*-phrase *which man* is interpreted as the subject of the non-finite embedded clause *to have injured himself*, just as *the man* is in (1), but it is displaced from the canonical embedded subject position after *expect*. In a context such as this, the *wh*-phrase *which man* must be the antecedent of *himself*, instead of the linearly closer noun phrase *Jane*. If *Jane* were chosen as the antecedent of the reflexive in either (1) or (2), the example would be predicted to be unacceptable due to the gender mismatch between the feminine name *Jane* and the masculine reflexive *himself*, contrary to fact.

(1) Jane$_i$ expected the man$_j$ to have injured himself$_{*i/j}$.

(2) Which man$_i$ did Jane$_j$ expect to have injured himself$_{i/*j}$?

[1]Here and throughout, subscript indices are used to indicate possible and impossible coreference relations: NPs bearing the same subscript index indicate an interpretation under which these NPs refer to the same entity, and impossible interpretations [e.g., *himself* = *Jane* in (1)] are prefixed with an asterisk.

Examples such as these, with nonfinite embedded clauses associated with sentence-initial *wh*-phrases, allow us to investigate whether the result of WhFGD resolution influences RD resolution. Without WhFGD resolution, in an example like (2) the closest potential candidate antecedent (*Jane*) for the reflexive mismatches with it in gender. If reflexive resolution operates over a representation that does not include information about WhFGDs, then in the course of finding the antecedent for the reflexive *himself* in (2) the parser may (at least transiently) attempt to associate *himself* with *Jane*, leading to processing difficulty and a possible slowdown due to the gender mismatch (Sturt, 2003).

If, however, the active process of WhFGD resolution alters the representation over which reflexive resolution operates, it may establish a new candidate antecedent for the reflexive that is closer to the reflexive than the ungrammatical antecedent *Jane*, by co-indexing the sentence-initial *wh*-phrase *which man* with the position of the gap (3)[2]. In this case, the closest candidate antecedent for the reflexive in (2) will be the gap linked to the (masculine) *wh*-phrase. We would thus not expect the parser to attempt to associate *himself* with *Jane* even temporarily, because there is a closer, grammatically acceptable antecedent. The parser should therefore experience no gender-mismatch effect when the reflexive mismatches in gender with an ungrammatical but linearly close candidate antecedent such as *Jane* in (2).

(3) Which man$_i$ did Jane expect /gap/$_i$ to have injured himself?

Thus, whether or not the parser experiences gender-mismatch effects from a linearly close but ungrammatical candidate antecedent like *Jane* in examples like (2) can tell us whether the process of reflexive resolution is sensitive to the presence of a WhFGD.

The plan for this paper is as follows. In the remainder of Section 1, we discuss the theoretical and empirical background of this line of research, focusing in turn on WhFGD resolution (Section 1.1), and reflexive antecedent search (Section 1.2). In Section 2 we report the results of four experiments to test whether the tail of a WhFGD is treated online as a potential antecedent for reflexive resolution. Section 3 discusses the implications of these results for theories of sentence-processing, and Section 4 concludes.

1.1. Active *wh*-dependency Resolution

The term *WhFGD resolution* refers to the process by which the parser interprets a left-peripheral *wh*-question element such as *what*, *who*, or *which NP* to correspond to appropriate sentence-internal material—approximately, to correspond to the position in which the *wh*-element's correlate would appear in an answer to the *wh*-question. The end result of this process is that the *wh*-element in a *wh*-question like (4-a) is interpreted as corresponding to the empty direct object position, such that an answer to (4-a) would include an element filling this position, as in (4-b).

[2]Whether *position* here is defined structurally, or in terms of verbal selection frames, or at the level of predicate-argument relations, is immaterial at this early point in the discussion.

(4) a. What did Mary devour?
b. Mary devoured fish.

Upon encountering a position in the input string in which a grammatically obligatory element is missing, [in the case of (4-a), the object position immediately after *devour*] the parser has bottom-up evidence that this is the position to be affiliated with the left-peripheral *wh*-word. In what follows, we refer to this empty position in the input corresponding to the *wh*-element's answer and to the variable in the sentence's interpretation as the *gap*.

Converging evidence from the psycholinguistic sentence processing literature, however, indicates that the parser is not as conservative in resolving WhFGDs as the above would suggest: instead the process of *wh*-dependency resolution is an *active* process (Frazier, 1987). In particular, upon encountering a *wh*-word, the parser does not wait to receive bottom-up information determining the location of the gap (viz. a position in the input string in which a grammatically obligatory element is missing), but actively posits or hypothesizes the existence of a gap whenever it detects an incoming position at which such a gap would be grammatically licit (Traxler and Pickering, 1996; Phillips, 2006; Omaki et al., 2015).

The principle line of evidence that *wh*-dependency resolution is an active process in this sense comes from the so-called filled-gap effect (FGE, Stowe, 1986 *et seq.*). The FGE is a reading-time slowdown observed at the positions of an overt NP in a sentence with a *wh*-element, such as the position of *the sushi* in (5). The object position after *eat* is a potential gap site, but not an actual gap site, the actual gap site being in the complement of the preposition *with*. The fact that reading-time slowdowns are observed at such positions is interpreted as an indication of the parser's having hypothesized a gap in the position occupied by the overt NP and subsequently, upon finding this prediction falsified, having to take time to correct its mistake[3].

(5) What did Mary eat the sushi with?

Precisely how active or predictive the process of WhFGD resolution may be is not directly relevant to the present study, because in all experiments reported here the WhFGD occurs substantially before the measurement regions of the reflexive and its spillover region. However, the general finding that gap-filling is an active, rapid process is strong evidence that this process will have completed by the time RD resolution is triggered, when the parser encounters the reflexive in examples like (3). This enables us to study whether reflexive resolution is sensitive to the presence of a WhFGD without the danger that the WhFGD has not been recognized by the parser at the point when RD resolution occurs.

1.2. Antecedent Retrieval

Because *wh*-words in English are under normal circumstances located at the left edge of the clause with which they are associated, *wh*-dependency resolution is in the general case a forward process in the sense that the cue to the existence of a long-distance dependency between two linguistic elements is encountered at the leftmost element. Reflexive antecedent search is quite different, because while a reflexive is overtly marked with the morpheme *self*, it generally occurs after its antecedent, which does not bear any marking indicating that it is the antecedent of an upcoming reflexive. Trivially, in example (6) below, *John* occurs in the same form whether it is the antecedent of a reflexive (6-b) or not (6-a). That is, while the presence of *wh*-morphology triggers an active search through subsequently-processed linguistic material for the tail of the *wh*-dependency, the presence of reflexive morphology (English *-self*) must instead trigger a backwards search through previously-processed material for its antecedent.

(6) a. John_i dislikes $\text{him}_{*i/j}$.
b. John_i dislikes $\text{himself}_{i/*j}$.

Additionally, the possibility of a RD is constrained in two ways. The first constraint, typically referred to as Condition A of the Binding Theory (Chomsky, 1981), states that reflexives must be locally bound, so non-clausemate NPs and those that do not c-command[4] the reflexive are illicit antecedents, as indicated by the unacceptability of the examples in (7). Second, in English, a reflexive must match its antecedent in number and gender, so that e.g., masculine NPs are illicit antecedents of feminine reflexives and vice versa, as indicated by the unacceptability of the examples in (8).

(7) a. *John hopes that the police won't find himself.
b. *Rumors about John bothered himself.

(8) a. *John injured herself.
b. *Mary injured himself.

Although the aim of this study is principally to determine how structure-sensitive the reflexive antecedent retrieval process is, rather than to distinguish between different mechanisms of antecedent retrieval, the parser's behavior in this context still has the potential to be informative about the retrieval mechanism itself, and so some discussion of different models of the retrieval of linguistic antecedents from memory bears inclusion here.

In cue-based models of antecedent retrieval like Lewis and Vasishth (2005), the antecedent retrieval mechanism is not crucially constrained by syntactic structure. Instead, upon encountering a word that requires an antecedent (in the present case, the reflexive), the parser performs a feature-matching operation in parallel on all the elements in a content-addressable memory store—roughly, all the words it has recently encountered. In a model like this, cues indexing the syntactic position of potential antecedents can interact with non-structural cues like agreement features, allowing ungrammatical

[3]While (Stowe, 1986) observed FGE-related slowdown only in the position of non-subjects, not of subjects, since then, other researchers, for example Lee (2004), have found that subject-FGEs do appear under slightly different experimental settings.

[4]C-command (Reinhart, 1976) is a notion of relative syntactic prominence; formally, in a tree structure, α c-commands β iff α does not dominate β and the node immediately dominating α also dominates β. For our purposes here, it suffices that subjects c-command their associated verb phrases and all the contents of their associated verb phrases, including subordinate clauses. Note also that additional NPs contained inside an NP subject do not c-command anything out of the NP subject, as in (7-b) below.

antecedents to potentially be retrieved. On such an account, both candidate reflexive antecedents are predicted to impact the reading-time measures of the reflexive in our experiments, because both of them will be simultaneously checked against the features (in particular the gender feature) of the reflexive when the parser encounters it.

Precisely how the candidate antecedents should affect reading-time measures of the reflexive depends upon the details of the cue-based model adopted. A simple possibility is that the parser should experience extra difficulty when no candidate antecedent is found in its memory store, leading to an interaction effect such that the reflexive regions of sentences like (3) but containing no gender-matching antecedent for the reflexive, such as (9), are read most slowly.

(9) Which woman did Jane expect to have injured himself?

More complex patterns are also possible, however. In the model of Lewis and Vasishth (2005), two distinct interference effects are predicted between the match/mismatch of the *wh*-NP and the linearly local NP. First, when both candidate antecedents match the feature specification of the reflexive, similarity-based interference is predicted, such that the reflexive regions of sentences like (3) but containing two gender-matching antecedent for the reflexive, such as (10), will exhibit reading-time slowdowns.

(10) Which man did John expect to have injured himself?

This is due to the mutual inhibition between the featurally similar candidates. Second, facilitation should occur where the accessible antecedent mismatches and the inaccessible antecedent matches the features of the reflexive. This would manifest as faster reading times. More complex models such as the one in Jäger et al. (2015) also predict that the gender congruency of the candidate antecedents should interact in modulating reading-times at the reflexive. In general, cue-based retrieval models that are not constrained by syntactic structure make the prediction that both candidate antecedents should affect reading times.

Cue-based models commonly incorporate a decay parameter, such that items that have been in memory longer are less salient and harder to retrieve, but a decay parameter does not predict effects of the *wh*-NP in the absence of effects of the more local candidate antecedent NP, since the *wh*-NP will have been in memory slightly longer. Even if the *wh*-NP is re-activated (and thus boosted) in memory at the position of the verb in the lower clause as Lewis and Vasishth (2005)'s model predicts, it should still be the case that the activation of the more local candidate antecedent remains strong enough to induce some effect at the reflexive. For this reason, theories of cue-based retrieval would predict an interference effect from the matrix subject NP *Jane* in (3).

Dillon et al. (2013) did not observe an interference effect of this kind in their experiments on sentences like (11), and performed computational simulations of the level of memory activation of the competing reflexive antecedents in order to determine whether such a reactivation-based account could explain the lack of interference effects. They compared the predictions of a cue-based system that was restricted to consider only syntactic information in reflexive antecedent retrieval with one that could consider all cues, including agreement information, where both systems incorporated memory reactivation of the grammatical antecedent [in (11), *the new executive*]) at the matrix verb (*doubted*), a point after the competing antecedent (*the middle managers*).

(11) *The new executive who oversaw the middle managers apparently doubted themselves …

They concluded, however, that a formal model of antecedent activation that was restricted to consider only syntactic cues in reflexive antecedent search provided a closer fit to their empirical findings than one that considered all cues (including morphological ones) and depended only upon relative activation level to modulate which candidate antecedent was retrieved. That is to say, cue-based models that were not restricted to consider only syntactic structural cues in reflexive antecedent retrieval predicted more interference than observed, even after accounting for reactivation of the grammatical antecedent.

While the sentences studied in Dillon et al. (2013) involve somewhat different long-distance dependencies, in both their sentences and ours the grammatical antecedent is reactivated after the ungrammatical candidate antecedent [at *doubted* in their (11) and at or around *have injured* in our (3)], and so a model such as theirs plausibly predicts no effect of the ungrammatical candidate antecedent in our experiments as well.

Many previous studies have investigated whether the antecedent retrieval process, whether cue-based or otherwise, is constrained by syntactic relations: namely, where a potential antecedent is located in the syntactic tree (e.g., Badecker and Straub, 2002; Sturt, 2003; Felser et al., 2009; Xiang et al., 2009; Cunnings and Felser, 2013; Dillon et al., 2013; Clackson and Heyer, 2014; Cunnings and Sturt, 2014).

Sturt (2003) investigated the on-line application of Condition A of the Binding Theory in sentence processing by cross-manipulating the (stereotypical) gender match/mismatch and structural accessibility/inaccessibility (in terms of c-command relations) of prior discourse referents. In his Experiment 1, two candidate antecedents, both c-commanding a reflexive, were cross-varied for gender congruency with the reflexive, as in (12). As expected, when the linearly closer and structurally accessible antecedent mismatched the anaphor in stereotypical gender, reading times on the anaphor/spill-over region were slower.

(12) He/she remembered that the surgeon had pricked himself/herself with a used syringe needle.

On the other hand, in Sturt (2003)'s Experiment 1, a significant effect of inaccessible-match/mismatch was found in later measures, such that mismatching inaccessible antecedents slowed down subsequent reading times on the anaphor. Sturt (2003) interpreted this result as evidence that the antecedent retrieval process is structurally constrained such that grammatical constraints act as a filter on interpretation during on-line reading, which can subsequently be violated by more general comprehension processes. Cunnings et al. (2015) and Kush et al.

(2015) found similar patterns in the case of pronoun binding, and both interpret them as evidence of later comprehension processes attempting to coerce an antecedence relation when none is permitted by the grammar, though the explanation offered by Kush et al. (2015) involves a number of additional complications.

Sturt (2003)'s Experiment 2 is similar in some ways to the present study in that it also tests reflexive resolution in configurations where a grammatically inaccessible antecedent is linearly closer to the reflexive than the grammatically accessible antecedent.

(13) The surgeon who treated Jonathan/Jennifer had pricked himself/herself with a used syringe needle.

The fact that Sturt found no result of the inaccessible antecedent in his Experiment 2, in contrast with his Experiment 1 where late interference effects were found, may indicate that such interference effects are confined to configurations in which the inaccessible antecedent c-commands the reflexive. If this is the case, interference effects similar to those in Sturt's Experiment 1 may be found in the present study. However, in this study, unlike in Sturt's Experiment 2, both the accessible and inaccessible antecedent c-command the reflexive. This difference allows the present study to serve as a kind of follow-up to Sturt (2003), distinguishing whether the parser's reflexive antecedent resolution system is sensitive to structural (rather than linear) locality of a potential antecedent separately from c-command.

Substantial additional evidence indicates that at least the structural relation of c-command affects dependency formation. Cunnings et al. (2015) and Kush et al. (2015), for example, both investigated the retrieval of antecedents of bound pronouns which, like reflexives, are referentially dependent upon other NPs. Both groups of researchers found evidence that the c-command relation constrains the antecedent search process, such that bound pronouns only trigger antecedent retrieval of possible binders in c-commanding positions.

However, theories taking account of only c-command do not predict that the parser should be able to effectively distinguish between the grammatical and ungrammatical antecedents in our experiments, since both candidate antecedents c-command the reflexive. For these accounts to make different predictions about these sentences, a notion of locality is needed as well—reflexives in English are more restricted than bound pronouns because their antecedents must c-command the reflexive *and* must be contained in the same immediate clause. If the retrieval system can take advantage of both of these structural properties (c-command and clausemate-hood), it should fail to exhibit interference effects from the linearly more local candidate antecedent in sentences like (3).

That is to say, the possible grammatical sensitivity of the parser investigated here is somewhat finer grained than that investigated in e.g., Cunnings et al. (2015), who investigated whether antecedent retrieval is sensitive to the c-command constraint on anaphora. Correctly resolving sentences like (3) requires the parser to attend to two grammatical constraints–the clausemate condition on reflexives, and the necessity of a WhFGD tail in the embedded clause in examples like (3)–and not merely retrieve a c-commanding antecedent, since both *which man* and *Jane* c-command the reflexive in sentences like (3). In our case, if the reflexive is linked to the c-commanding linearly local antecedent, a gender mismatch effect is expected based upon the gender match/mismatch of this NP with the reflexive. On the other hand, if the reflexive is linked to the tail of the WhFGD, due to the parser's respecting the structural constraint on the WhFDG, we should observe a gender mismatch effect due to the *wh*-NP's match/mismatch with the gender of the reflexive.

Dillon et al. (2013) directly compared reflexive antecedent retrieval with a somewhat similar dependency, subject-verb agreement, that also requires feature congruency between words that may be linearly distant from one another. They investigated whether the interference effects found in subject-verb agreement, where an illicit potential antecedent can cause the verb to mistakenly bear incorrect agreement morphology, were also found in reflexive antecedent retrieval, and did not find evidence that they were. Dillon et al. (2013) proposed that, unlike in subject-verb agreement, the antecedents of reflexives are retrieved using solely syntactic cues, with other kinds of cues, such as grammatical gender, being checked against retrieved antecedents only later. On their account, the retrieval system has access to information about which NP in its memory store is the local subject, thus enabling it to be sensitive to c-command as well as locality. On an account of this kind, we would not expect to see an effect of the inaccessible antecedent on reading times of the reflexive.

However, there is a caveat to the preceding discussion. Even if the retrieval system is able to track the identity of the current local subject, it is possible for it to be misled by examples like (3). The grammatically accessible antecedent for the reflexive in (3) is the *wh*-phrase, which is not located inside the immediate local clause containing the reflexive, *to have injured himself*. No local subject is overtly present in this clause at all. The sentences in our experiments thus probe one further level of syntax-sensitivity on the part of the antecedent retrieval system: whether it is able to access the result of the WhFGD resolution process, a posited gap in the subject of the infinitive clause, as a potential reflexive antecedent. There are at least two reasons it might fail to do so.

First, it is possible that the results of WhFGD resolution are simply not represented in a way that is accessible to the reflexive resolution process. This might be the case if the gap/tail of the WhFGD was simply not present in its memory store. Second, it might be the case that the parser is susceptible to what are known as *local coherence effects*, where a parse is adopted that is suitable for only a substring of the input. Note that in examples like (3), if the *wh*-NP is disregarded, the result is the possible sentence *did Jane expect to have injured himself*, in which *Jane* must be the antecedent of the reflexive, contrary to gender congruency. There is evidence that in some contexts the parser can be misled by local coherence effects (e.g., Ferreira et al., 2002; Tabor et al., 2004; Konieczny et al., 2010), and if the result of WhFGD formation is not accessible to the reflexive antecedent retrieval system, it may exhibit such effects in this context as well.

Prior research on reflexive antecedent retrieval in configurations similar to WhFGDs in that they involve an NP associated with a subsequent, unpronounced position

similar to a gap has found mixed results. Kwon and Sturt (2015) investigated reflexive antecedent retrieval in the context of nominal control constructions like (14). Control constructions of this kind resemble WhFGDs in that they can involve a dependency between a displaced NP [*Luke* in (14-b)] and a position in an embedded clause [the subject position of *to photograph himself* in (14-b)].

(14) a. Luke's order to Sophia to photograph *himself/herself . . .
b. Luke's promise to Sophia to photograph himself/*herself . . .

Kwon and Sturt (2015) found an effect of antecedent-reflexive gender mismatch both in nominal control constructions like (14-a), where the accessible antecedent (*Sophia*) was closer to the reflexive than the inaccessible candidate, and in nominal control constructions like (14-b), where the accessible antecedent (*Luke*) was more distant, though the former effect was reliable for more reading-time measures. They interpret this result to indicate that the control relation is processed early on and used to constrain subsequent RD formation.

For our purposes here, the fact that an effect was observed in the control constructions most similar to WhFGDs, those like (14-b) where the antecedent is distant from the embedded clause, suggests that the antecedent retrieval system may be sensitive to agreement mismatch in resolving reflexive-antecedent dependencies even when the antecedent is related to the reflexive via the mediation of a long-distance dependency.

Sturt and Kwon (2015) presented additional results on reflexive antecedent retrieval in nominal control as well as the related construction of *raising*, illustrated in (15).

(15) John seemed to Amy to be kind to himself . . .

They found evidence of retrieval interference by grammatically inaccessible antecedents for both raising and nominal control, casting further doubt on the possibility that reflexive antecedent search can find an antecedent online whose relation to the reflexive is mediated by a long-distance dependency. Like Sturt (2003)'s early measures, however, they did not find evidence for interference from grammatically inaccessible antecedents in reflexive-antecedent configurations not involving raising or control.

1.3. Summary

The resolution of a WhFGD is an active process by which the parser posits the tail of a *wh*-dependency upon encountering grammatically licit positions for it in the input. Similarly, the application of binding conditions in reflexive resolution occurs rapidly in on-line reading. Because of this, sentences containing a *wh*-dependency whose tail constitutes the grammatically-licit antecedent for a reflexive pronoun are an ideal environment for a test for the time course of structure-sensitivity in on-line sentence processing. In particular, these kinds of sentences allow us to test whether backward antecedent search processes are sensitive to fine-grained details of the grammatical representation containing the candidate antecedent NPs. Furthermore, if the timing of effects of accessible and inaccessible antecedents differs, they may may be informative about whether grammatical sensitivity constrains the antecedent search process itself or whether the antecedent search process is itself insensitive to fine-grained syntactic details and syntactic knowledge becomes operative only later as a supplementary cue to filter out impossible antecedent-reflexive relations generated by the antecedent search process. The experiments described below constitute such a test.

2. Experiments

2.1. Experiment 1

2.1.1. Introduction

Experiment 1 is the principal experiment of this study and serves to test whether reflexive resolution is sensitive to the result of WhFGD resolution. Experiments 2–4, which are reported in subsequent sections below, serve as follow-up experiments to Experiment 1, intended to clarify the interpretation of Experiment 1's results. Like all of the experiments in this study, Experiment 1 tests sentences in which a *wh*-NP occurs at the left edge of a complex sentence involving a matrix clause and an embedded clause, the latter of which contains a reflexive pronoun. In Experiment 1, these sentences look like (16), and all follow the basic template in (17).

(16) [Which cowgirl did Mary/David expect [to have injured herself/himself due to negligence?]]

(17) [Which NP_1 did NP_2 VERB [to have VERB-ed him-/herself SPILLOVER REGION]]

By independently varying the reflexive's gender congruency with the linearly local, grammatically inaccessible candidate antecedent NP on the one hand, and with the grammatically accessible (but linearly more distant) *wh*-NP on the other, we use on-line eye-tracking reading measurements of sentences like (16) to investigate whether reflexive antecedent search is immediately sensitive to the presence of a WhFGD or whether it initially considers linearly local but grammatically impossible antecedents.

Much previous work using the gender-mismatch effect as a probe for the parser's establishment of a long-distance dependency has utilized gender-stereotypic nouns like *doctor* and *nurse*. In contrast to this, the experiments reported here all use gender-categorical nouns like *cowgirl* or *uncle* and strongly gendered personal names like *Mary* or *Steven*. The reason for this design decision is that in piloting work, the subject population (Northwestern University undergraduates) was not found to exhibit a measurable gender-mismatch effect in response to gender-stereotypic nouns associated with (stereotypic-)gender mismatched reflexives. We do not speculate here as to the reason for this difference from previously studied populations except to say that it may be connected to changing social attitudes about the appropriateness of different professions for individuals of one or another gender. For our purposes it is sufficient that the study population *does* exhibit a gender-mismatch effect in response to gender-categorical nouns and strongly

gendered personal names associated with gender-mismatched reflexives[5].

2.1.2. Participants

Forty English speaking undergraduates from the Northwestern University community volunteered to participate in this experiment in return for course credit or a small monetary compensation. This experiment, and all experiments reported below, were approved by the Northwestern University Institutional Review Board as compliant with ethical standards for research on human subjects and were run under the protocol *Meaning in Language: Words, Sentences and Inferences (STU00025908)* or *Clausal Ellipsis: Its Structure and Online Processing (STU00082465)*.

2.1.3. Design and Materials

Materials consisted of 24 sentences like (18), with a complex *wh*-NP at the left edge associated with the subject position of an embedded non-finite clause, plus 140 filler sentences from unrelated, non-competing experiments. Comprehension questions were asked after 25% of trials in order to motivate the participants to attend to the experiment. This procedure is used in all following experiments as well. The gender match of the reflexive with the *wh*-NP and the linearly closer matrix-clause subject was independently varied in a two-by-two factorial design.

(18) Sample Stimuli

a. Which cowgirl did Mary expect to have injured herself due to negligence? // *wh*-NP match, local NP match.

b. Which cowgirl did David expect to have injured herself due to negligence? // *wh*-NP match, local NP mismatch.

c. Which cowgirl did David expect to have injured himself due to negligence? // *wh*-NP mismatch, local NP match.

d. Which cowgirl did Mary expect to have injured himself due to negligence? // *wh*-NP mismatch, local NP mismatch.

In this experiment, the embedded clause is non-finite (marked with the infinitival marker *to* and without agreement or independent tense-marking) and the tail of the *wh*-dependency headed by the *wh*-NP terminates in the embedded clause, after the position of the subject of the matrix clause. Although the embedded verbs were not formally normed for transitive or reflexive frame probabilities, they are all judged by the consensus of the native English speaking authors to be obligatorily transitive or highly transitively biased, and none are inherently reflexive. The subject of the matrix clause is thus the closest overt NP to the reflexive, and will consequently be referred to as the *linearly local* candidate antecedent, but because of the long-distance WhFGD between the *wh*-NP and the subject position of the embedded clause (17), only the *wh*-NP can be adopted as the antecedent for the reflexive in the final interpretation of the example[6]. For this reason the *wh*-NP is a *grammatically accessible* antecedent in the terminology we adopt here, and likewise the linearly local NP (the subject of the matrix clause) is a *grammatically inaccessible* antecedent.

In conditions (a) and (c), the reflexive matches the gender of the linearly closer but grammatically inaccessible matrix-clause subject. In conditions (a) and (b), the reflexive matches the gender of the linearly more distant but grammatically accessible *wh*-NP. Full experimental materials for this and all subsequent experiments are available in the online Supplementary Materials.

2.1.4. Predictions

If the process of antecedent search involved in reflexive resolution is sensitive to the output of WhFGD resolution, an early gender-mismatch effect should be observed when the gender of the grammatically accessible *wh*-NP mismatches that of the reflexive [i.e., in conditions (c) and (d)], and no gender-mismatch effect should be observed when the grammatically inaccessible, linearly local NP mismatches the gender of the reflexive, at least in early reading-time measures.

In contrast, if reflexive antecedent search is not sensitive to the output of WhFGD resolution and consists of a retrieval system that is not constrained to consider only grammatical antecedents, early gender-mismatch effects should be observed when the gender of the grammatically inaccessible, linearly local NP mismatches the gender of the reflexive. Depending upon the naive retrieval model adopted, several patterns of effects from the grammatically accessible *wh*-NP might be observed. Effects of gender mismatch with the *wh*-NP may be predicted to be observed only in later measures, if subjects select the linearly closest candidate antecedent on their initial parse. This might be the case if subjects are initially misled into a locally-coherent but globally ungrammatical parse (Ferreira et al., 2002; Tabor et al., 2004; Konieczny et al., 2010) in which the *wh*-phrase is not assigned an interpretation, as discussed above. In this case effects of the *wh*-NP would be expected to follow those of the linearly local candidate antecedent. On the other hand, if the reflexive antecedent retrieval system is a cue-based system that is not constrained to consider only grammatical antecedents, an effect of gender mismatch with the linearly local but grammatically inaccessible candidate antecedent should interact with that of the grammatically accessible *wh*-NP. In particular, the slowdown effect due to gender mismatch with the *wh*-NP should be ameliorated in the presence of a gender-matching inaccessible antecedent, because a cue-based retrieval system that is not restricted to consider only syntactically accessible candidate

[5]Note that a consequence of this experimental manipulation is that conditions in which the only grammatically-accessible antecedent for the reflexive mismatches it in gender are prima facie ungrammatical because, for example, *The cowgirl injured himself* or *Steven injured herself* may not be grammatical reflexive-antecedent dependencies. However, it is arguable whether examples like *The cowgirl injured himself* or *Steven injured herself* are genuinely ungrammatical, rather than simply unacceptable in the majority of contexts: it is not unimaginable that a woman might be named *Steven*, merely very unexpected, and likewise in the context of a costume party, the individual picked out by the referring expression *the cowgirl* could conceivably be male.

[6]Modulo dispreferred intensifier readings of the reflexive which will be addressed in the discussion below.

antecedents should be able to retrieve the gender-matching but grammatically inaccessible candidate antecedent.

Importantly however, a linguistically-naive antecedent retrieval process, whether cue-based or otherwise, should always show effects of the grammatically inaccessible, linearly local NP if any gender-mismatch effects are measurable at all. This is because such a process can by definition not distinguish potential candidate antecedents based upon the syntactic configurations in which they occur. For this reason, an effect of the grammatically accessible *wh*-NP in this experiment in the absence of an effect of the grammatically inaccessible, linearly local NP should be a clear signal of structure-sensitivity in the reflexive antecedent search mechanism.

2.1.5. Data Analysis

Using a tower-mounted EyeLink1000 eye-tracker, gaze was recorded and manually corrected for vertical drift. Fixations shorter than 80 ms were incorporated into adjacent fixations, and fixations longer than 2000 ms were excluded from analysis. The following analysis is based on four eye-tracking measures: first fixation duration, first pass duration, regression path duration, and re-read time. First fixation measures are based on the duration of the first time a fixation occurs within the region. First pass times include all time spent within the region before the first instance of the gaze exiting the region, either to the left or the right. Regression path duration is calculated by summing the times spent within the region and all time after exiting the region to the left until the first instance that the gaze exits to the right of the region. Re-read time is the sum of time spent within the region after the first time the gaze exits the region.

For the purposes of this study, we will concentrate on two regions of interest: the critical region containing the reflexive anaphor [e.g., *herself* in (19)], and the spillover region containing the remaining words on that line before the carriage return [e.g., *for unimportant* in (19)]. The stimuli were all displayed on two lines, due to character length limitations of the presentation software. The carriage returns were all in the same location and included in the post-spillover region, which is not analyzed in this study due to the complexity of interpreting fixations in regions that contain line breaks.

(19) Which saleswoman did Margaret presume to have excused herself for unimportant reasons?

In line with discussion in Barr et al. (2013), analyses were conducted by comparing a converging maximally inclusive linear mixed effects regression (LMER) model to a reduced model, i.e., a model with the same structure as the maximal model but with a single effect of interest removed from the fixed effects structure. Intercepts (β) and standard error (S.E.) were calculated from the maximal model. Maximal and reduced models were then compared by ANOVA to calculate the χ^2 and significance ($\alpha = 0.05$), reported in **Table 2**. The ideal maximal model for the critical region consisted of two independent factors (gender congruency with the *wh*-phrase; gender congruency with the local NP), and one additional fixed factor (presentation order).

TABLE 1 | Means (and Standard Errors) for Experiment 1.

Region		Critical region	Spillover region
Wh-phrase	**Local NP**		
FIRST FIXATION			
Match	Match	207 (6)	199 (5)
Match	Mismatch	203 (5)	214 (6)
Mismatch	Match	212 (5)	209 (7)
Mismatch	Mismatch	221 (6)	212 (7)
FIRST PASS			
Match	Match	221 (7)	312 (14)
Match	Mismatch	225 (7)	338 (15)
Mismatch	Match	238 (8)	335 (16)
Mismatch	Mismatch	247 (9)	343 (15)
REGRESSION PATH			
Match	Match	314 (23)	666 (64)
Match	Mismatch	293 (17)	680 (63)
Mismatch	Match	331 (25)	757 (57)
Mismatch	Mismatch	404 (34)	931 (78)
RE-READ TIME			
Match	Match	395 (28)	434 (42)
Match	Mismatch	297 (23)	391 (28)
Mismatch	Match	376 (28)	463 (37)
Mismatch	Mismatch	397 (25)	457 (31)

Intercepts were allowed to vary across subjects and items. We also allowed for the slopes of the following effects to vary across subjects and items: gender congruency of the *wh*-phrase, gender congruency of the local antecedent, the interaction of these two factors, and the presentation order. In cases where the maximal model failed to converge, the random effects correlation parameters were removed from the random effects structure (thus necessitating removal from reduced models as well). All models converged with either the ideal maximal model[7], or with the random effects correlations removed, as suggested in Bates et al. (2015). Data were contrast coded with conditions summing to 0 (i.e., *wh*- and local congruency conditions were coded as 0.5 or −0.5, respectively.) This coding scheme and analytical method is used for all experiments in this study. **Table 1** contains the means and standard errors in milliseconds of reading times. These measures were calculated after manual vertical alignment of fixations. For statistical analysis, converging maximal linear mixed effect models were compared via ANOVA to depleted models of the same structure, but with a term of interest removed. χ^2-values and their corresponding *p*-values are reported in **Table 2**, alongside the estimates and standard errors calculated from the corresponding maximal model. Bold values indicate that the comparison reached significance.

[7] For example: $lmer(rt \sim wh{*}lc + ord + (1 + wh{*}lc + ord|subj) + (1 + wh{*}lc + ord|item), data = data)$, where *rt* is the reading time, the predictors *wh* and *lc* are the gender match/mismatch of the *wh*-NP and the local candidate antecedent NP, respectively, and *ord* is the presentation order.

TABLE 2 | Combined ANOVA and LME results for Experiment 1.

Region	Effect	Estimate	Std error	χ^2 (df)	*p*-value
FIRST FIXATION					
Critical	*wh*–NP	−12.71	6.30	3.71 (1)	0.054
	local NP	−4.15	6.21	0.44 (1)	>0.1
	interaction	16.09	12.11	1.69 (1)	>0.1
Spillover	*wh*–NP	−4.23	7.03	0.36 (1)	>0.1
	local NP	−8.17	6.01	1.80 (1)	>0.1
	interaction	−10.05	13.80	0.52 (1)	>0.1
FIRST PASS					
Critical	*wh*–NP	**−19.55**	**8.32**	**6.65** (1)	**0.010**
	local NP	−6.15	8.84	1.36 (1)	>0.1
	interaction	9.58	16.83	0.36 (1)	>0.1
Spillover	*wh*-NP	−10.00	14.34	0.47 (1)	>0.1
	local NP	−21.83	13.83	2.32 (1)	>0.1
	interaction	−19.29	25.59	0.56 (1)	>0.1
REGRESSION PATH					
Critical	*wh*–NP	−59.88	30.65	3.43 (1)	0.064
	local NP	−24.55	24.95	0.94 (1)	>0.1
	interaction	90.75	49.05	3.40 (1)	0.065
Spillover	*wh*-NP	**−159.70**	**72.40**	**4.33** (1)	**0.037**
	local NP	−102.67	78.77	1.62 (1)	>0.1
	interaction	121.97	115.73	1.09 (1)	>0.1
RE-READ TIME					
Critical	*wh*-NP	−39.57	27.45	2.00 (1)	>0.1
	local NP	24.80	27.82	0.78 (1)	>0.1
	interaction	91.31	51.72	2.87 (1)	0.090
Spillover	*wh*-NP	−29.80	34.60	0.71 (1)	>0.1
	local NP	26.54	34.26	0.59 (1)	>0.1
	interaction	27.19	68.70	0.16 (1)	>0.1

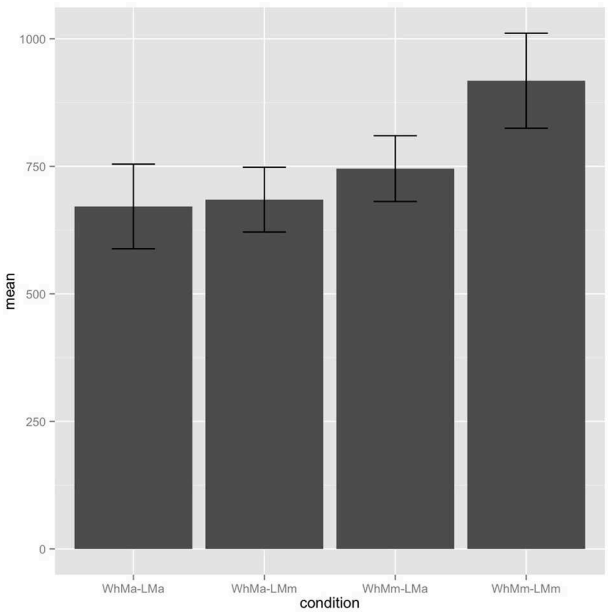


FIGURE 1 | Experiment 1 spillover region regression path durations.

2.1.6. Results

In the critical region, i.e., at the reflexive pronoun, we found a significant main effect of gender congruency between the *wh*-phrase and the reflexive, with matched gender read faster than mismatched gender, for first pass reading time [$\beta = -19.55$, S.E. $= 8.32$, $\chi^2(1) = 6.65$, $p = 0.010$]. This suggests that the parser is trying to form a dependency between the *wh*-phrase and the reflexive pronoun. When it successfully forms the dependency in the *wh*-phrase gender match condition, the reading time at the critical region is faster than when it is unsuccessful in the *wh*-phrase gender mismatch condition.

In the spillover region, we observe a significant main effect of gender congruency between the *wh*-phrase and the reflexive, with matched gender read faster than mismatched gender for regression path duration [$\beta = -159.70$, S.E. $= 72.40$, $\chi^2(1) = 4.33$, $p = 0.037$], (**Figure 1**). No other effects reached significance.

There were, however, marginal interactions of *wh*-phrase gender congruence with local NP congruence in the regression path duration and re-reading time in the critical region, such that the mismatch-mismatch condition was read more slowly. Although this interaction was not statistically significant it is consistent with the predictions of some unconstrained cue-based models of antecedent retrieval. On an explanation of this kind, the parser would attempt to associate the reflexive with all possible candidate antecedents in parallel and experience extra difficulty when no gender-congruent antecedent is found in its memory store. In the absence of a significant effect, this is of course a purely speculative suggestion.

The main effect we observe in the spillover region is consistent with the effect at the critical region and supports the hypothesis that the parser represents the tail of the *wh*-dependency and is thus able to connect the *wh*-phrase and the reflexive pronoun. This suggests that the presence of the WhFGD is accessible to the process of RD resolution. In other words, since the parser has already linked the *wh*-phrase with the gap, the search for the RD does not allow the parser to consider the interpretation in which the linearly closest antecedent (i.e., the proper name) is linked with the gap. The effect of gender mismatch of the *wh*-phrase in the absence of an effect of the linearly local but grammatically inaccessible candidate antecedent supports the hypothesis that the reflexive antecedent retrieval system is constrained to consider only grammatically accessible antecedents. However, the marginal interaction with the gender-match/mismatch of the linearly local candidate antecedent suggests a possible signature of a cue-based retrieval system that is not constrained to consider only grammatically accessible antecedents, which forms much of the motivation for Experiment 2.

In addition to the reflexive interpretation proper, the English pronouns ending in *-self* have at least two other interpretations which are subject to different syntactic constraints[8]. In an *emphatic* reading of a *-self* -type pronoun in English, the pronoun, though formally reflexive, does not have a properly reflexive

[8]Thanks to Dave Kush (*personal communication*) for pointing out the possibility of this reading for the stimuli in Experiment 1.

reading (roughly, indicating that the object of the verb refers to the same entity as the subject). Emphatic reflexives instead have a focus-related meaning emphasizing that some entity referred to by an NP associated with the reflexive was involved in the event described by the sentence, rather than any other entity that might have been involved in the event. So in (20-a), the emphatic reflexive *himself* is associated with the matrix subject *John* and serves to emphasize that John's expectation was that he himself, and not someone else, would have injured the cowgirl.

(20) a. John expected to have injured the cowgirl himself.
b. *John expected had injured the cowgirl himself.

In an *anti-assistive* reading of a *-self*-type pronoun, the *-self*-type pronoun serves to indicate that the agent of the sentence performed the action in question *without help*, so in (20-a), such a reading would mean that John expected to have received no assistance in injuring the cowgirl. Because control into finite embedded clauses is impossible in English (20-b), emphatic and anti-assistive reflexive readings for sentences like the stimuli for Experiment 2 (discussed below), with finite embedded clauses, are not possible.

2.2. Experiment 2

2.2.1. Introduction

In order to demonstrate that the effect of the *wh*-NP observed in Experiment 1 is, in fact, a consequence of the *wh*-dependency and not some other factor, we should replicate these results in a syntactically different context, but one that is similar in all respects that this account predicts to be relevant for the pattern of results observed in Experiment 1: namely, the presence of a dependency tail associated with the sentence-initial *wh*-NP after the linearly closest candidate antecedent. This is the primary purpose of Experiment 2.

Experiment 2 also serves to distinguish the possibility that the marginal interactions with the gender congruence of the local NP result from retrieval difficulty from the possibility that they result from the parser's later consideration of the dispreferred non-reflexive readings for the *-self* pronoun.

2.2.2. Participants

Forty English speaking undergraduates from the Northwestern University community volunteered to participate in this experiment in return for course credit or a small monetary compensation.

2.2.3. Design and Materials

Materials for Experiment 2 consisted of 24 target sentences, plus 90 filler sentences from unrelated experiments. The target stimuli used in Experiment 2 are based upon those used in Experiment 1, with one relevant difference. While the target stimuli from Experiment 1 include non-finite embedded clauses, those in Experiment 2 use finite embedded clauses, as exemplified in (21).

(21) Sample Stimuli
a. Which cowgirl did Mary expect had injured herself due to negligence? // *wh*-NP match, local NP match.
b. Which cowgirl did David expect had injured herself due to negligence? // *wh*-NP match, local NP mismatch.
c. Which cowgirl did David expect had injured himself due to negligence? // *wh*-NP mismatch, local NP match.
d. Which cowgirl did Mary expect had injured himself due to negligence? // *wh*-NP mismatch, local NP mismatch.

This difference has two related effects on the possible behavior of the parser in these sentences. First, because finite complement clauses in English do not permit control readings (22), there is no potential locally coherent substring of these examples in which the grammatically inaccessible, linearly local candidate antecedent NP is a grammatical antecedent for the reflexive. Given that effects of the linearly local candidate antecedent were not observed in Experiment 1, this difference is not expected to influence reading time measures.

(22) *Susan expected had injured herself.

A related but more important difference is that, precisely *because* a control reading is not possible for examples like those in (21), these examples do not admit of intensifier readings for the reflexive. For this reason, then, there is no grammatical possibility of linking the reflexives in the embedded clause with the matrix subject.

It is not clear how the possibility of an intensifier reflexive reading might have contaminated the primary results of Experiment 1, given that the observed effects were not compatible with such a reading (i.e., they did not indicate that participants were attempting to associate the reflexive with the matrix subject rather than with the *wh*-NP). However, because intensifier reflexives are subject to somewhat different syntactic constraints than reflexives proper, it was deemed worthwhile to ensure that a similar pattern of results obtained in the absence of any possibility of such a reading. Moreover, if the marginal interactions reported above do result from the parser's consideration of an intensifier reading for the reflexive, they should disappear in a context where this is not possible. In contrast, if they arise from interference in the antecedent retrieval process proper, they should be expected to persist.

2.2.4. Predictions

The results of Experiment 2 are predicted to be broadly similar to those of Experiment 1: namely, if reflexive resolution is sensitive to presence of a WhFGD and constrained to consider only grammatically accessible antecedents, a gender-mismatch effect should be observed when the gender of the grammatically accessible *wh*-NP mismatches that of the reflexive [i.e., in conditions (c) and (d)], and no gender-mismatch effect should be observed when the grammatically inaccessible, linearly local NP mismatches the gender of the reflexive. If reflexive antecedent search is not constrained to consider only grammatically accessible antecedents, gender-mismatch effects should be observed when the gender of the grammatically inaccessible, linearly local NP mismatches the gender of the

reflexive. If this is because of the antecedent search process's susceptibility to linear closeness, gender mismatch effects from the linearly local candidate antecedent should precede any from the *wh*-NP. If instead antecedent retrieval consists of a cue-based retrieval system that is able to consider ungrammatical reflexive antecedents, gender mismatch effects of both candidate antecedents should interact in such a way that the slowdown effect induced by mismatch with the *wh*-NP is ameliorated in the presence of a gender-matching ungrammatical candidate antecedent.

However, because in the finite embedded clauses used in Experiment 2 no control reading is possible, it is not possible to interpret the reflexive in the examples used in Experiment 2 as an intensifier reflexive linked to the matrix subject, so this experiment may constitute a cleaner test of the role of the binding constraints in reflexive antecedent search. It is not expected that the pattern of effects in this experiment will differ from that in Experiment 1; if it does, this would cast doubt upon an explanation of the effect of the *wh*-NP in Experiment 1 in terms of the parser's online sensitivity to find-grained syntactic constraints.

2.2.5. Data Analysis

The analysis of the data gathered in this experiment was carried out in much the same way as in Experiment 1. The critical region corresponds to the reflexive pronoun (*herself*) and the spillover region corresponds to *for unimportant* in the example below. Since the stimuli used in this experiment are adapted from Experiment 1, the same limitations on region size due to line breaks constrained the spillover region.

(23) Which saleswoman did Margaret presume had excused herself for unimportant reasons?

Table 3 contains the means and standard errors in milliseconds of reading times. These measures were calculated after manual vertical alignment of fixations. For statistical analysis, converging maximal linear mixed effect models were compared via ANOVA to depleted models of the same structure, but with a term of interest removed. χ^2-values and their corresponding p-values are reported in **Table 4**, alongside the estimates and standard errors calculated from the corresponding maximal model. The ideal maximal structure contains the same terms as in Experiment 1, and in cases where a maximal or depleted model did not converge, additional terms were removed in the order specified above.

2.2.6. Results

In the critical region, the only observed effects are in re-read time. We observe a significant main effect of the *wh*-phrase, with the gender matched condition read faster than gender mismatched conditions [$\beta = -129.22$, S.E. $= 31.34$, $\chi^2(1) = 12.43$, $p < 0.001$]. This is consistent with the observations in Experiment 1, that the local NP is not considered as a candidate antecedent for the reflexive pronoun, despite its linear proximity. No other effects reached significance.

The spillover region displays a similar pattern of effects, with the addition of significant main effects of *wh*-phrase observed in first pass reading time, regression path duration (**Figure 2**), and re-read time, with gender match between the *wh*-phrase and the reflexive pronoun read faster than gender mismatch [first pass: $\beta = -42.53$, S.E. $= 13.92$, $\chi^2(1) = 8.21$, $p = 0.004$; regression path: $\beta = -831.51$, S.E. $= 140.41$, $\chi^2(1) = 21.11$, $p < 0.001$], as in the case for re-read time [$\beta = -107.00$, S.E. $= 43.43$, $\chi^2(1) = 5.13$, $p = 0.023$].

As before, this indicates that the gender of the *wh*-phrase is somehow represented at the tail of the WhFGD, which is then being accessed during the reflexive antecedent search. These results are compatible with our observations in Experiment 1. As such, we can confirm that the gender mismatch effects observed in the critical region and spillover region in both Experiments 1 and 2 are due to the ability of the parser to form a dependency between the reflexive and the gap, although for different reasons. There were no marginal effects of the linearly local candidate antecedent NP in this experiment, unlike in Experiment 1, suggesting that the intensifier reading explanation for those effects in Experiment 1 may be on the right track, rather than an interpretation in terms of failed cue-based retrieval. This is of course merely speculation, given that the effects in question do not reach statistical significance.

TABLE 3 | Means (and Standard Errors) for Experiment 2.

Region		Critical region	Spillover region
Wh-phrase	**Local NP**		
FIRST FIXATION			
Match	Match	209 (6)	204 (5)
Match	Mismatch	214 (5)	208 (6)
Mismatch	Match	222 (6)	219 (6)
Mismatch	Mismatch	213 (4)	213 (6)
FIRST PASS			
Match	Match	223 (7)	293 (12)
Match	Mismatch	238 (8)	319 (14)
Mismatch	Match	250 (9)	356 (18)
Mismatch	Mismatch	237 (8)	348 (18)
REGRESSION PATH			
Match	Match	461 (55)	1860 (119)
Match	Mismatch	454 (45)	1867 (112)
Mismatch	Match	620 (66)	2656 (171)
Mismatch	Mismatch	611 (73)	2744 (197)
RE-READ TIME			
Match	Match	300 (32)	382 (38)
Match	Mismatch	335 (26)	381 (30)
Mismatch	Match	463 (33)	477 (35)
Mismatch	Mismatch	426 (28)	494 (34)

2.3. Experiment 3

2.3.1. Introduction

Experiment 3 (as well as Experiment 4, discussed below) serves as a check to ensure that the difference observed in Experiments 1 and 2 between the effect of gender match/mismatch of the reflexive and the *wh*-NP and of the reflexive and the linearly closer NP is not due to some difference between the way *wh*-NPs

TABLE 4 | Combined ANOVA and LME results for Experiment 2.

Region	Effect	Estimate	Std error	χ^2 (df)	*p*-value
FIRST FIXATION					
Critical	*wh*-NP	−6.30	5.05	1.55 (1)	>0.1
	local NP	3.19	5.48	0.34 (1)	>0.1
	interaction	−11.09	10.79	1.02 (1)	>0.1
Spillover	*wh*-NP	−8.79	7.92	1.20 (1)	>0.1
	local NP	3.31	6.05	0.30 (1)	>0.1
	interaction	−13.48	12.70	1.11 (1)	>0.1
FIRST PASS					
Critical	*wh*-NP	−13.11	7.44	3.08 (1)	0.079
	local NP	−0.05	8.30	0.0001 (1)	>0.1
	interaction	0.22	7.75	2.60 (1)	>0.1
Spillover	*wh*-NP	**−42.53**	**13.92**	**8.21** (1)	**0.004**
	local NP	−11.07	13.60	0.66 (1)	>0.1
	interaction	−33.21	31.13	1.11 (1)	>0.1
REGRESSION PATH					
Critical	*wh*-NP	−140.07	81.96	2.80 (1)	0.094
	local NP	−13.73	60.67	0.05 (1)	>0.1
	interaction	41.56	132.41	0.10 (1)	>0.1
Spillover	*wh*-NP	**−831.51**	**140.41**	**21.11** (1)	**<0.001**
	local NP	−91.67	135.46	0.45 (1)	>0.1
	interaction	152.58	309.65	0.24 (1)	>0.1
RE-READ TIME					
Critical	*wh*-NP	**−129.22**	**31.34**	**12.43** (1)	**<0.001**
	local NP	−7.94	30.75	0.06 (1)	>0.1
	interaction	−80.55	62.04	1.61 (1)	>0.1
Spillover	*wh*-NP	**−107.00**	**43.43**	**5.13 (1)**	**0.023**
	local NP	−14.26	40.13	0.12 (1)	>0.1
	interaction	−6.61	82.07	0.006 (1)	>0.1

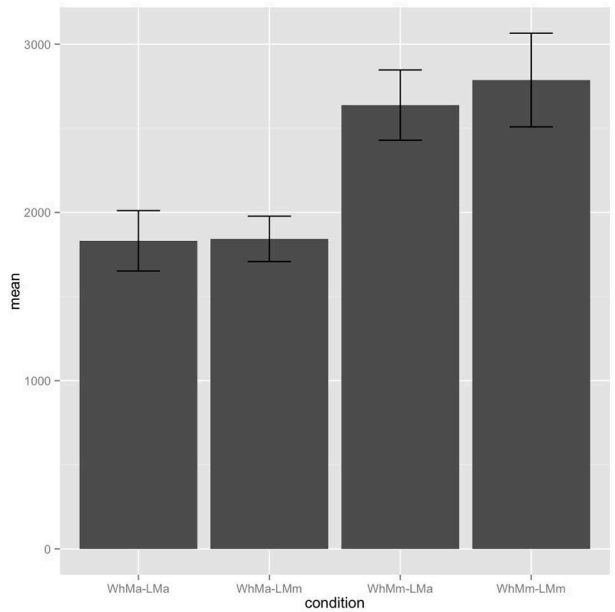


FIGURE 2 | Experiment 2 spillover region regression path durations.

and personal names are processed in general. For example, the results of Martin and McElree (2011) indicate that *wh*-NPs may have a higher prominence in memory, inasmuch as they are candidates for antecedent retrieval, than other categories. Therefore, there is a possibility that the results of Experiments 1 and 2 are not demonstrating grammar-sensitivity on the part of the parser's reflexive antecedent search process, but are instead merely demonstrating that *wh*-NPs are treated differently in memory than other NPs in some way that causes them to induce gender mismatch effects on subsequently encountered reflexives.

For this reason, in Experiment 3, the WhFGD originating in the sentence-initial *wh*-NP does not span across the linearly local NP but terminates before it, in the matrix clause, as in (24).

(24) Which cowgirl expected Mary to have injured herself due to negligence?

This has the effect that the *wh*-NP, though equally distant from the reflexive, is not its grammatical antecedent. If the effect of the *wh*-NP observed in Experiments 1 and 2 is due to a general high salience of *wh*-NPs in memory, the pattern of effects in this experiment should be largely the same here. On the other hand, if the effect of the *wh*-NP on RTs at and following the reflexive in Experiments 1 and 2 is due to the parser's sensitivity to the presence of a WhFGD intervening between the more linearly local candidate antecedent and the reflexive, the linearly local candidate antecedent should modulate RTs at the reflexive in this experiment rather than the *wh*-NP.

2.3.2. Participants

Forty English speaking undergraduates from the Northwestern University community volunteered to participate in this experiment in return for course credit or a small monetary compensation.

2.3.3. Design and Materials

Materials for Experiment 3 consist of 24 target sentences and 88 filler sentences from unrelated experiments. Experiment 3 (as well as Experiment 4, discussed below) serves as a check to ensure that the difference observed in Experiments 1 and 2 between the effect of gender match/mismatch of the reflexive and the *wh*-NP and of the reflexive and the linearly closer NP is not due to some difference between the way *wh*-NPs and personal names are processed in general. That is, there is a possibility that the results of Experiments 1 and 2 are not demonstrating grammar-sensitivity on the part of the parser's reflexive antecedent search process, but are instead merely demonstrating that *wh*-NPs are treated differently in memory than other NPs in some way that causes them to induce gender mismatch effects on subsequently encountered reflexives. For this reason, in Experiment 3, the WhFGD originating in the sentence-initial *wh*-NP does not span across the linearly local NP but terminates before it, as in (25).

(25) Sample Stimuli

a. Which cowgirl expected Mary to have injured herself due to negligence? // *wh*-NP match, local NP match.

b. Which cowgirl expected David to have injured herself due to negligence? // *wh*-NP match, local NP mismatch.
c. Which cowgirl expected David to have injured himself due to negligence? // *wh*-NP mismatch, local NP match.
d. Which cowgirl expected Mary to have injured himself due to negligence? // *wh*-NP mismatch, local NP mismatch.

If the effect of the gender match of the *wh*-NP in Experiments 1 and 2 is to be attributed to the parser's sensitivity to the WhFGD between the *wh*-NP and the embedded clause, this effect should go away when the WhFGD is not associated with the embedded clause but instead with the matrix clause, as in (25). On the other hand, if the role of the *wh*-NP in modulating reading times of the reflexive is due to a high overall salience of *wh*-NPs in memory, it should persist in this experiment.

2.3.4. Predictions

If the patterns of effects observed in Experiments 1 and 2–broadly, effects of the *wh*-NP's gender match/mismatch with the reflexive on the reading times of the reflexive–is due to the parser's grammatical sensitivity to the presence of a long-distance WhFGD between the sentence-initial *wh*-word and the embedded clause, the result in this experiment should be very different. In particular, because no such long-distance WhFGD between the sentence-initial *wh*-word and the embedded clause is present in the stimuli used in Experiment 3, no effect of the *wh*-NP's gender match/mismatch with the reflexive should be observed in this experiment. On the other hand, if the effect of the *wh*-NP on the reading time of the reflexive in Experiments 1 and 2 is due, in whole or in part, to a difference between the way that the parser treats previously-processed *wh*-NPs and the way it treats previously-processed personal names, an effect of the gender match/mismatch of the *wh*-NP should be observed in this experiment as well. If the results of Experiments 1 and 2 are due entirely to a difference between the behavior of previously-processed *wh*-NPs and personal names, then, the results of this experiment should be the same as those of Experiments 1 and 2. If a difference between the behavior of previously-processed *wh*-NPs and personal names is a contributor to the pattern of results in Experiments 1 and 2 but not the sole driver of the effect, with grammar-sensitivity of the parser also being implicated, then an effect of the gender match/mismatch of both candidate antecedent NPs, the *wh*-NP and the linearly local NP, should be observed. As above, if the antecedent retrieval system is a cue-based retrieval system that is not constrained to consider only grammatical antecedents, an interaction effect should be observed such that the gender mismatch effect due to the grammatically accessible antecedent (in this case, the linearly local antecedent rather than the *wh*-NP) should be ameliorated provided the other candidate antecedent is gender-matched with the reflexive.

2.3.5. Data Analysis

The analysis of the data gathered in this experiment was carried out in much the same way as in the previous two experiments. The critical region corresponds to the reflexive pronoun (*herself*) and the spillover region corresponds to *for unimportant* below. The stimuli used in this experiment are again adapted from Experiment 1 and the same limitations on region size due to line breaks constrained the spillover region. The critical difference between the stimuli in Experiments 1 and 2 and the current set is that the *wh*-phrase is no longer accessible to the reflexive pronoun. Rather, the local antecedent is the globally coherent and accessible antecedent.

(26) Which saleswoman presumed Margaret to have excused herself for unimportant reasons?

Table 5 displays the means and standard errors in milliseconds of reading times, calculated after manual vertical alignment of fixations. For statistical analysis, converging maximal linear mixed effect models were compared via ANOVA to depleted models of the same structure, but with a term of interest removed. χ^2-values and their corresponding *p*-values are reported in **Table 6**, alongside the estimates, and standard errors calculated from the corresponding maximal model. The ideal maximal structure contains the same terms as in previous experiments, and in cases where a maximal or depleted model did not converge, additional terms were removed in the order previously specified.

2.3.6. Results

In this experiment, we observe the expected reverse in effect source, now with the gender of the local NP influencing reading times in the critical region. Here, we observe a main effect of local NP in the critical region's re-read time, with the gender matched conditions read faster than gender mismatched

TABLE 5 | Means (and Standard Errors) for Experiment 3.

Region		Critical region	Spillover region
Wh-phrase	**Local candidate**		
FIRST FIXATION			
Match	Match	191 (4)	195 (5)
Match	Mismatch	200 (5)	212 (6)
Mismatch	Match	198 (4)	200 (5)
Mismatch	Mismatch	200 (4)	207 (5)
FIRST PASS			
Match	Match	205 (5)	319 (13)
Match	Mismatch	220 (7)	332 (13)
Mismatch	Match	217 (6)	307 (11)
Mismatch	Mismatch	211 (5)	345 (16)
REGRESSION PATH			
Match	Match	397 (41)	2068 (109)
Match	Mismatch	527 (59)	2907 (146)
Mismatch	Match	412 (52)	2117 (116)
Mismatch	Mismatch	423 (42)	2968 (181)
RE-READ TIME			
Match	Match	305 (21)	333 (25)
Match	Mismatch	413 (26)	482 (31)
Mismatch	Match	339 (35)	394 (31)
Mismatch	Mismatch	447 (33)	568 (69)

TABLE 6 | Combined ANOVA and LME results for Experiment 3.

Region	Effect	Estimate	Std error	χ^2 (df)	*p*-value
FIRST FIXATION					
Critical	*wh*-NP	−4.64	4.76	0.93 (1)	>0.1
	local	−7.79	4.25	3.24 (1)	0.072
	interaction	−8.00	10.64	0.55 (1)	>0.1
Spillover	*wh*-NP	−0.35	6.10	0.003 (1)	>0.1
	local	−10.98	6.27	2.76 (1)	0.097
	interaction	−8.22	10.47	0.60 (1)	>0.1
FIRST PASS					
Critical	*wh*-NP	−0.57	6.29	0.01 (1)	>0.1
	local	−5.28	7.01	0.55 (1)	>0.1
	interaction	−24.08	16.89	1.95 (1)	>0.1
Spillover	*wh*-NP	2.24	13.87	0.026 (1)	>0.1
	local	−25.92	13.24	3.44 (1)	0.064
	interaction	23.27	26.80	0.73 (1)	>0.1
REGRESSION PATH					
Critical	*wh*-NP	43.67	44.18	0.98 (1)	>0.1
	local	−74.27	45.67	2.58 (1)	>0.1
	interaction	−111.73	96.88	1.29 (1)	>0.1
Spillover	*wh*-NP	−29.80	120.90	0.06 (1)	>0.1
	local	**−844.80**	**117.30**	**27.58** (1)	**<0.001**
	interaction	113.20	253.30	0.20 (1)	>0.1
RE-READ TIME					
Critical	*wh*-NP	−44.40	29.08	2.18 (1)	>0.1
	local	**−100.51**	**35.59**	**6.62** (1)	**0.010**
	interaction	3.84	81.52	0.002 (1)	>0.1
Spillover	*wh*-NP	−79.11	59.80	1.64 (1)	>0.1
	local	**−164.91**	**69.38**	**4.97** (1)	**0.026**
	interaction	27.40	141.89	0.04 (1)	>0.1

condition [$\beta = -100.51$, S.E. $= 35.59$, $\chi^2(1) = 6.62$, $p = 0.010$]. We also observe a main effect of gender mismatch in the regression path duration (**Figure 3**) and re-read time) in the spillover region [regression path: $\beta = -844.80$, S.E. $= 117.30$, $\chi^2(1) = 27.58$, $p < 0.001$; re-read time: $\beta = -164.91$, S.E. $= 69.38$, $\chi^2(1) = 4.97$, $p = 0.026$]. No other effects reached significance. Note however that all marginal effects are of the local NP, consistent with the parser only considering this NP as a potential reflexive antecedent. Thus, this supports the hypothesis that the results of Experiments 1 and 2 are due to the RD resolution process being sensitive to the presence of the WhFGD, rather than being due to some general property of *wh*-NPs as candidate antecedents.

2.4. Experiment 4

2.4.1. Introduction

Experiment 4 serves primarily to complete the paradigm explored in Experiments 1–3, so that over the course of all four experiments all combinations of *finite vs. nonfinite embedded clause* and *matrix interpretation of wh-word vs. embedded WhFGD tail* are investigated. The results of this experiment are not expected to differ from those of Experiment 3 except that, because of certain differences between finite and non-finite embedded clauses, as discussed below, the effect of the local candidate antecedent may be stronger in Experiment 4 than in Experiment 3.

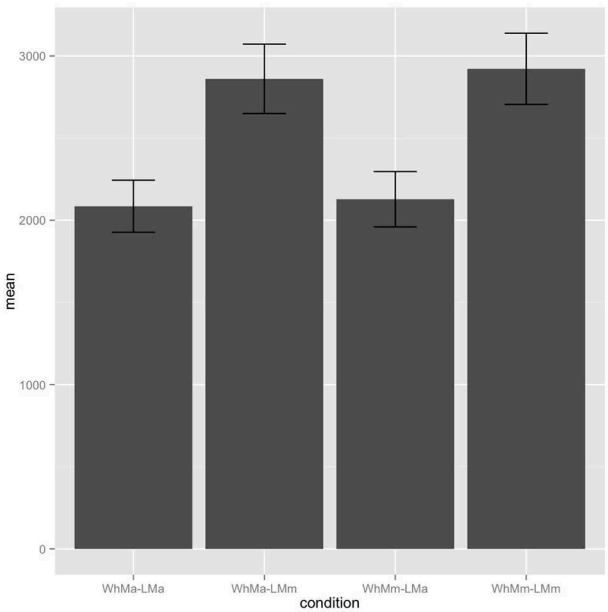


FIGURE 3 | Experiment 3 spillover region regression path durations.

2.4.2. Participants

Twenty English speaking undergraduates from the Northwestern University community volunteered to participate in this experiment in return for course credit or a small monetary compensation[9].

2.4.3. Design and Materials

The design of Experiment 4 is substantially the same as that of Experiment 3. The materials consist of 24 target sentences, plus 144 filler sentences from unrelated experiments. Like in Experiment 3, the sentence-initial *wh*-word is associated not with the embedded clause but with the matrix clause, and consequently it is not associated with a dependency tail intervening between the linearly local candidate antecedent and the reflexive. The difference between Experiments 3 and 4 is that in Experiment 4, as in Experiment 2, the embedded clause is finite rather than non-finite, as in (27).

(27) Sample Stimuli

a. Which cowgirl expected Mary had injured herself due to negligence? // *wh*-NP match, local NP match.

b. Which cowgirl expected David had injured herself due to negligence? // *wh*-NP match, local NP mismatch.

[9]The smaller number of participants in this experiment is due to accidental exclusion of the target stimuli during compiling for experimental presentation in half of the presentation orders. Fortunately, the conditions remain properly counterbalanced.

c. Which cowgirl expected David had injured himself due to negligence? // *wh*-NP mismatch, local NP match.
d. Which cowgirl expected Mary had injured himself due to negligence? // *wh*-NP mismatch, local NP mismatch.

2.4.4. Predictions

Because the only difference between Experiments 4 and 3 is the finiteness of the embedded clause, the result of this experiment is not expected to differ from that of Experiment 3. In particular, in this experiment as well, no long-distance WhFGD between the sentence-initial *wh*-word and the embedded clause is present in the stimuli used. Therefore, no effect of the *wh*-NP's gender match/mismatch with the reflexive should be observed in this experiment if the effect of the *wh*-NP's gender match/mismatch with the reflexive observed in the results of Experiments 1 and 2 is due to the parser's grammatical sensitivity to the presence of a long-distance WhFGD whose tail intervenes between the linearly closer NP and the reflexive. Likewise, if the effect of the *wh*-NP's gender match/mismatch with the reflexive is due to a general processing difference between *wh*-NPs and other NPs, it should be observed in this experiment as well. As in the preceding experiments, if the antecedent retrieval system is a cue-based retrieval system that is not constrained to consider only grammatical antecedents, an interaction effect should be observed such that the gender mismatch effect due to the grammatically accessible antecedent should be ameliorated provided the other candidate antecedent is gender-matched with the reflexive.

However, one possible small difference may be observed because of the similarity of embedded finite clauses to matrix clauses in English. Note that in example (27), if the initial words *which cowgirl expected* were omitted, the example would be the entirely grammatical matrix declarative sentences *Mary/David had injured himself/herself due to negligence*, until the presence of the question mark. It is conceivable that in these examples, for this reason, the association of the reflexive with the linearly local NP may be easier for the parser to detect, because of the similarity of these examples to simple matrix sentences in which there is only one candidate antecedent. If something like this is the case, we might expect the effect of the local candidate NP to reach significance for more reading-time measures than in Experiment 3.

2.4.5. Data Analysis

The analysis of the data gathered in this experiment was carried out in much the same way as in the previous three experiments. The critical region corresponds to the reflexive pronoun (*herself*) and the spillover region corresponds to *for unimportant*. Using the same design as in Experiment 3, the *wh*-phrase (i.e., *Which saleswoman*) is inaccessible to the reflexive pronoun as an antecedent, while the local antecedent (i.e., *Margaret*) is accessible. The limitations on region size due to line breaks constrained the spillover region, as in the previous experiments.

(28) Which saleswoman presumed Margaret had excused herself for unimportant reasons?

Table 7 displays the means and standard errors in milliseconds of reading times, calculated after manual vertical alignment of fixations. For statistical analysis, converging maximal linear mixed effect models were compared via ANOVA to depleted models of the same structure, but with a term of interest removed. χ^2-values and their corresponding *p*-values are reported in **Table 8**, alongside the estimates and standard errors calculated from the corresponding maximal model. The ideal maximal structure contains the same terms as in previous experiments, and in cases where a maximal or depleted model did not converge, additional terms were removed in the order previously specified.

2.4.6. Results

The results of Experiment 4 reveal a significant effect in regression path duration and re-read time, consistent with Experiment 3. This main effect of local NP in the critical region reveals that gender incongruency between the local NP and the reflexive pronoun led to an increased duration than when the gender matched [regression path: $\beta = -81.82$, S.E. $= 38.71$, $\chi^2(1) = 3.95$, $p = 0.047$; re-read time: $\beta = -222.01$, S.E. $= 44.71$, $\chi^2(1) = 13.25$, $p < 0.001$].

The pattern of increased durations in local mismatches is also observed in the spillover region [regression path: $\beta = -251.72$, S.E. $= 96.04$, $\chi^2(1) = 6.19$, $p = 0.013$, (**Figure 4**); re-read time: $\beta = -137.90$, S.E. $= 56.17$, $\chi^2(1) = 4.68$, $p = 0.031$]. This result is consistent with our claim that the parser

TABLE 7 | Means (and Standard Errors) for Experiment 4.

Region		Critical region	Spillover region
Wh-phrase	**Local candidate**		
FIRST FIXATION			
Match	Match	216 (7)	196 (6)
Match	Mismatch	229 (8)	215 (10)
Mismatch	Match	220 (8)	202 (7)
Mismatch	Mismatch	231 (9)	218 (9)
FIRST PASS			
Match	Match	233 (8)	387 (30)
Match	Mismatch	260 (12)	329 (22)
Mismatch	Match	241 (9)	371 (27)
Mismatch	Mismatch	259 (13)	385 (25)
REGRESSION PATH			
Match	Match	319 (26)	527 (59)
Match	Mismatch	352 (35)	843 (94)
Mismatch	Match	287 (23)	577 (83)
Mismatch	Mismatch	412 (46)	830 (153)
RE-READ TIME			
Match	Match	262 (29)	337 (59)
Match	Mismatch	468 (49)	514 (60)
Mismatch	Match	248 (24)	390 (39)
Mismatch	Mismatch	449 (54)	468 (55)

TABLE 8 | Combined ANOVA and LME results for Experiment 4.

Region	Effect	Estimate	Std error	χ^2 (df)	*p*-value
FIRST FIXATION					
Critical	*wh*-NP	−2.12	10.12	0.04 (1)	>0.1
	local	−9.65	12.43	0.58 (1)	>0.1
	interaction	−1.01	8.14	0.03 (1)	>0.1
Spillover	*wh*-NP	−4.33	7.32	0.35 (1)	>0.1
	local	−15.31	9.74	2.30 (1)	>0.1
	interaction	−1.13	15.72	0.005 (1)	>0.1
FIRST PASS					
Critical	*wh*-NP	−1.66	11.87	0.02 (1)	>0.1
	local	−24.01	14.30	2.57 (1)	>0.1
	interaction	−7.99	24.89	0.10 (1)	>0.1
Spillover	*wh*-NP	−19.31	24.88	0.59 (1)	>0.1
	local	19.00	27.92	0.46 (1)	>0.1
	interaction	62.83	51.55	1.42 (1)	>0.1
REGRESSION PATH					
Critical	*wh*-NP	−22.48	38.62	0.33 (1)	>0.1
	local	**−81.82**	**38.71**	**3.95** (1)	**0.047**
	interaction	88.63	69.32	1.56 (1)	>0.1
Spillover	*wh*-NP	0.99	126.68	0.0001 (1)	>0.1
	local	**−251.72**	**96.04**	**6.19 (1)**	**0.013**
	interaction	−138.84	212.14	0.42 (1)	>0.1
RE-READ TIME					
Critical	*wh*-NP	−13.29	57.62	0.052 (1)	>0.1
	local	**−222.01**	**44.71**	**13.25 (1)**	**<0.001**
	interaction	49.51	141.01	0.12 (1)	>0.1
Spillover	*wh*-NP	7.50	56.04	0.018 (1)	>0.1
	local	**−137.90**	**56.17**	**4.68** (1)	**0.031**
	interaction	−37.89	110.71	0.11 (1)	>0.1

searches a sophisticated structural representation during reflexive antecedent dependency formation. That is, the presence of an effect from the local NP supports the results of Experiment 3 in demonstrating that the parser is sensitive to the tail of the WhFGD as well as other rich, phonologically null representations in the parse tree.

3. Discussion

The current study sought to investigate the interaction of the resolution of two non-local dependencies, *wh*-filler-gap dependencies (WhFGD) and reflexive-antecedent dependencies (RD). In particular, we investigated the time course of the online resolution of a RD in the context of grammatically accessible and inaccessible possible antecedents. Experiments 1 and 2 examined whether the RD resolution process would target a linearly closer but grammatically illicit antecedent, or whether instead the grammatically licit tail of a WhFGD would be selected as the reflexive antecedent. Experiments 3 and 4 examined whether a possible antecedent that was both grammatically illicit and linearly more distant from a grammatically licit antecedent would influence the reflexive antecedent search. Results from these four eye-tracking text-reading experiments indicates that the RD

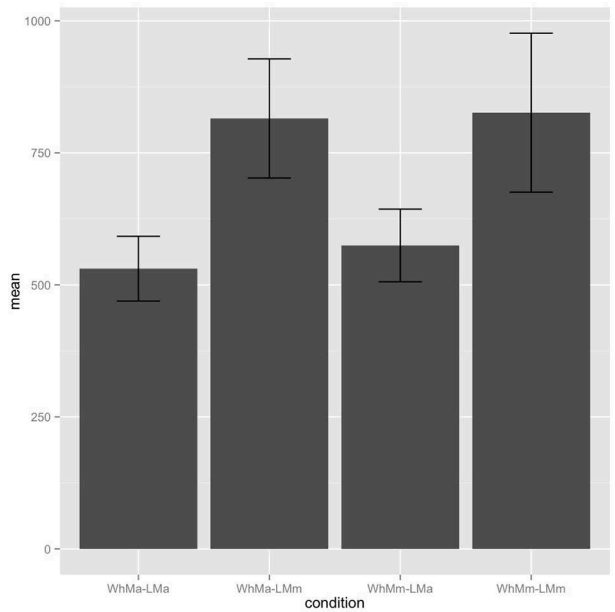


FIGURE 4 | Experiment 4 spillover region regression path durations.

resolution process is sensitive to grammatical structure, not local linear order.

Experiment 1 examined the time course of online reading of examples from the paradigm illustrated in (16). Crucially, such sentences are locally ambiguous; the string subsequent to the *wh*-phrase could be a coherent, grammatical utterance in which the antecedent for the reflexive would be a proper name. Globally, however, such a parse, with the proper name serving as antecedent for the reflexive, is unavailable; the only globally coherent parse is one in which the *wh*-phrase serves as the reflexive antecedent. Thus, if the search for reflexive antecedence is insensitive to syntactic structure, selecting either any feature-matching possible antecedent without regard to its structural position or the linearly closest possible antecedent, the proper name should be identified as the antecedent. Consequently, if such a theory is true, we expected to find a reading time slowdown at the reflexive if the gender of the reflexive and the proper name mismatched. Conversely, if the search for the reflexive antecedent is sensitive to grammatical configuration, we expected to see a reading time slowdown just in case the gender of the reflexive mismatched with that of the grammatically licit antecedent, the *wh*-phrase.

Results support the hypothesis that the parser is sensitive to global structural information during the reflexive antecedent search process. At or immediately after the reflexive, conditions in which the gender of the reflexive mismatched with that of the *wh*-phrase were read slower than those conditions in which the genders matched. At the spillover region the same effect of reflexive gender congruence with the *wh*-phrase was found, with the match condition read faster than the mismatch condition, in the regression path measure. We conclude that the parser attempts to form the RD with the grammatically accessible

antecedent, the *wh*-phrase, and not with the grammatically inaccessible antecedent, consistent with the findings of Sturt (2003). This is despite the fact that in the configuration in question the grammatically inaccessible antecedent is linearly closer to the reflexive.

Furthermore, in Experiment 1, the string including the grammatically inaccessible antecedent and the reflexive is locally coherent if the initial *wh*-phrase is disregarded as indicated in (29-a). Theories of sentence processing where the parser builds the structure based on the information available within a linearly local span (e.g., Ferreira et al., 2002; Tabor et al., 2004; Konieczny et al., 2010) would predict that the parser could be subject to confusion in these contexts and select the linearly closer candidate antecedent, but we do not find evidence for this behavior. Given the experimental support for the existence of local coherence effects of this kind elsewhere, a possible explanation for why they do not occur in this context is that the parser operates over a representation containing the unpronounced tail of the WhFGD after the linearly local candidate antecedent. If the RD formation process operates over such a representation, local coherence effects may be blocked here because the true closest candidate antecedent is the tail of the WhFGD, i.e., there is no actual locally coherent substring in the examples because the WhFGD tail disrupts the potential local coherency (29-b).

(29) a. …did Mary/David expect to have injured herself/himself…
b. …did Mary/David expect /gap/$_i$ to have injured herself/himself…

In addition to the main effect of *wh*-phrase gender congruence found in Experiment 1, marginal interactions of *wh*-phrase gender congruence with local NP congruence were found in the regression path duration and re-read time in the critical region, such that the mismatch-mismatch conditions were read more slowly. This interaction, while not statistically significant, could be consistent with the predictions of an unconstrained cue-based model of antecedent retrieval under which the parser attempts to associate the reflexive with all possible candidate antecedents and experiences extra difficulty when no gender-congruent antecedent is found in memory.

Another interpretation for this interaction relies upon the observation that the examples used in Experiment 1 have another, less easily accessible parse in which the reflexive receives a non-argument, intensifier interpretation. Such a parse can be paraphrased with the intensifier reflexive located in another position possible for such intensifiers: *Which cowgirl did Mary herself expect to have injured due to negligence?*. It is possible that the marginal interaction with the local NP is due to the parser considering this alternative parse. Experiment 2 is an attempt to distinguish these explanations by testing configurations in which this alternative parse is unavailable.

Experiment 2 examines the reading time course of examples from the example paradigm in (21). These examples are similar to those used in Experiment 1, with the exception that the embedded clause is finite. The consequence of this change is that the examples are no longer locally ambiguous in the substring subsequent to the *wh*-phrase. However, as in Experiment 1, these examples include a grammatically accessible antecedent for the reflexive, the *wh*-phrase, and a grammatically inaccessible, but linearly closer possible antecedent, the proper name. Consequently, the gender mismatch manipulation yields the same two sets of predictions in this experiment: if the search for the reflexive antecedent is structure-insensitive, we would expect to see a gender mismatch effect on the linearly closer but grammatically inaccessible antecedent, the local NP. Conversely, if the reflexive antecedent search is sensitive to grammatical structure, we would expect to see the gender mismatch effect on the grammatically accessible but linearly further *wh*-phrase.

The results again support the hypothesis that the parser only considers the grammatically accessible *wh*-phrase when attempting to identify the reflexive antecedent. On the critical region, in regression path duration and re-read times, we saw a main effect of gender congruence with the *wh*-phrase, with the match conditions read faster than the mismatch conditions. In the spillover region, we see the same effect in first pass, regression path and re-read times. Additionally, here we failed to see any effect, even marginal, of the grammatically inaccessible antecedent. As this alternative parse is unavailable for the stimuli used in Experiment 2, this suggests that the marginal interaction with the local NP in Experiment 1 may indeed have been due to the alternative *intensifier reflexive* parse discussed above, rather than being evidence for a cue-based retrieval system experiencing difficulty in the absence of a gender-congruent candidate antecedent.

In Experiments 1 and 2, we saw that the parser considered just the grammatically licit anaphor antecedent, despite the presence of another possible antecedent intervening between the grammatically licit *wh*-phrase antecedent and the anaphor. One may wonder, however, whether these results are the result of the *wh*-phrases having a special status in working memory, or having a particularly high prominence relative to other potential antecedents (Martin and McElree, 2011). If this were the case, the results from Experiments 1 and 2 might simply be the result of this high prominence; the parser attended to the *wh*-phrase as a potential antecedent not because it was a grammatically licit antecedent and the local NP an illicit antecedent, but rather because the *wh*-phrase was the most prominent possible antecedent.

Experiments 3 and 4 were designed to test this alternative hypothesis through the examination of the reading time-course of examples from the paradigms illustrated in (25) and (27). In these examples, the *wh*-phrase is no longer a grammatically accessible antecedent for the reflexive. Instead, the local NP serves as the sole grammatically licit antecedent. Thus, if the parser considered the *wh*-phrase as an antecedent regardless of whether it is a grammatically licit antecedent, we would expect that a mismatch in gender between *wh*-phrase and anaphor would induce a slowdown in reading times.

The results of Experiments 3 and 4 do not support this alternative hypothesis. In both experiments, the conditions in which the local NP mismatched in gender with the reflexive

were read slower than the conditions in which the local NP and reflexive matched in gender. In Experiment 3, this effect was found on the critical region in the first fixation, first pass, regression path, and re-read durations. In Experiment 4, the effect was found on the critical region in the regression path reading times.

The combination of Experiments 1 and 3 shows that merely having a *wh*-NP present in a sentence does not automatically cause it to be considered as the antecedent of a reflexive—that is, a general notion of the prominence of a potential antecedent, such that a *wh*-NP is checked as a potential antecedent of any dependency encountered in later processing, is insufficient to explain the observed pattern of results.

If the prominence of a potential antecedent were the source of the interference effects, we would expect to observe the same pattern in Experiments 1 and 3. But instead the observed pattern is that, when the tail of the WhFGD intervenes between the linearly local embedded subject NP and the reflexive, the *wh*-NP's gender congruency with the reflexive modulates the presence or absence of gender mismatch effects, whereas when the the *wh*-NP is not associated with a tail intervening between the subject and reflexive, the linearly local embedded subject NP's gender congruency with the reflexive modulates the presence or absence of gender mismatch effects. Thus, while it is plausibly true that *wh*-NPs are highly "prominent" candidate antecedents, the parser still appears to be guided by syntactic structure in the course of reflexive resolution and is not "confused" by the presence of an irrelevant *wh*-NP.

The principal significance of this set of findings is to provide evidence for quite sophisticated structure-sensitivity on the part of the antecedent retrieval system. Whatever mechanism subserves reflexive antecedent retrieval, whether cue-based or otherwise, must be able to exhibit online sensitivity to fine-grained syntactic structure of at least two kinds. First, it must be able to respect the clausemate condition on anaphora: that reflexives are unable to find their antecedent outside of their immediate clause. Second, it must be sensitive to the presence and location of WhFGD tails: the presence of a WhFGD must be visible to the antecedent retrieval system, whether in the form of reactivation of a previously processed NP upon gap-detection or via the positing of gaps/dependency tails as candidate antecedents in the representation over which reflexive antecedent retrieval operates. For this reason, this study constitutes evidence against unrestricted versions of cue-based retrieval, and in favor of models like that in Dillon et al. (2013) that constrain the antecedent retrieval process to respect syntactic structure.

Why might the reflexive antecedent retrieval system fail to experience interference from grammatically inaccessible antecedents in WhFGD contexts, while showing evidence of such interference in reflexive antecedent retrieval mediated by the related long-distance dependencies of raising and control studied in Kwon and Sturt (2015) and Sturt and Kwon (2015)? We speculate that the active nature of the WhFGD formation process may provide an explanation for this difference. Encountering a *wh*-NP triggers the parser to initiate an active search for its corresponding gap site, while control and raising dependencies cannot be identified until later in a sentence. If active search behavior on the part of the parser involves positing a WhFGD dependency tail within the local domain of the reflexive, this element may be an accessible retrieval candidate for a syntactically constrained antecedent retrieval system. Future research could address this question by investigating reflexive antecedent retrieval in the context of other long-distance dependencies whose leftmost element is a strong cue to the existence of the long-distance dependency, perhaps topicalization.

4. Conclusion

In this study we have investigated whether the process of reflexive-antecedent resolution is sensitive, in on-line measures, to the presence of a WhFGD dependency whose tail is the grammatically licit antecedent of the reflexive. The fact that Experiments 1 and 2 found gender mismatch effects between the *wh*-NP and the reflexive, and not between a linearly local NP and the reflexive, strongly supports the position that the tail of a WhFGD can be accessed rapidly online as a candidate antecedent in reflexive antecedent search.

This effect of the *wh*-NP is not compatible with an account where *wh*-NPs are simply highly prominent candidate antecedents regardless of the grammatical possibility of such an antecedent-reflexive relationship because, if this were the explanation for the effect of gender mismatch between *wh*-NP and reflexive in Experiments 1 and 2, Experiments 3 and 4 should have shown the same pattern. Instead, in Experiments 3 and 4, gender mismatch effects of the linearly local non-*wh*-NP and the reflexive were observed, consistent with an account on which the parser's reflexive antecedent search is grammatically guided. In general, then, we conclude that the parser's reflexive antecedent search is rapidly sensitive to such fine-grained syntactic details as the presence and location of a WhFGD. We take this to be evidence that, whatever mechanism is implicated in reflexive antecedent retrieval, it must be able to exhibit online sensitivity to the binding constraints and to treat the tail of a WhFGD as a potential candidate antecedent.

Acknowledgments

We would like to thank two reviewers for their detailed and helpful discussions. We are in debt to the following colleagues for their helpful discussion: Brian Dillon, Dave Kush, Nina Kazanina, Colin Phillips, Martin Pickering, Lars Konieczny, Patrick Sturt, and Whitney Tabor. Thanks also to the audiences of AMLaP2013, AMLaP2014, CUNY2014, and CUNY2015. This work has been supported in part by NSF grant BCS-1323245 and DDRIG:1348677. All the remaining errors are, of course, our own.

Supplementary Material

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpsyg.2015.01504

References

Badecker, W., and Straub, K. (2002). The processing role of structural constraints on interpretation of pronouns and anaphors. *J. Exp. Psychol.* 28:748. doi: 10.1037/0278-7393.28.4.748

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: keep it maximal. *J. Mem. Lang.* 68, 255–278. doi: 10.1016/j.jml.2012.11.001

Bates, D., Kliegl, R., Vasishth, S., and Baayen, H. (2015). Parsimonious mixed models. *arXiv*. arXiv:1506.04967

Chomsky, N. (1981). *Lectures on Government and Binding*. Dordrecht: Foris.

Clackson, K., and Heyer, V. (2014). Reflexive anaphor resolution in spoken language comprehension: structural constraints and beyond. *Front. Psychol.* 5:904. doi: 10.3389/fpsyg.2014.00904

Cunnings, I., and Felser, C. (2013). The role of working memory in the processing of reflexives. *Lang. Cogn. Process.* 28, 188–219. doi: 10.1080/01690965.2010.548391

Cunnings, I., Patterson, C., and Felser, C. (2015). Structural constraints on pronoun binding and coreference: evidence from eye movements during reading. *Front. Psychol.* 6:840. doi: 10.3389/fpsyg.2015.00840

Cunnings, I., and Sturt, P. (2014). Coargumenthood and the processing of reflexives. *J. Mem. Lang.* 75, 117–139. doi: 10.1016/j.jml.2014.05.006

Dillon, B., Mishler, A., Sloggett, S., and Phillips, C. (2013). Contrasting intrusion profiles for agreement and anaphora: experimental and modeling evidence. *J. Mem. Lang.* 69, 85–103. doi: 10.1016/j.jml.2013.04.003

Felser, C., Sato, M., and Bertenshaw, N. (2009). The on-line application of Binding Principle A in English as a second language. *Bilingualism* 12, 485–502. doi: 10.1017/S1366728909990228

Ferreira, F., Bailey, K. G., and Ferraro, V. (2002). Good-enough representations in language comprehension. *Curr. Dir. Psychol. Sci.* 11, 11–15. doi: 10.1111/1467-8721.00158

Frazier, L. (1987). Syntactic processing: evidence from Dutch. *Nat. Lang. Linguist. Theory* 5, 519–559.

Jäger, L. A., Engelmann, F., and Vasishth, S. (2015). Retrieval interference in reflexive processing: experimental evidence from Mandarin, and computational modeling. *Front. Psychol.* 6:617. doi: 10.3389/fpsyg.2015.00617

Konieczny, L., Weldle, H., Wolfer, S., Müller, D., and Baumann, P. (2010). "Anaphora and local coherences," in *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, eds S. Ohlsson, and R. Catrambone (Austin, TX: Cognitive Science Society), 1204–1209.

Kush, D., Lidz, J., and Phillips, C. (2015). Relation-sensitive retrieval: evidence from bound variable pronouns. *J. Mem. Lang.* 82, 18–40. doi: 10.1016/j.jml.2015.02.003

Kwon, N., and Sturt, P. (2015). Processing control information in a nominal control construction: an eye-tracking study. *J. Psycholinguist. Res.* doi: 10.1007/s10936-015-9374-2. [Epub ahead of print].

Lee, M.-W. (2004). Another look at the role of empty categories in sentence processing (and grammar). *J. Psycholinguist. Res.* 33, 51–73. doi: 10.1023/B:JOPR.0000010514.50468.30

Lewis, R. L., and Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cogn. Sci.* 29, 375–419. doi: 10.1207/s15516709cog0000_25

Martin, A. E., and McElree, B. (2011). Direct-access retrieval during sentence comprehension: evidence from sluicing. *J. Mem. Lang.* 64, 327–343. doi: 10.1016/j.jml.2010.12.006

Nicol, J., and Swinney, D. (1989). The role of structure in coreference assignment during sentence comprehension. *J. Psycholinguist. Res.* 18, 5–19.

Omaki, A., Lau, E. F., Davidson White, I., Dakan, M. L., Apple, A., and Phillips, C. (2015). Hyper-active gap filling. *Front. Psychol.* 6:384. doi: 10.3389/fpsyg.2015.00384

Phillips, C. (2006). The real-time status of island phenomena. *Language* 82, 795–823. doi: 10.1353/lan.2006.0217

Reinhart, T. M. (1976). *The Syntactic Domain of Anaphora*. Ph.D. thesis, Massachusetts Institute of Technology.

Stowe, L. A. (1986). Parsing wh-constructions: evidence for on-line gap location. *Lang. Cogn. Process.* 1, 227–245.

Sturt, P. (2003). The time-course of the application of binding constraints in reference resolution. *J. Mem. Lang.* 48, 542–562. doi: 10.1016/S0749-596X(02)00536-3

Sturt, P., and Kwon, N. (2015). The processing of raising and nominal control: an eye-tracking study. *Front. Psychol.* 6:331. doi: 10.3389/fpsyg.2015.00331

Tabor, W., Galantucci, B., and Richardson, D. (2004). Effects of merely local syntactic coherence on sentence processing. *J. Mem. Lang.* 50, 355–370. doi: 10.1016/j.jml.2004.01.001

Traxler, M. J., and Pickering, M. J. (1996). Plausibility and the processing of unbounded dependencies: an eye-tracking study. *J. Mem. Lang.* 35, 454–475.

Xiang, M., Dillon, B., and Phillips, C. (2009). Illusory licensing effects across dependency types: ERP evidence. *Brain Lang.* 108, 40–55. doi: 10.1016/j.bandl.2008.10.002

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2015 Frazier, Ackerman, Baumann, Potter and Yoshida. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 29 July 2014
doi: 10.3389/fpsyg.2014.00818

A fan effect in anaphor processing: effects of multiple distractors

Kevin S. Autry* and William H. Levine*

Department of Psychological Science, University of Arkansas, Fayetteville, AR, USA

Edited by:
Colin Phillips, University of Maryland, USA

Reviewed by:
Claudia Felser, University of Potsdam, Germany
Dan Parker, University of Maryland, USA

****Correspondence:***
Kevin S. Autry and William H. Levine, Department of Psychological Science, University of Arkansas, Memorial Hall 216, Fayetteville, AR 72701, USA
e-mail: ksautry@gmail.com; whlevine@uark.edu

Research suggests that the presence of a non-referent from the same category as the referent interferes with anaphor resolution. In five experiments, the hypothesis that multiple non-referents would produce a cumulative interference effect (i.e., a fan effect) was examined. This hypothesis was supported in Experiments 1A and 1B, with subjects being less accurate and slower to recognize referents (1A) and non-referents (1B) as the number of potential referents increased from two to five. Surprisingly, the number of potential referents led to a decrease in anaphor reading times. The results of Experiments 2A and 2B replicated the probe-recognition results in a completely within-subjects design and ruled out the possibility that a speeded-reading strategy led to the fan-effect findings. The results of Experiment 3 provided evidence that subjects were resolving the anaphors. These results suggest that multiple non-referents do produce a cumulative interference effect; however, additional research is necessary to explore the effect on anaphor reading times.

Keywords: comprehension, memory, fan effect, reading, anaphor resolution, antecedent, distractor

INTRODUCTION

Many theorists have argued that language comprehension processes can be explained in large part by appealing to general memory processes (e.g., Lewis, 1996; Gerrig and McKoon, 1998; Myers and O'Brien, 1998; Lewis and Vasishth, 2005; van den Broek et al., 2005); this hypothesis has been widely supported by empirical evidence. For example, general theories of memory processes have been shown to provide explanations for linguistic tasks such as establishing common ground between multiple parties (Horton and Gerrig, 2005) and resolving anaphors (O'Brien et al., 1990; Almor, 1999). Anaphor comprehension (often called anaphor resolution) in particular appears to rely heavily upon memory to determine co-reference between an anaphor and antecedent. Even within a sentence, limitations on working memory capacity induce the need for retrieval of referents (McElree, 2000). There are also instances, such as pronouns that refer to implicit referents (Greene et al., 1994) and bridging inferences (Garrod and Sanford, 1981), where anaphors are resolved even though the intended referent has not been explicitly mentioned. Such processes clearly rely on memory to produce an acceptable referent. Further evidence for the relationship between memory and anaphor resolution is provided by the findings that many factors affecting memory also affect anaphor resolution, including distance and elaboration (O'Brien et al., 1990), salience of the anaphor (Klin et al., 2004), salience of the referent (Foraker and McElree, 2007), and frequency (van Gompel and Majid, 2004). In the research reported here, we focus on anaphor resolution across sentences. Nevertheless, models of retrieval processes both across (Myers and O'Brien, 1998) and within (e.g., Lewis and Vasishth, 2005) sentences have many commonalities, which we highlight below.

Of particular interest for the current research are studies that have examined the effects of multiple potential referents on anaphor resolution (e.g., Corbett and Chang, 1983; Corbett, 1984; Mason, 1997; Levine et al., 2000; Wiley et al., 2001; Badecker and Straub, 2002; Klin et al., 2004, 2006; Ditman et al., 2007; Levine and Hagaman, 2008). In one of the first studies examining the effect of multiple potential referents, Corbett found longer reading time for an anaphoric noun phrase (e.g., *the frozen vegetable*) that included a category label when a text contained two members of that category (e.g., *fresh corn and frozen asparagus*) than when there was only a single category member (e.g., *frozen asparagus*). Badecker and Straub similarly found an increase in reading time shortly after subjects read reflexives when multiple gender-matched referents had been mentioned (e.g., *John thought that Bill owed himself another opportunity to solve the problem*). Levine et al. (see also Klin et al., 2004, 2006) found evidence suggesting that under some conditions anaphors (e.g., *the dessert*) appear not to be resolved at all when a text contains two potential referents from the same category (e.g., *tart* and *cake*), likely due to the increased difficulty in identifying a unique referent. The increased difficulty in processing anaphors in these studies suggests that readers engage in additional processing when a distractor (i.e., a non-referent) is present. Presumably this occurs because the both nouns are considered as potential referents, a process that is initiated by simple memory matching and that leads to retrieval-based interference. This explanation follows straightforwardly from global memory models (e.g., Ratcliff, 1978; Gillund and Shiffrin, 1984; Hintzman, 1986), which assume that stored memory representations that are related to a memory cue are activated in parallel and to the degree that they share features with the memory cue. Somewhat surprisingly, this

additional processing appears to occur regardless of disambiguating material that should identify the proper referent, such as a prenominal adjective like *frozen* or the grammatical constraints that govern interpretation of reflexives (e.g., Reinhart, 1983). The reliability and time course of distractor interference, especially for within-sentence retrieval, is a matter of debate. Recent evidence is consistent with a very early role for grammatical constraints in retrieval. For example, Chow et al. (2014) were unable to replicate Badecker and Straub's results, and they found evidence that grammatical constraints prevent distractor interference (see also Dillon et al., 2013). Across sentence boundaries, some features, such as parallel structure (e.g., *Josh criticized Paul. Then Marie insulted him.*), may play an early role in limiting referent search (Chambers and Smyth, 1998). Nevertheless, for definite noun-phrase anaphors like *the dessert*, reported findings suggest that retrieval processes rely on semantic matching between an anaphor and potential referents, with no evidence as yet indicating that there are grammatical constraints on this process.

Whereas results like those from Badecker and Straub (2002), Corbett (1984), and Klin and colleagues (Levine et al., 2000; Klin et al., 2004, 2006) illustrate indirectly that distractors are considered during anaphor resolution, direct evidence that distractors are activated during anaphor resolution comes from results reported by O'Brien et al. (1990). O'Brien et al. had subjects read passages with two potential antecedents (e.g., *train* and *plane*), which appeared early and late in a passage and were sometimes described elaborately. At the end of a passage, a sentence (e.g., *Mark's neighbor asked him how he had traveled to his parent's*) appeared that required retrieval of only one of the antecedents. Following this sentence, subjects had to name aloud one of the potential antecedent nouns. Relative to a no-anaphor control condition, referent nouns were named more quickly, replicating findings that suggest that referents are activated by anaphor resolution processes (e.g., Dell et al., 1983). Of perhaps greater interest was the finding that non-referent concepts were also activated relative to a control condition, especially when they were elaborated and appeared in the late position in the passage, between the anaphor and the correct antecedent. These results are consistent with the hypothesis that an anaphor acts like any other memory cue, activating related information in parallel. The finding that non-referent concepts were activated, especially when they occurred late and were elaborated, again fits very well with well-established findings from the memory literature that recency and elaboration lead to easier memory access.

Taken together, these studies demonstrate that people consider multiple potential referents when resolving anaphors, and further, that the resolution of the anaphor increases activation for the referent. However, studies involving distractors have typically been limited to situations with a single distractor. Therefore, the effect of additional distractors remains an open empirical question. A yet-stronger case that general memory processes govern anaphor resolution can be made if there is a cumulative effect of additional distractors. Both Myers and O'Brien's (1998) resonance model and Lewis and Vasishth's (2005) implementation of ACT-R (e.g., Anderson, 2005) as a theory of memory-retrieval in sentence-processing make similar predictions about the effect of multiple distractors. The resonance model states that elements in the mental representation resonate to signals from retrieval cues. In the case of anaphor resolution, the retrieval cue is the anaphor and the resonating elements are related items in the mental representation. Critically, the signal (i.e., resonance strength) of any item in the representation is divided among receiving elements, and only a subset of the elements with the strongest signal enter working memory (WM). Thus, the strength of a referent will be reduced in the presence of related distractors, reducing the probability that the correct referent will be selected into WM. Similarly, Lewis and Vasishth's model states that the activation that a chunk in memory will receive is reduced as there are more chunks in memory associated with a particular cue. Given the assumption that activation determines retrieval latency and the probability of the retrieval of a memory chunk, there should be greater difficulty in retrieving the correct referent with every additional distractor.

We can also draw on the memory literature to provide empirical guidance about the possible effects of multiple distractors. Specifically, research has shown that reading sentences that pair a person with multiple locations (or a location with multiple people) slows later recognition of the sentences (Anderson, 1974; Radvansky, 1998; Anderson and Reder, 1999). This result, known as the fan effect, is hypothesized to occur because of interference among competing associations in memory. Unlike the anaphor literature, which has focused on single distractors, the fan effect literature has explored situations with more than two associations and has demonstrated a cumulative effect, such that additional associations cause additional interference.

In the original demonstration of the fan effect (Anderson, 1974), subjects studied sentences in which a person was paired with a location (see 1–4 below).

(1) A hippie is in the park.
(2) A hippie is in the church.
(3) A policeman is in the park.
(4) A sailor is in the park.

Importantly, some people were associated with more than one location and some locations were associated with more than one person. For example, the sailor was associated only with the park (i.e., a fan of one), the hippie was associated with both the park and the church (i.e., a fan of two), and the park was associated with hippie, the policeman, and the sailor (i.e., a fan of three). Thus, the nouns varied in the number of associations with other nouns. After the study phase, subjects read another set of sentences, some of which were the same as those studied previously and some of which were novel pairings of people and locations that the subjects had not seen. For each sentence, subjects indicated whether it was the same as one they had read during the study phase or not. Consistent with the hypothesis that multiple associations interfere with one another, subjects were slower to recognize sentences with nouns that were associated with more nouns compared to sentences with nouns associated with fewer nouns. That is, subjects were slower to respond as the size of the noun's fan increased.

If anaphor resolution relies on general memory processes, and increasing the number of associations with a noun increases

interference, then we can predict that increasing the total number of potential referents for an anaphor should also show a cumulative retrieval-interference effect (i.e., a fan effect). The present study tested this prediction across five experiments by exploring the effects of multiple distractors on anaphor resolution and the subsequent activation levels of referents and distractors. In particular, we used a probe recognition task after anaphor sentences to measure the relative activation of an anaphoric referent when there were a variable number of distractors. We also used the probe task to measure activation of those distractors as a function of the number of distractors. Our results demonstrate evidence of a fan effect in anaphor resolution.

EXPERIMENT 1A

In Experiment 1A, subjects read pairs of sentences. The first provided an antecedent and one or more distractors in a serial list, and the second included an anaphoric noun phrase that co-referred with the antecedent; these were followed by a probe recognition task that was used to measure the activation of the referent concept (see **Table 1** for a sample passage and Appendix A in Supplementary Materials for a full list of experimental passages). In particular, the first sentence ended with a list of two, three, four, or five potential referents from the same taxonomic category, and the second sentence referred with a disambiguating adjective and categorical anaphor to a single item mentioned in the list. Following each sentence-pair, subjects completed a probe recognition task to measure the activation level of the referent following the anaphor. For example, the first sentence in the example in **Table 1** describes a person looking through a toolbox with a number of tools in it. The last tool mentioned in the sentence, a saw, is the antecedent concept. The second sentence then describes the person fixing a table using *the cutting tool*. The latter noun phrase serves as an unambiguous reference to the entity introduced by the antecedent. After the second sentence was completed, the word *saw* was presented in an old-new recognition task, the correct response for which is "old." We assume that reaction time and accuracy in responding to the probes will reflect the ease or difficulty the subjects have in selecting the correct referent (cf. Dell et al., 1983; Levine et al., 2000) from the list of potential referents, including the distractors and the referent.

Table 1 | Sample passage.

List sentence	Amelia's new table was wobbling, so she looked in her toolbox and found . . .
Two-noun	. . . a hammer and a saw. (all experiments)
Three-noun	. . . a screwdriver, a hammer, and a saw. (Experiments 1A and 1B only)
Four-noun	. . . a level, a screwdriver, a hammer, and a saw. (Experiments 1A and 1B only)
Five-noun	. . . a wrench, a level, a screwdriver, a hammer, and a saw. (all experiments)
REFERENCE SENTENCE	
Anaphor	She fixed it with the cutting tool before it broke. (all experiments)
No anaphor	She fixed the table all by herself before it broke. (Experiment 3 only)
PROBE WORD	
Referent	SAW (Experiments 1A, 2A, 2B, and 3)
Distractor	HAMMER (Experiments 1B, 2A, and 2B)
Comprehension question	Did Amelia use the saw? (all experiments)

We hypothesized that increasing the number of distractors would lead activation from the anaphor to spread among the referent and distractor concepts (Kintsch, 1988; Myers and O'Brien, 1998; Lewis and Vasishth, 2005). It was expected that the spread of activation from the anaphor to all conceptually-related potential referents would cause the referent to be less active following anaphor resolution as the number of distractors increased (i.e., a monotonic increasing trend in reaction time and decreasing trend in accuracy was expected), resulting in lower probe accuracy and longer probe recognition times. Additionally, this spread of activation should interfere with the selection of the appropriate referent during anaphor resolution, thus slowing reading of the reference sentence, replicating several findings (e.g., Corbett and Chang, 1983; Corbett, 1984; Mason, 1997; Badecker and Straub, 2002). Alternatively, it is possible that a backward, parallel-search process occurs such that the earlier-occurring distractors have little or no detectable impact on anaphor resolution (O'Brien, 1987). A backward, serial, self-terminating search would also predict no impact of early distractors on resolution of later referents. This latter strategy seems attractive especially in short passages with a list-like first sentence (cf. Townsend and Fific, 2004).

METHOD

Subjects

Ninety-five students enrolled in a general psychology course at the University of Arkansas participated in the experiment to partially fulfill a research requirement. All subjects were native-English speakers. Informed consent was obtained from all subjects in this and all subsequent experiments.

Materials and design

There were 31[1] experimental sentence-pairs that appeared in one of four conditions (see **Table 1**). Each sentence-pair began with a list sentence that introduced a character by proper name (half stereotypically male, half stereotypically female) and ended in a list of either two, three, four, or five nouns from the same taxonomic category. The nouns were all single words, common, and were selected to be roughly equal in typicality as judged by the first author and several research assistants. Furthermore, each of the last two nouns in the list was able to be distinguished from the other nouns by means of an adjective (e.g., saws can be distinguished from the other tools in the list using the adjective *cutting*). The list sentence was followed by a reference sentence that unambiguously referred to the final item in the list using an adjective and a categorical anaphor (e.g., *cutting tool*) that was the same for all conditions. The anaphor always occurred three words prior to the end of the reference sentence to ensure that there was enough time for the anaphor to be resolved by the time the sentence was fully read (i.e., by the time the probe-word task was presented).

[1]Experimenter error resulted in an odd number of experimental items in this experiment and in Experiment 1b.

In addition, there were 68 filler sentence-pairs that each included a list sentence with two to five nouns but that were not limited by the same restrictions on nouns in the experimental lists (e.g., the nouns could be proper or multiple words). As with the experimental sentence pairs, the filler reference sentences also included a categorical anaphor modified by an adjective; however, the referent of the anaphor was not always completely unambiguous. Moreover, the referent of the anaphor was not always the last item in the list. These two features of the fillers were expected to encourage subjects to put forth more effort in resolving anaphors across all trials.

Each experimental and filler sentence-pair also had a corresponding recognition probe and comprehension question. Following the reference sentence, subjects completed a probe recognition task in which they indicated whether a word on the screen had occurred in the previous sentence-pair. For experimental sentence-pairs, the probe word was always the final noun from the list, which required a "yes" response. To ensure an equal number of "yes" and "no" responses across the experiment, the majority of the filler probe tasks presented a word that did not occur in the sentence pair and therefore required a "no" response. Other fillers presented a probe word that was not the final noun from the list, requiring a "yes." Finally, a comprehension question was presented following the probe recognition task, half of which required a "yes" response and half of which required a "no" response. Comprehension questions frequently, but not always, focused on correct resolution of the anaphor (e.g., *Did Amelia use the saw?*).

Subjects saw each experimental sentence-pair in one of the four conditions along with all filler sentence-pairs. Four counterbalanced lists were created with the following constraints: one quarter of the list sentences had two nouns, one quarter had three nouns, one quarter had four nouns, and one quarter had five nouns. Furthermore, a second set of materials[2] was created that reversed the order of the final two nouns in the list, such that final noun in the first set of materials (e.g., *saw*) became the penultimate noun and the formerly penultimate noun (e.g., *hammer*) became the final noun. This also required a change in the disambiguating adjective in the reference sentence (e.g., *cutting* changed to *pounding*) such that the referent of the categorical anaphor was always the final noun. The manipulation of these factors resulted in a design that was 4 (nouns: 2, 3, 4, 5) × 2 (noun order: order 1, order 2).

Procedure

The experiment began with three practice blocks of five trials each, which were intended to familiarize the subject with the yes/no response keys, the probe recognition task, and the comprehension question, respectively. For all practice trials, feedback about the correctness of subjects' responses was provided.

Subjects then began the experimental session. Subjects were instructed to read the sentences as they normally would for comprehension and to respond to the probe words as quickly and accurately as possible. Each trial consisted of a list sentence, a reference sentence, a probe word, and a comprehension question. At the beginning of each trial, subjects were given the instruction "PRESS THE SPACEBAR WHEN READY" centered on a computer monitor. When they pressed the spacebar, the list sentence appeared left-justified in the middle of the screen. Subjects pressed the spacebar to indicate when they had finished reading the list sentence, which removed the list sentence from the screen and replaced it with the reference sentence. Subjects pressed the spacebar again to indicate when they had finished reading the reference sentence, which removed the reference sentence from the screen and replaced it with a probe word in all capital letters in the center of the screen. Subjects used the left and right arrow keys labeled "Y" and "N" for yes and no, respectively, to respond to the probe task. This removed the probe word and replaced it with a comprehension question in the center of the screen; no feedback about correctness was provided for probes or questions. Subjects again used the yes and no keys to respond to the comprehension question, which ended the trial.

The experimental session consisted of 99 trials (31 experimental and 68 fillers) in three blocks of 25 trials and one block of 24 trials. The order of the blocks, as well as the order of the trials within each block, was randomized with the restriction that the first sentence-pair of each block was always a filler sentence-pair, to allow time for the subjects to fully return their attention to the task after a mandatory 10 s break between blocks. Subjects were free to take breaks between trials. The experiment lasted approximately 30 min. The procedure for this and all subsequent experiments were approved by the University of Arkansas Institutional Review Board.

RESULTS

Data exclusion and general analytic considerations

A subject's data were excluded from further analysis if they met any of the following criteria: (1) they had more than 30% of reading times less than 1000 ms or greater than 7500 ms; (2) they had lower than 70% probe recognition accuracy; (3) they had more than 30% of probe reaction times less than 500 ms or greater than 2500 ms; (4) they had no non-outlying probe recognition observations in at least one condition; or (5) they had less than 70% comprehension question accuracy. Based on these criteria, the data from eight subjects were excluded from further analysis. Additionally, two experimental items were removed from further analysis due to counterbalancing errors. Therefore, the reported analyses include 85 subjects and 29 items.

For all experiments reported in this paper, subject and item condition means were analyzed separately; a subscript of 1 indicates that subjects were treated as a random-effects variable, whereas a subscript of 2 indicates that items were treated as a random-effects variable. For all significance tests, an alpha level of 0.05 was used. Predictions about monotonic increasing and decreasing trends were tested using polynomial contrasts. For all repeated-measures effects with more than one numerator *df*, Huynh-Feldt adjusted *p*-values are reported to correct for sphericity violations. Effect-size measures that are reported are based on the subject analyses, and all within-subject standard

[2]Probe length and frequency was similar for the two sets of materials (length: Set 1 $M = 6.5$ letters, $SD = 1.5$; Set 2 $M = 6.4$ letters, $SD = 1.7$; log frequency (Lund and Burgess, 1996; Balota et al., 2007): Set 1 $M = 7.4$, $SD = 2.4$; Set 2 $M = 7.2$, $SD = 2.9$).

errors in figures and tables were computed using the method recommended by Loftus and Masson (1994).

Comprehension

In general, the number of nouns did not affect comprehension (see **Table 2** for comprehension results across all experiments). The linear trend was non-significant, $F_{1(1,\ 84)} = 0.34$, $p = 0.56$, $F_{2(1,\ 56)} = 0.07$, $p = 0.79$, with no significant higher-order trends. (See Appendix C in Supplementary Materials for the results of the noun-order factor in this experiment and Experiment 1B.)

Probe accuracy

Figure 1 presents mean probe word accuracy and reaction times along with mean reference-sentence reading times as a function of the number of referents. In general, accuracy decreased as the number of nouns in the list sentence increased. The linear trend was significant, $F_{1(1,84)} = 9.63$, $p = 0.003$, $F_{2(1,28)} = 9.99$, $p = 0.004$, $\eta_p^2 = 0.10$, with no significant higher-order trends.

Probe reaction times (RT)

Only correct probes were analyzed. Outliers were first classified as RTs that were less than 400 ms or greater than 3000 ms. Remaining reaction times more extreme than 1.5 times the interquartile range above the 75th percentile or below the 25th percentile for each subject were classified as outliers (Tukey, 1977), resulting in 8.6% of the data being excluded from further analyses. In general, reaction time increased as the number of nouns in the list sentence increased (see **Figure 1**). The linear trend was significant, $F_{1(1,\ 84)} = 8.03$, $p = 0.006$, $F_{2(1,28)} = 6.68$, $p = 0.02$, $\eta_p^2 = 0.09$, with no significant higher-order trends.

Table 2 | Mean comprehension for all experiments (with standard errors of the mean).

	Two-noun	Three-noun	Four-noun	Five-noun
Experiment 1A	0.88 (0.014)	0.86 (0.014)	0.87 (0.014)	0.87 (0.014)
Experiment 1B	0.89 (0.014)	0.87 (0.015)	0.83 (0.017)	0.86 (0.016)
Experiment 2A				
Referent	0.95 (0.012)	–	–	0.92 (0.015)
Distractor	0.93 (0.012)	–	–	0.91 (0.014)
Experiment 2B				
Referent	0.93 (0.013)	–	–	0.88 (0.021)
Distractor	0.92 (0.015)	–	–	0.91 (0.015)
Experiment 3				
Anaphor	0.88 (0.019)	–	–	0.90 (0.014)
No anaphor	0.96 (0.011)	–	–	0.91 (0.015)

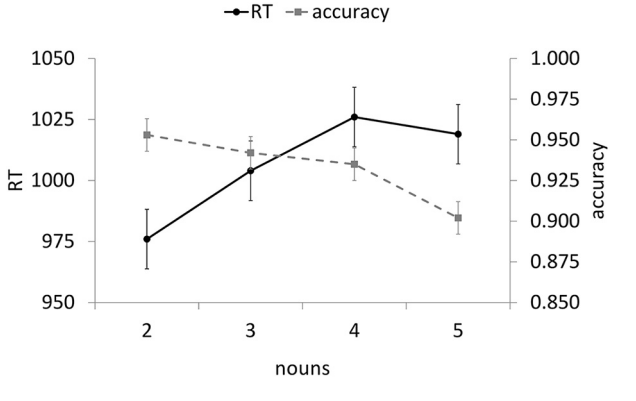


FIGURE 1 | Experiment 1A antecedent probe reaction times and accuracies by noun condition (error bars indicate SE of the mean).

Reference-sentence reading times

Reference-sentence reading times were transformed to per-character reading times by dividing the full-sentence reading time by the number of characters in the sentence, not counting spaces and punctuation (see **Table 3**). Outliers were first identified as trials with less than 15 ms/char or more than 150 ms/char. Outliers among the remaining reading times were then identified within each subject based on Tukey's (1977) criteria. 7.6% of the trials were excluded from further analysis. In general, reading time on the reference sentence *decreased* as the number of nouns in the list sentence increased. The linear trend was significant, $F_{1(1,84)} = 19.55$, $p < 0.001$, $F_{2(1,30)} = 10.87$, $p = 0.003$, $\eta_p^2 = 0.19$, with no significant higher-order trends.

DISCUSSION

The results of the probe word analyses were consistent with the fan-effect hypothesis and generally favor models of anaphor resolution that posit a parallel-search mechanism in retrieval of the correct referent. As predicted, the presence of distractors interfered with the probe recognition task. Increasing the number of distractors in the list sentence decreased recognition accuracy and increased reaction times for referents, which suggests that the activation level of referents decreased as the number of distractors increased. The existing literature has shown via a variety of measures and paradigms that the presence of one distractor interferes with anaphor resolution (e.g., Corbett and Chang, 1983; Corbett, 1984; Mason, 1997; Levine et al., 2000; Wiley et al., 2001; Klin et al., 2004, 2006; Ditman et al., 2007; Levine and Hagaman, 2008). The present results extend this finding by demonstrating a cumulative effect of distractors.

The effect of additional nouns on the subsequent reference-sentence reading times, however, was unexpected. It was predicted, based on previous research (e.g., Corbett, 1984), that anaphor resolution would be slowed by the presence of distractors, resulting in longer full-sentence reading times as the number of distractors increased. However, the results were exactly the opposite, indicating that the subjects actually read the reference sentences more quickly as the number of distractors

Table 3 | Experiment 1A mean per-character reading times in ms (with standard errors of the mean).

	List sentence	Reference sentence
Two-noun	73.2 (0.9)	58.3 (0.7)
Three-noun	74.0 (0.9)	57.5 (0.7)
Four-noun	76.8 (0.9)	55.4 (0.7)
Five-noun	77.4 (0.9)	53.9 (0.7)

increased. Assuming this is not a Type I error, one plausible explanation for this result is that subjects adopted a strategy of speeding through the reference sentence to reduce the time between the referents and the probe recognition task when there were more distractors. A similar finding was reported by Van Dyke and McElree (2006), who had subjects reading sentences of variable complexity while holding or not holding a memory load and found that reading was faster for more-complex sentences with a memory load than without one. This speeded-reading strategy as a potential alternative explanation for the fan effect was explored in further detail in Experiments 2A and 2B; we defer discussion until the presentation of those experiments.

EXPERIMENT 1B

Experiment 1A established that referents were less active following anaphor resolution when there were more potential referents available in the discourse. Experiment 1B replicated Experiment 1A but used distractors as the probe words to test the effect of multiple distractors on the activation level of a distractor. As in Experiment 1A, it was hypothesized that additional distractors would decrease probe accuracy and slow probe recognition times. If anaphors act like any other cue to memory, the activation of the referent and distractors should be split (Kintsch, 1988; Myers and O'Brien, 1998; Lewis and Vasishth, 2005), resulting in less activation to go around (i.e., a fan effect) as there are more related concepts in the list sentence. Because the anaphor contains two cues (i.e., adjective plus noun) to retrieve the referent but only one (i.e., the noun) that matches the distractors, referents should become more active and experience less interference (i.e., a reduced fan effect) than distractors following anaphor resolution. Moreover, later items may overwrite or displace earlier items, leading to degraded representations of the referent and especially earlier-occurring distractors (Nairne, 1990; Lewis, 1996). We examine these predictions in a cross-experiment comparison after presenting the results of Experiment 1B, and then examine them more directly (i.e., in a completely within-subjects design) in Experiments 2A and 2B.

METHOD

Subjects

Seventy-eight students enrolled in a general psychology course at the University of Arkansas participated in the experiment to partially fulfill a research requirement. All subjects were native-English speakers.

Materials, design, and procedure

Experiment 1B was identical to Experiment 1A except that the probe words in the probe recognition task for experimental trials were distractors (i.e., the penultimate word in the list).

RESULTS

Data exclusion and general analytic considerations

Based on the data exclusion criteria detailed in Experiment 1A, the data from eight subjects were excluded from further analysis. Sixteen more subjects were removed from further analysis for a systematic misunderstanding of the instructions. These subjects consistently responded "no" to distractors on the probe task when they should have been responding "yes." This pattern of responding suggests that these subjects were correctly identifying the correct referent of the anaphor, but misunderstanding that this identification was unrelated to the probe task. Therefore, the comprehension accuracy, probe accuracy, and reading time analyses included 54 subjects and 31 items.

Comprehension

In general, comprehension (see **Table 2**) decreased as the number of nouns increased. The linear trend was significant in the subject analysis, $F_{1(1,\ 53)} = 5.36$, $p = 0.025$, $\eta_p^2 = 0.09$, but non-significant in the items analysis, $F_{2(1,\ 60)} = 2.91$, $p = 0.093$, with no significant higher-order trends.

Probe accuracy

Figure 2 presents mean probe word accuracy and reaction times along with mean reference-sentence reading times as a function of the number of referents. In general, accuracy decreased as the number of nouns in the list sentence increased. The linear trend was significant, $F_{1(1,\ 53)} = 39.08$, $p < 0.001$, $F_{2(1,\ 30)} = 45.28$, $p < 0.001$, $\eta_p^2 = 0.42$, with no significant higher-order trends.

Probe reaction times

Based on outlier exclusion criteria, 9.6% of the data were excluded from further analyses. In general, reaction time increased as the number of nouns in the list sentence increased (see **Figure 2**). The linear trend was significant, $F_{1(1,\ 53)} = 16.79$, $p < 0.001$, $F_{2(1,\ 30)} = 6.59$, $p = 0.01$, $\eta_p^2 = 0.24$. There was also an unexpected cubic trend, $F_{1(1,\ 53)} = 12.81$, $p = 0.001$, $F_{2(1,\ 30)} = 3.13$, $p = 0.09$. There was no theoretical expectation of this effect, and it did not appear in Experiment 1A, so we did not try to interpret it.

Reference-sentence reading times

Based on outlier exclusion criteria, 5.2% of the data were excluded from further analyses. In general, as in Experiment 1A, reading time (see **Table 4**) on the reference sentence decreased as

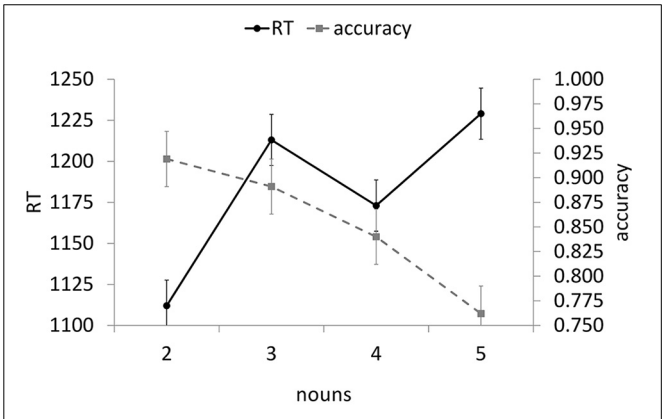


FIGURE 2 | Experiment 1B distractor probe reactions times and accuracies by noun condition (error bars indicate SE of the mean).

Table 4 | Experiment 1B mean per-character reading times in ms (with standard errors of the mean).

	List sentence	Reference sentence
Two-noun	77.0 (1.11)	60.9 (0.88)
Three-noun	80.3 (1.11)	59.8 (0.88)
Four-noun	80.3 (1.11)	58.3 (0.88)
Five-noun	82.5 (1.11)	57.4 (0.88)

the number of nouns in the list sentence increased. The linear trend was significant, $F_{1(1,53)} = 11.74$, $p = 0.001$, $F_{2(1,\ 30)} = 11.52$, $p = 0.002$, $\eta_p^2 = 0.18$, with no significant higher-order trends.

DISCUSSION

The probe word results were again consistent with the fan-effect hypothesis. As predicted, the presence of distractors interfered with the probe recognition task. Increasing the number of referents in the list sentence decreased recognition accuracy and increased reaction times for distractors similar to the effect found for referents in Experiment 1A. This result extends the findings of Experiment 1A to show that distractors also decrease in activation as the number of referents increases.

As in Experiment 1A, the reading-time results did not support the fan-effect hypotheses. Subjects again read the reference sentence more quickly as the number of distractors increased. This replication provides additional confidence that the unexpected results were not occurring due to chance. This issue was explored in further detail in Experiments 2A and 2B.

EXPERIMENTS 1A AND 1B COMBINED ANALYSIS

As noted in the introduction to Experiment 1B, the effect of fan size should be different for referents (Experiment 1A) and distractors (Experiment 1B). To compare the magnitude of the effect of the number of nouns on referents and distractors, an additional analysis was conducted for the probe reaction times from Experiments 1A and 1B. Probe reaction times for each subject in both experiments were first linearly regressed on the number of nouns (cf. Lorch and Myers, 1990), and the slopes were then examined in an independent-samples t-test with experiment (i.e., probe: referent vs. distractor) as a between-subjects variable. This analysis revealed a non-significant effect of probe in the expected direction, with a substantially smaller mean slope among subjects responding to referents in Experiment 1A ($M_{\text{slope}} = 15.2$ ms/noun, $SE = 5.4$) than among subjects responding to distractors in Experiment 1B ($M_{\text{slope}} = 31.3$ ms/noun, $SE = 7.6$), $t_{(137)} = 1.77$, $p = 0.08$, $d = 0.30$.

A similar analysis performed on the accuracy data revealed a large and significant effect of probe, with a substantially smaller mean slope among subjects responding to referents in Experiment 1A ($M_{\text{slope}} = -0.014$ accuracy/noun, $SE = 0.0046$) than among subjects responding to distractors in Experiment 1B ($M_{\text{slope}} = -0.052$ accuracy/noun, $SE = 0.0084$), $t_{(137)} = 4.33$, $p < 0.001$, $d = 0.74$. Although referents likely gained an advantage in both accuracy and speed of responding due to having appeared more recently than distractors, these analyses focused on the linear trends in which distance from the probe were equal. Therefore, these results provide evidence that the interference effect is greater for distractors than referents; this effect was tested more directly in Experiments 2A and 2B.

EXPERIMENT 2A

The procedure for Experiment 2A was modified from that in Experiments 1A and 1B such that subjects read the reference sentence one word at a time. This allowed for a more detailed analysis of the reading times, which was necessary to help understand the unexpected reference-sentence reading time results of Experiments 1A and 1B. The prediction that additional distractors should slow reading of the reference sentence was based on the hypothesis that multiple distractors would interfere with anaphor resolution. This means that the expected slowdown should occur specifically on the anaphor or immediately after the anaphor during spillover processing. According to this hypothesis, it was expected that there should be no difference in reading times on the reference-sentence until subjects reach the anaphor and post-anaphor regions, where they were expected to read more slowly as the number of distractors increased. However, if the results of Experiments 1A and 1B are reliable, then there should be longer reading times when there are more distractors at some point in the reference sentence prior to the anaphor.

In addition, Experiments 1A and 1B demonstrated that the presence of multiple distractors made recognition of both referents and distractors more difficult, as indexed by both reaction time and accuracy. Experiments 2A and 2B were designed to manipulate the probe word within subjects to address potential concerns about comparing results across experiments. Thus, in these experiments, probe word (referent vs. distractor) and number of distractors (two vs. five) were manipulated within subjects. The fan-effect hypothesis predicts that additional distractors would slow recognition and decrease accuracy for both referents and distractors. Moreover, to the extent that anaphor resolution focuses activation on the referent, thereby minimizing interference, the degree of interference should be greater for distractors than for referents.

METHOD

Subjects

Seventy-five students enrolled in a general psychology course at the University of Arkansas participated in the experiment to partially fulfill a research requirement. All subjects were native-English speakers.

Materials and design

Thirty of the experimental materials from Experiment 1 were used and appeared in only the two- and five-noun list conditions. This also required some modification of the list length in the filler sentences to maintain an equal distribution of list lengths across the entire experiment. In addition, the probe words were manipulated within subjects, such that each subject saw an equal number of referent and distractor probes following experimental items.

Subjects saw each experimental sentence pair in one of the four conditions along with all filler sentence pairs. Four

counterbalanced lists were created with the following constraints: approximately (i.e., 7 or 8 items) one quarter of the list sentences had two nouns followed by a referent probe, approximately one quarter had two nouns followed by a distractor probe, approximately one quarter had five nouns followed by a referent probe, and approximately one quarter had five nouns followed by a distractor probe. Because counterbalancing order did not have any important effects in Experiments 1A and 1B, order was no longer manipulated, resulting in a 2 (nouns: 2, 5) × 2 (probe word: referent, distractor) completely within-subjects design.

Procedure

The experiment was conducted using Linger (Rohde, 2003) to present the materials using a moving window (Just et al., 1982). Before starting the experiment, subjects completed three practice trials to familiarize themselves with the procedure. Each trial began with two rows of dashes, centered on the left-hand side of the screen, with each dash replacing a character or space in the sentences. Subjects pressed the spacebar to initially present the list sentence in its entirety. When they finished reading the list sentence, subjects pressed the spacebar again which replaced the list sentence with dashes and revealed the first word of the reference sentence. Subjects continued to press the spacebar to advance from one word to the next, with each press replacing the previous word with dashes and revealing the next word in the sentence. Pressing the spacebar after the final word of the reference sentence removed all of the dashes from the screen and presented a probe word in all capital letters in the center of the screen. Subjects responded to the probe word using the F key for yes and the J key for no. The response removed the probe word from the screen and replaced it with a comprehension question. Subjects again responded using the F and J keys, which advanced the screen to the next trial.

The experimental session consisted of 98 trials (30 experimental and 68 fillers) in two blocks of 49 trials each with the order of the trials completely randomized. Subjects were instructed to read the sentences as they normally would for comprehension and to respond to the probe words as quickly and accurately as possible. Subjects were free to take breaks between trials. The experiment lasted approximately 30 min.

RESULTS

Data exclusion and general analytic considerations

Based on the data exclusion criteria, the data from six subjects were excluded from further analysis. Therefore, the reported analyses included 69 subjects and 30 items.

Comprehension

In general, comprehension (see **Table 2**) decreased as the number of nouns increased. A 2 (nouns: 2, 5) × 2 (noun probed: referent, distractor) repeated-measures ANOVA revealed a main effect of nouns that was non-significant in the subject analysis, $F_{1(1,\ 68)} = 3.38$, $p = 0.07$, $\eta_p^2 = 0.05$, but significant in the items analysis, $F_{2(1,\ 29)} = 4.82$, $p = 0.04$. The main effect of noun probed was non-significant, $F_{1(1,\ 68)} = 2.62$, $p = 0.11$, $F_{2(1,\ 29)} = 1.07$, $p = 0.31$, and the interaction between number of nouns and noun probed was also non-significant, $F_{1(1,\ 68)} = 0.01$, $p = 0.92$, $F_{2(1,\ 29)} = 0.14$, $p = 0.71$.

Probe accuracy

Table 5 presents mean accuracy and probe reaction times as a function of the number of nouns and the noun probed. In general, accuracy was higher for referents than for distractors and when there were two nouns in the list sentence than when there were five. A 2 (nouns: 2, 5) × 2 (noun probed: referent, distractor) repeated-measures ANOVA revealed a significant main effect of the number of nouns, $F_{1(1,\ 68)} = 28.34$, $p < 0.001$, $F_{2(1,\ 29)} = 25.46$, $p < 0.001$, $\eta_p^2 = 0.29$, as well as a significant main effect of the noun probed, $F_{1(1,\ 68)} = 17.62$, $p < 0.001$, $F_{2(1,\ 29)} = 28.98$, $p < 0.001$, $\eta_p^2 = 0.21$. There was also a significant interaction between the number of nouns in the sentence and the noun being probed, $F_{1(1,\ 68)} = 4.51$, $p = 0.04$, $F_{2(1,\ 29)} = 4.37$, $p = 0.05$, $\eta_p^2 = 0.06$, with a greater 2- vs. 5-noun difference for distractors than for referents, replicating the effect seen in the between-experiments comparison presented above. Planned pairwise comparisons revealed a significant effect of the number of nouns for both the referent probes, $t_{1(68)} = 3.04$, $p = 0.003$, $t_{2(29)} = 3.69$, $p = 0.001$, $d = 0.37$, and the distractor probes, $t_{1(68)} = 4.53$, $p < 0.001$, $t_{2(29)} = 4.17$, $p < 0.001$, $d = 0.55$.

Probe reaction times

Based on outlier exclusion criteria, 7.8% of the data were excluded from further analyses. Like the accuracy results, reaction time tended to be shorter for referents than for distractors and when there were two nouns in the list sentence than when there were five. A 2 (nouns: 2, 5) × 2 (noun probed: referent, distractor) repeated-measures ANOVA revealed a significant main effect of the number of nouns, $F_{1(1,\ 68)} = 4.20$, $p = 0.04$, $F_{2(1,\ 29)} = 5.99$, $p = 0.02$, $\eta_p^2 = 0.06$, as well as a significant main effect of the noun probed, $F_{1(1,\ 68)} = 12.73$, $p = 0.001$, $F_{2(1,\ 29)} = 19.18$, $p < 0.001$, $\eta_p^2 = 0.16$. Despite the pattern of means replicating the cross-experiment interaction seen in Experiments 1A and 1B, there was not a significant interaction between the number of nouns in the sentence and the noun being probed, $F_{1(1,68)} = 0.28$, $p = 0.60$, $F_{2(1,\ 29)} = 2.64$, $p = 0.12$. Planned pairwise comparisons revealed a non-significant 46 ms effect of the number of nouns for the antecedents, $t_{1(68)} = 1.35$, $p = 0.18$, $t_{2(29)} = 0.80$, $p = 0.43$, but the 73 ms effect of the number of nouns

Table 5 | Experiments 2A and 2B mean probe word responses (with standard errors of the mean).

	Experiment 2A			
	Accuracy		Reaction time (ms)	
	Referent	Distractor	Referent	Distractor
Two-noun	0.97 (0.013)	0.92 (0.013)	1553 (25.1)	1632 (25.1)
Five-noun	0.93 (0.013)	0.83 (0.013)	1599 (25.1)	1705 (25.1)
	Experiment 2B			
	Accuracy		Reaction time (ms)	
	Referent	Distractor	Referent	Distractor
Two-noun	0.95 (0.017)	0.92 (0.019)	1151 (21.5)	1273 (21.5)
Five-noun	0.94 (0.013)	0.72 (0.022)	1245 (21.5)	1403 (21.5)

for distractor probes, though not significant by subjects, $t_{1(68)} = 1.73$, $p = 0.09$, was significant by items, $t_{2(29)} = 3.00$, $p = 0.005$, $d = 0.21$. For the sake of comparison with Experiments 1A and 1B, in Experiment 2A the slope of the number of nouns among the referents was 15.4 ms/noun, whereas the slope of the number of nouns among the distractors was 24.2 ms/noun. These values were 15.2 and 31.3, respectively, in Experiments 1A and 1B.

Reference-sentence reading times

Outliers were first identified as words read for less than 150 ms or more than 700 ms; different criteria were used in this experiment to try to approximate in a per-word measure the per-character measures used in the previous experiments. Outliers among the remaining reading times were then identified within each subject based on Tukey's (1977) criteria. This resulted in 8.1% of the trials being excluded from further analysis[3].

The individual-word reading times were combined into three regions of three words each. The pre-anaphor region consisted of the three words prior to the anaphor; the anaphor region consisted of the three-word noun phrase involving the determiner, adjective, and anaphor (e.g., *the cutting tool*); and the post-anaphor region consisted of the three words following the anaphor. Although some items had more than three words prior to the anaphor noun phrase, the analysis was restricted to this point because there was a dramatic drop in the number of observations starting four words prior to the anaphor region. The post-anaphor region was always the final three words of the anaphor sentence. Thus, each region consisted of three words, making their reading times roughly comparable.

In general, reading time on the reference sentence decreased as the number of nouns in the list sentence increased (see **Figure 3**); this effect occurred most strongly in the pre-anaphor region. A 2 (nouns: 2, 5) × 3 (region: pre-anaphor, anaphor, post-anaphor) repeated measures ANOVA revealed a significant main effect of the number of nouns only in the items analysis, $F_{1(2,\ 68)} = 2.68$, $p = 0.11$, $F_{2(1,\ 29)} = 4.66$, $p = 0.04$, $\eta_p^2 = 0.04$. There was also a significant main effect of region, $F_{1(2,\ 136)} = 29.9$, $p < 0.001$, $F_{2(1,\ 58)} = 16.8$, $p < 0.001$, $\eta_p^2 = 0.31$, but the interaction between the number of nouns and region was non-significant, $F_{1(2,\ 136)} = 0.65$, $p = 0.53$, $F_{2(1,\ 58)} = 1.03$, $p = 0.36$. Planned pairwise comparisons revealed that subjects read the pre-anaphor region significantly faster in the five noun condition compared to the two noun condition ($p = 0.02$ by subjects, $p = 0.05$ by items), but this effect was non-significant in the anaphor region (p =0.26 by subjects, $p = 0.16$ by items) and the post-anaphor region ($p = 0.51$ by subjects, $p = 0.21$ by items).

DISCUSSION

As predicted by the fan-effect hypothesis, and consistent with Experiments 1A and 1B, probe word accuracy was higher and responses were made faster in the two-noun condition than in the five-noun condition for both referents and distractors. Moreover, the cross-experiment interaction of number of nouns and probe type was replicated; the fan effect is larger for distractors. The reading time results replicated those from Experiments 1A and 1B: subjects read the reference sentence faster in the five-noun condition than in the two-noun condition. However, measuring reading time per-word enabled a more detailed analysis of the reference-sentence reading times and revealed that the faster reading primarily occurred in the pre-anaphor region. Because this region was identical across conditions and made no reference to the list sentence, there is no theoretical reason to expect this difference based on anaphor resolution processes. Instead, these results support the speeded-reading explanation suggested in the discussion of Experiment 1A, that subjects may have adopted a particular strategy in order to mitigate the increased difficulty of the probe-word task in the five-noun condition by reaching the probe word task and comprehension questions more quickly. Furthermore, per-character reading times on the list sentence (see Appendix B in Supplementary Material) increased as the number of nouns increased, suggesting that the speeded-reading strategy was adopted only on the reference sentence after subjects became aware of the increased difficulty imposed by the longer lists.

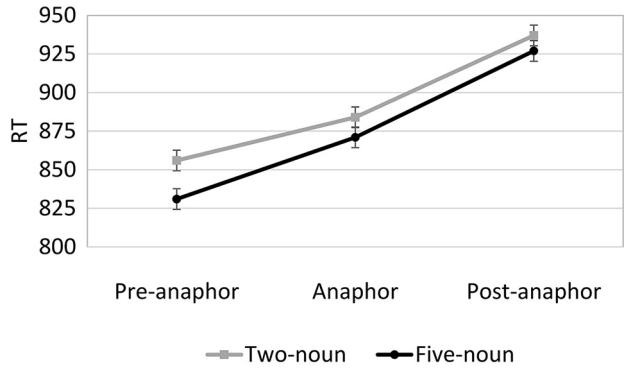


FIGURE 3 | Experiment 2A mean anaphor sentence reading times per region (error bars indicate SE of the mean).

EXPERIMENT 2B

Because subjects appeared to be adjusting their reading speed to accommodate the difficulty of representing multiple referents, it was important to assess whether the probe word results were dependent on this apparent strategy. Experiment 2B was thus a replication of Experiment 2A using a fixed-rate presentation of the passages. By controlling the pace of reading, any effects found on the probe recognition task can be assumed to reflect processes that occurred independent of subjects' variable reading speed. Holding reading-rate constant was not expected to change the probe-word results, so it was expected that responses to both referents and distractors would be faster and more accurate when there were two referents in the list sentence than when there were five referents. Moreover, this experiment provided one more opportunity to examine the prediction that the effect of fan would be greater among distractors than among referents. In the accuracy data, the fan effect has been reliably much stronger for distractors than it has been among referents. In the reaction-time data, between Experiments 1A and 1B, this effect was significant

[3]The pattern of results remained similar using a less-strict cutoff of 1500 ms/word.

only in a one-tailed test, and in Experiment 2A, the same pattern emerged but it was not reliable.

METHOD

Subjects

Sixty-six students enrolled in a general psychology course at the University of Arkansas participated in the experiment to partially fulfill a research requirement. All subjects were native-English speakers.

Materials, design, and procedure

The materials, design, and procedure were identical to Experiment 2A except that the materials were presented at a fixed pace of 450 ms per word[4].

RESULTS

Data exclusion and general analytic considerations

Based on the data exclusion criteria, the data from 10 subjects were excluded from further analysis. Four more subjects were removed from further analysis for a systematic misunderstanding of the instructions. Therefore, the analyses included 50 subjects and 30 items.

Comprehension

In general, comprehension (see **Table 2**) decreased as the number of nouns increased. A 2 (nouns: 2, 5) × 2 (noun probed: referent, distractor) repeated-measures ANOVA revealed a main effect of nouns that was significant in the subject analysis, $F_{1(1,\ 49)} = 4.59$, $p = 0.04$, $\eta_p^2 = 0.09$, but non-significant in the items analysis, $F_{2(1,\ 29)} = 2.53$, $p = 0.12$. The main effect of noun probed was non-significant, $F_{1(1,\ 49)} = 0.20$, $p = 0.66$, $F_{2(1,\ 29)} = 0.23$, $p = 0.63$, and the interaction between number of nouns and noun probed was also non-significant, $F_{1(1,\ 49)} = 1.50$, $p = 0.23$, $F_{2(1,\ 29)} = 1.31$, $p = 0.26$.

Probe accuracy

Table 5 presents mean accuracy and probe reaction times as a function of the number of nouns and the noun probed. In general, accuracy was higher for referents than for distractors and when there were two nouns in the list sentence than when there were five, once again replicating the pattern seen in Experiments 1A, 1B, and 2A. A 2 (nouns: 2, 5) × 2 (noun probed: referent, distractor) repeated-measures ANOVA revealed a significant main effect of the number of nouns, $F_{1(1,\ 49)} = 53.9$, $p < 0.001$, $F_{2(1,\ 29)} = 27.9$, $p < 0.001$, $\eta_p^2 = 0.52$, as well as a significant main effect of the noun probed, $F_{1(1,\ 49)} = 53.0$, $p < 0.001$, $F_{2(1,\ 29)} = 31.8$, $p < 0.001$, $\eta_p^2 = 0.52$. Additionally, there was a significant interaction between the number of nouns in the sentence and the noun being probed, $F_{1(1,\ 49)} = 29.2$, $p < 0.001$, $F_{2(1,\ 29)} = 23.1$, $p < 0.001$, $\eta_p^2 = 0.37$, with a greater 2- vs. 5-noun difference for distractors than for referents, the third time this pattern has been replicated.

[4]Due to limitations of the Linger program, words were presented at a fixed rate instead of using a variable rate dependent on the length of each word (cf. Gernsbacher, 1989).

Probe reaction times

Based on outlier exclusion criteria, 8.7% of the data were excluded from further analyses. Like the accuracy results, reaction time tended to be shorter for referents than for distractors and when there were two nouns in the list sentence than when there were five. A 2 (nouns: 2, 5) × 2 (noun probed: referent, distractor) repeated-measures ANOVA showed that there was a significant main effect of the number of nouns, $F_{1(1,\ 49)} = 25.8$, $p < 0.001$, $F_{2(1,\ 29)} = 19.4$, $p < 0.001$, $\eta_p^2 = 0.35$, as well as a significant main effect of the noun probed, $F_{1(1,\ 49)} = 44.1$, $p < 0.001$, $F_{2(1,\ 29)} = 31.9$, $p < 0.001$, $\eta_p^2 = 0.47$. Once again, the pattern of means replicated the cross-experiment pattern seen in Experiments 1A and 1B as well as that seen in Experiment 2A, with the effect of number of nouns being larger for distractors than for referents. Despite this, there was not a significant interaction between the number of nouns in the sentence and the noun being probed, $F_{1(1,\ 49)} = 0.70$, $p = 0.41$, $F_{2(1,\ 29)} = 2.03$, $p = 0.17$. The effect of the number of nouns was significant among the referents, $t_{1(49)} = 2.86$, $p = 0.006$, $t_{2(29)} = 2.54$, $p = 0.02$, as well as among the distractors, $t_{1(49)} = 4.55$, $p < 0.001$, $t_{2(29)} = 4.87$, $p < 0.001$; this effect was numerically smaller for referents (94 ms, $d = 0.40$) than for distractors (130 ms, $d = 0.64$). The slopes corresponding to these effects, 31.2 ms/noun for referents and 43.2 ms/noun for the distractors, were substantially larger than the respective slopes seen in the previous experiments, possibly due to the change in the presentation of the passages to experimenter-paced.

DISCUSSION

The results confirmed the predictions of the fan-effect hypothesis, and the probe-word results were conceptually identical to Experiment 2A. Although subjects in the previous experiments seemed to be adopting a special strategy of reading the reference sentence more quickly when there were more distractors, the results of Experiment 2B indicate that this strategy was not necessary for the emergence of the probe-word results we had previously observed because subjects did not have the opportunity to employ it. The replication of the finding that the activation level of nouns decreases as the number of distractors increases therefore appears to be the result of a diffusion of activation to all potential referents.

However, this conclusion relies on the assumption that subjects were resolving the anaphor and that the anaphor processing affected the activation level of the referents. There is some evidence, however, that anaphor resolution may not always occur during reading (Greene et al., 1992; Levine et al., 2000; Klin et al., 2004, 2006; Love and McKoon, 2011), making it possible that the present results could be occurring independent of anaphor resolution. The effect of nouns may have been caused by the increasing memory demands incurred as the number of referents increased regardless of whether the subjects attempted to resolve the anaphors. It is possible that as the amount of information in the subjects' mental representations increased, the probability of the correct referent being activated even by the probe word itself, independent of anaphor resolution processes, decreased, resulting in slower reaction times as the number of referents.

EXPERIMENT 3

Experiment 3 was designed to address the possibility that anaphors were not being resolved in the prior experiments. To do so, the reference sentence was modified such that it contained an anaphor or not (see **Table 1**), a manipulation that has been used many times in the anaphor resolution literature (e.g., Dell et al., 1983; Levine et al., 2000). As in Experiments 2A and 2B, there were either two nouns (i.e., a referent and one distractor) or five nouns (i.e., a referent and four distractors) in the list sentence that preceded the reference sentence. The referent was used as the probe word to provide an index of the activation of this concept at the end of the anaphor or no-anaphor sentence. According to the fan-effect hypothesis, it is activation from the anaphor as a memory cue that is divided among the referent and the distractors that is the source of the effect of the number of nouns. Thus, when there is an anaphor, the fan-effect hypothesis predicts an effect of the number of nouns like that seen in the previous experiments. Whatever pattern emerges for the effect of the number of nouns in the anaphor condition, because anaphor resolution involves reactivation of the correct referent (e.g., Dell et al., 1983), there should be an overall accuracy and reaction time advantage in the anaphor over the no-anaphor control condition.

METHOD

Subjects

Seventy students enrolled in a general psychology course at the University of Arkansas participated in the experiment to partially fulfill a research requirement. All subjects were native-English speakers.

Materials and design

Experiment 3 used the same set of materials as Experiments 2A and 2B with the exception that the reference sentence was manipulated (see **Table 1**) such that it included an anaphor (i.e., Anaphor condition) or not (i.e., No Anaphor condition), while equating for length (i.e., the mean length for both the anaphor and no anaphor conditions was 61.5 characters). Finally, the probe words were limited to referents only, as in Experiment 1A. The manipulation of these factors resulted in a 2 (nouns: 2, 5) × 2 (reference: anaphor, no anaphor) completely within-subjects design.

Procedure

The procedure of Experiment 3 was identical to that of Experiments 1A and 1B, except that it included only 98 trials (30 experimental and 68 fillers), as in Experiments 2A and 2B.

RESULTS

Data exclusion and general analytic considerations

Based on outlier identification and comprehension and probe accuracy, the data from 5 subjects were excluded from further analysis. Therefore, the reported analyses included 65 subjects and 30 items.

Comprehension

In general, comprehension (see **Table 2**) decreased as the number of nouns increased and accuracy was greater in the anaphor condition than in the no anaphor condition. A 2 (nouns: 2, 5) × 2 (reference: anaphor, no anaphor) repeated measures ANOVA revealed a non-significant main effect of nouns, $F_{1(1,\ 64)} = 0.62$, $p = 0.44$, $F_{2(1,\ 30)} = 0.90$, $p = 0.35$, and a significant main effect of reference, $F_{1(1,\ 64)} = 9.28$, $p = 0.003$, $F_{2(1,\ 30)} = 6.15$, $p = 0.019$, $\eta_p^2 = 0.13$. In addition, there was a significant interaction between the number of nouns and reference, $F_{1(1,\ 64)} = 4.83$, $p = 0.032$, $F_{2(1,\ 30)} = 5.23$, $p = 0.029$, $\eta_p^2 = 0.07$, with a 7.3% accuracy advantage for the 2-noun condition compared to the 5-noun condition in the anaphor condition but only a 1.3% accuracy advantage in the no anaphor condition. However, the comprehension questions differed between the anaphor and no anaphor conditions, making this the likely cause of the observed effect.

Probe accuracy

Table 6 presents mean probe word accuracy and reaction times along with mean reference-sentence reading times as a function of the number of referents and whether the reference sentence contained an anaphor. In general, subjects responded more accurately in the anaphor condition than in the no anaphor condition and when there were two nouns in the list sentence than when there were five nouns. A 2 (nouns: 2, 5) × 2 (reference: anaphor, no anaphor) repeated measures ANOVA revealed a significant main effect of the number of nouns, $F_{1(1,\ 64)} = 8.43$, $p = 0.005$, $F_{2(1,\ 29)} = 14.14$, $p = 0.001$, $\eta_p^2 = 0.12$; however, the simple effect of the number of nouns for the anaphor condition was not significant, $t_{1(64)} = 1.11$, $p = 0.27$, $t_{2(29)} = 1.21$, $p = 0.24$. There was also a significant main effect of reference, $F_{1(1,\ 64)} = 7.53$, $p = 0.008$, $F_{2(1,\ 29)} = 6.56$, $p = 0.02$, $\eta_p^2 = 0.11$, but the interaction between the number of nouns and reference was non-significant, $F_{1(1,\ 64)} = 2.70$, $p = 0.11$, $F_{2(1,\ 29)} = 2.84$, $p = 0.10$.

Probe reaction times

Based on outlier exclusion criteria, 7.5% of the data were excluded from further analyses. Reaction times (see **Table 6**) tended to be faster in the anaphor condition than in the no anaphor condition and when there were two nouns in the list sentence than when there were five nouns. A 2 (nouns: 2, 5) × 2 (reference: anaphor, no anaphor) repeated measures ANOVA revealed that the main effect of nouns was non-significant, $F_{1(1,\ 64)} = 1.26$,

Table 6 | Experiment 3 mean probe word responses and per-character reading times (with standard errors of the mean).

	Anaphor condition			
	Probe word responses		Per-character reading time	
	Accuracy	Reaction time	List sentence	Reference sentence
Two-noun	0.96 (0.01)	976 (9.9)	70.7 (1.32)	55.7 (0.64)
Five-noun	0.94 (0.01)	997 (9.9)	73.7 (1.32)	54.0 (0.64)
	No anaphor condition			
	Probe word responses		Per-character reading time	
	Accuracy	Reaction time	List sentence	Reference sentence
Two-noun	0.95 (0.01)	1009 (9.9)	–	51.2 (0.64)
Five-noun	0.89 (0.01)	1013 (9.9)	–	50.1 (0.64)

$p = 0.27$, $F_{2(1, 29)} = 0.24$, $p = 0.63$. Because of the prediction of the fan-effect hypothesis, the effect of the number of nouns was examined for the anaphor condition. The noun-effect was 23 ms but was also not significant, $t_{1(64)} = 1.26$, $p = 0.21$, $t_{2(29)} = 0.80$, $p = 0.43$. The main effect of reference was nearly significant in the subjects analysis, $F_{1(1, 64)} = 3.38$, $p = 0.07$, $\eta_p^2 = 0.05$, but non-significant in the items analysis, $F_{2(1, 29)} = 2.32$, $p = 0.14$, and the interaction between reference and nouns was not significant, $F_{1(1, 64)} = 0.65$, $p = 0.42$, $F_{2(1, 29)} = 1.09$, $p = 0.31$.

Reference-sentence reading times

Based on outlier exclusion criteria, 6.5% of the data were excluded from further analyses. In general, reading time (see **Table 6**) was longer when the sentence contained an anaphor than when it did not. A 2 (nouns: 2, 5) × 2 (reference: anaphor, no anaphor) repeated measures ANOVA revealed a significant main effect of reference, $F_{1(1, 64)} = 23.29$, $p < 0.001$, $F_{2(1, 29)} = 13.01$, $p = 0.001$, $\eta_p^2 = 0.27$. The main effect of nouns was not quite significant, $F_{1(1, 64)} = 3.08$, $p = 0.08$, $F_{2(1, 29)} = 1.56$, $p = 0.22$, although the pattern observed in Experiments 1A, 1B, and 2A appeared once again, with shorter reading times when there were more nouns. There was not a significant interaction between reference and nouns, $F_{1(1, 64)} = 0.23$, $p = 0.64$, $F_{2(1, 29)} = 0.55$, $p = 0.46$.

DISCUSSION

The results from Experiment 3 provided some evidence that subjects were in fact resolving the anaphors when reading the passages. Probe accuracy was better after reading a sentence with an anaphoric reference than after reading a sentence that did not make an anaphoric reference. Additional evidence that subjects were resolving the anaphors comes from the reading-time data. Controlling for length, the reference sentences were read more slowly when they contained an anaphor than when they did not, consistent with the hypothesis that subjects were engaging in additional processing to resolve the anaphor. This conclusion is tentative, though, as there were more explicit references[5] (e.g., pronouns, specifiers, definite noun phrases) to entities in the prior sentence in the reference sentences in the anaphor condition ($M = 2.8$, $SD = 0.8$) than in the no-anaphor condition ($M = 1.6$, $SD = 0.8$), $t_{(29)} = 6.27$, $p < 0.001$. In most cases (25 of 30 passages), these additional references were not to any of the list items; excluding the five passages with a second reference to list items does not change the pattern of results for probe accuracy or reaction time reported above.

The results of this experiment's anaphor condition were less consistent with the fan-effect hypothesis than the results from prior experiments, although the general pattern of degraded recognition performance with more nouns persisted; we return to this issue in the General Discussion.

GENERAL DISCUSSION

Explanations of how anaphoric expressions are understood have frequently appealed to general memory processes. Consistent with theories of comprehension that place memory at their center (e.g., Kintsch, 1988; Myers and O'Brien, 1998; Lewis and Vasishth, 2005), anaphor resolution is more difficult when factors are present that make retrieving a unique item from memory more difficult, such as when there is similarity between a desired target and some distractor. Prior research that has produced findings that are consistent with this hypothesis (e.g., Corbett and Chang, 1983; Corbett, 1984; O'Brien, 1987; O'Brien et al., 1990; Greene et al., 1992; Levine et al., 2000; Badecker and Straub, 2002; Klin et al., 2004, 2006) have used stimuli with one distractor and one antecedent, and by a variety of measures anaphor resolution has been shown to be more difficult because of the distractor. In five experiments, we examined the hypothesis that a greater number of distractors would lead to a fan effect (Anderson, 1974) in anaphor resolution, that is, if with each additional distractor there would be additional difficulty in identifying the correct referent of the anaphor. We also examined the effect of additional distractors on the activation of those distractors. Our subjects read pairs of sentences, the first of which provided a variably-long list of concepts from the same taxonomic category and the second of which made unambiguous reference to one of the items in the list with an adjective-modified definite noun phrase; this was followed by a probe recognition task that should provide an index of how active the probed concept is in the text representation.

Collectively, the probe word results from the present experiments supported the hypothesis that distractors have a cumulative effect on antecedent activation levels. Although the effect of the distractors on reaction time varied in size and significance from experiment to experiment, it is overall a robust effect. The two- and five-noun conditions with a referent probe were present in Experiments 1A, 2A, 2B, and 3. The subject data from these four experiments were combined and submitted to a 2 (nouns: 2, 5) × 4 (Experiments: 1A, 2A, 2B, 3) mixed-factor ANOVA with repeated-measures on the first factor. The effect of nouns was significant, $F_{(1, 265)} = 15.56$, $p < 0.001$, and the interaction was not, $F_{(3, 265)} = 1.18$, $p = 0.32$, suggesting that there was not significant variability in the effect of nouns across experiments. Cohen's d for the effect of nouns was 0.24 (95% confidence interval: 0.12, 0.36; Smithson, 2003), demonstrating a small but reliable effect. Whereas previous research has shown that the presence of a single distractor interferes with the activation of the antecedent, the present research extends this finding by demonstrating that each additional distractor further reduces the activation level of the antecedent and other distractors. This effect is akin to a set size effect (Sternberg, 1966), with larger lists leading to longer reaction times; however, the difference in the size of the effect for referents and distractors suggests that an additional process related to anaphor resolution is also occurring.

The present results are conceptually similar to the fan effect where delayed recognition [i.e., the recognition task occurring after the presentation of all of the materials as in Anderson (1974)] slows as the number of facts associated with a noun increases. This effect is generally attributed to the reduction in the probability of the correct item in memory being activated at the time of retrieval, thus slowing responses. The present experiments demonstrate an earlier effect, with the number of distractors affecting the activation level of nouns immediately after each

[5]Associative anaphora (e.g., referring to *the test* after a sentence mentioning studying) were not counted.

trial. In this case, the categorical anaphor (e.g., *tool*) acts as a retrieval cue, with activation being split among all of the concepts associated with the category (i.e., the referent and distractor[s]). Increasing the number of distractors should therefore increase the time required to resolve the anaphor. This increased retrieval time effect was not observed in the present experiments, although this was likely due to subjects adopting a speeded-reading strategy (see the discussion of the reading-time results below). As a consequence of multiple potential antecedents, activation should be divided among the concepts, limiting the activation for each one (see spreading activation theory; Collins and Loftus, 1975; Anderson, 1983). This prediction was supported by the slowed reaction times and the reduced accuracy resulting from increasing the number of distractors. The present results further demonstrate that activation does not spread equally to all category members when there is disambiguating information (e.g., an adjective modifier like *cutting* in *the cutting tool*). Increasing the number of nouns led to a consistently greater reduction in probe accuracy and increase in reaction time for distractors than for referents in the Experiments 1A and 1B combined analysis and Experiments 2A and 2B, suggesting that activation was spreading disproportionately to the referent.

We have framed the current results as primarily being an effect that occurs at the time of retrieval (i.e., upon reading the anaphor). It is possible that these effects are also influenced by encoding or storage interference. Upon reading multiple items with many shared features, like our list-sentence items, the mental representation of these items may be overwritten (Nairne, 1990) or degraded due to repeated reactivation by similar items (Estes, 1997). The methodology used in the current research does not allow for delineation between a storage-based and a retrieval-based explanation. Ferreting out the relative contributions of storage- and retrieval-interference processes would likely require careful parametric manipulation of feature overlap among distractors and the referent as well as precise control over not only timing of reading and probes but also time elapsed between storage and retrieval, as well as manipulation of serial position of distractors and referents. Attempting to work out these details is a promising avenue for future research.

Turning to the reading-time results, we found no evidence that additional distractors led to more difficulty processing anaphoric reference. By contrast, we consistently found that our subjects read faster as there were more distractors. We believe that this is the result of subjects adopting a speeded-reading strategy on difficult trials (i.e., trials with longer lists of nouns), which counteracted the predicted increase in anaphor reading time. This is similar to Van Dyke and McElree's (2006) finding that, while reading grammatically-complex sentences, subjects read faster and had worse comprehension while holding a memory load (i.e., a list of three words) than when not holding a memory load, suggesting a dual-task strategic trade-off. Our subjects also had lower comprehension with greater list length (see **Table 2**), suggesting that there was possibly a task demand that shifted attention somewhat from the comprehension aspect of the task to the memory aspect of the task. In no case, however, was comprehension lower than about 83%. Moreover, there is no theoretical reason to expect anaphor resolution to take less time as the number of candidate antecedents increases unless subjects were giving up on trying to identify the correct antecedent (Levine et al., 2000). There are a few arguments consistent with the notion that subjects were in fact resolving the anaphors in the current research. First, correctly answering a large majority of the comprehension questions required the anaphors to be resolved, which some have suggested is necessary to get subjects to resolve anaphors in anaphor resolution research (Foertsch and Gernsbacher, 1994). Second, some subjects, especially in Experiment 1B, spontaneously adopted the strategy of labeling distractors as new in the probe recognition task, which suggests that they had selected the referent as the "correct" answer and distractors as the "incorrect" answer to the probe task. Third, Experiment 3 provides tentative evidence that subjects were resolving the anaphor, even on five-noun trials. Given these arguments and findings, we believe that our subjects were resolving anaphors even when it was difficult to do so. Therefore, the speeded-reading strategy appears to be the most parsimonious explanation of these unexpected results. Furthermore, the fixed-pace presentation of the sentence in Experiment 2B prevented subjects from engaging in the speeded-reading strategy, demonstrating that the probe word effects do not rely on such a strategy. Future research should attempt to prevent the speeded-reading strategy while maintaining naturalistic reading (e.g., introducing a substantial delay between the passages and the probe task or eliminating the probe task entirely) in order to better evaluate the anaphor reading time hypothesis.

Finally, returning to the fan-effect hypothesis, the original explanation offered for the fan effect by Anderson (1974) was based on Anderson and Bower's (1973) theory of memory, which assumed that memory retrieval was based on search cues being used to identify, in parallel, matching elements in memory, which were then serially examined, resulting in an increase in reaction time with each additional matching element. In the former detail (i.e., a parallel matching), this theory is in the same family as other global-matching memory theories like those of Ratcliff (1978), Gillund and Shiffrin (1984), and Hintzman (1986), upon which memory-based text processing frameworks like Myers and O'Brien's (1998) resonance model are based. In this sense, the results of our experiments are confirmation of both theories of memory search and the hypothesis that at least some aspects of comprehension may be explained by general memory processes. However, other research into the fan effect has shown that there are circumstances under which there is no fan effect despite there being multiple associations with a single memory cue (Myers et al., 1984; Radvansky, 1998; Radvansky et al., 1998). Myers et al. found no fan effect when memory elements could be integrated causally. For example, reading the elements *the doctor went to the racetrack*, *the doctor studied the odds*, and *the doctor made a selection* may be readily integrated into a causally-coherent narrative representation about events occurring at a racetrack. Similarly, Radvansky and colleagues showed that the fan effect is reduced or even eliminated when potentially-competing memory elements can be readily integrated. One feature that makes elements easy to integrate is if they can occur at the same time (e.g., the grocer was folding a towel; the grocer was clearing his throat), whereas elements that are in different locations may not be integrated (e.g.,

the welcome mat is in the cocktail lounge; the welcome mat is in the office building). Radvansky et al. observed a fan-effect in recognition of hard-to-integrate elements, but not for easy-to-integrate elements. Given that there are boundary conditions for the fan-effect in memory experiments, a natural question to ask is if there are circumstances under which the search process in anaphor resolution might occur without interference. Across sentences, one such circumstance might be if the items in a list occur in more-naturalistic texts, allowing for an integrated situation model to be constructed, as suggested by both Myers et al. (1984) and Radvansky (1998; Radvansky et al., 1998). By contrast, within sentences, one condition that has been shown to limit the search for referents is when there are strong grammatical constraints on reference. Recent evidence from Dillon et al. (2013; see also Chow et al., 2014) suggests that syntactic principles may guide retrieval in a constrained manner for some linguistic dependencies, such as reflexives (but see Badecker and Straub, 2002; Kennison, 2003 and Sturt, 2003, for further complexities), leading to retrieval without interference from distractors; syntactic constraints may play an especially critical role in directing the retrieval processes that occur within a sentence. These types of findings are representative of two distinct research literatures have arisen over the past few decades, one focused on retrieval across sentences, and the other focused on retrieval within sentences. Integration of these theories and findings holds out the promise of yet further integration of theories of memory and comprehension.

AUTHOR NOTE

Portions of these data were presented at the 52nd Annual Meeting of the Psychonomic Society, 2011, Seattle, Washington, and at the 22nd Annual Meeting of the Society for Text and Discourse, 2012, Montreal, Canada.

ACKNOWLEDGMENTS

We thank Katie Berghorn, Willie Curry, Justin Dollman, Audrey Dunn, Alisha Foster, Samantha Herrera, Kelsey Lovewell, Mollie Price, Alicia Small, Colby Thompson, Jessica Turner, and James Wages for their assistance with stimulus preparation and data-collection, and Celia Klin for her comments on a previous draft of this paper. Two anonymous reviewers provided excellent constructive feedback on an earlier draft of this paper.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fpsyg.2014.00818/abstract

REFERENCES

Almor, A. (1999). Noun-phrase anaphora and focus: the informational load hypothesis. *Psychol. Rev.* 106, 748–765. doi: 10.1037/0033-295X.106.4.748

Anderson, J. R. (1974). Retrieval of propositional information from long-term memory. *Cogn. Psychol.* 6, 451–474. doi: 10.1016/0010-0285(74)90021-8

Anderson, J. R. (1983). A spreading activation theory of memory. *J. Verbal Learn. Verbal Behav.* 22, 251–295. doi: 10.1016/S0022-5371(83)90201-3

Anderson, J. R. (2005). Human symbol manipulation within an integrated cognitive architecture. *Cogn. Sci.* 29, 313–341. doi: 10.1207/s15516709cog0000_22

Anderson, J. R., and Bower, G. H. (1973). *Human Associative Memory*. Washington, DC: Winston and Sons.

Anderson, J. R., and Reder, L. M. (1999). The fan effect: new results and new theories. *J. Exp. Psychol. Gen.* 128, 186–197. doi: 10.1037/0096-3445.128.2.186

Badecker, W., and Straub, K. (2002). The processing role of structural constraints on the interpretation of pronouns and anaphors. *J. Exp. Psychol. Learn. Mem. Cogn.* 28, 748–769. doi: 10.1037/0278-7393.28.4.748

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., et al. (2007). The English lexicon project. *Behav. Res. Methods* 39, 445–459. doi: 10.3758/BF03193014

Chambers, C. G., and Smyth, R. (1998). Structural parallelism and discourse coherence: a test of centering theory. *J. Mem. Lang.* 39, 593–608. doi: 10.1006/jmla.1998.2575

Chow, W.-Y., Lewis, S., and Phillips, C. (2014). Immediate sensitivity to structural constraints in pronoun resolution. *Front. Psychol.* 5:630. doi: 10.3389/fpsyg.2014.00630

Collins, A. M., and Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychol. Rev.* 82, 407–428. doi: 10.1037/0033-295X.82.6.407

Corbett, A. T. (1984). Prenominal adjectives and the disambiguation of anaphoric nouns. *J. Verbal Learn. Verbal Behav.* 23, 683–695. doi: 10.1016/S0022-5371(84)90418-3

Corbett, A. T., and Chang, F. R. (1983). Pronoun disambiguation: accessing potential antecedents. *Mem. Cognit.* 11, 283–294. doi: 10.3758/BF03196975

Dell, G. S., McKoon, G., and Ratcliff, R. (1983). The activation of antecedent information during the processing of anaphoric reference in reading. *J. Verbal Learn. Verbal Behav.* 22, 121–132. doi: 10.1016/S0022-5371(83)80010-3

Dillon, B., Mishler, A., Sloggett, S., and Phillips, C. (2013). Contrasting intrusion profiles for agreement and anaphora: experimental and modeling evidence. *J. Mem. Lang.* 69, 85–103. doi: 10.1016/j.jml.2013.04.003

Ditman, T., Holcomb, P. J., and Kuperberg, G. R. (2007). The contributions of lexico-semantic and discourse information to the resolution of ambiguous categorical anaphors. *Lang. Cogn. Process.* 22, 793–827. doi: 10.1080/01690960601057126

Estes, W. K. (1997). Processes of memory loss, recovery, and distortion. *Psychol. Rev.* 104, 148–169. doi: 10.1037/0033-295X.104.1.148

Foertsch, J., and Gernsbacher, M. A. (1994). In search of comprehension: getting "minimalists" to work. *Discourse Process.* 18, 271–296. doi: 10.1080/01638539409544896

Foraker, S., and McElree, B. (2007). The role of prominence in pronoun resolution: active versus passive representations. *J. Mem. Lang.* 56, 357–383. doi: 10.1016/j.jml.2006.07.004

Garrod, S., and Sanford, A. J. (1981). "Bridging inferences and the extended domain of reference," in *Attention and Performance IX*, eds J. Long and A. Baddeley (Hillsdale, NJ: Erlbaum), 331–346.

Gernsbacher, M. A. (1989). Mechanisms that improve referential access. *Cognition* 32, 99–156. doi: 10.1016/0010-0277(89)90001-2

Gerrig, R., and McKoon, G. (1998). The readiness is all: the functionality of memory-based text processing. *Discourse Process.* 26, 67–86. doi: 10.1080/01638539809545039

Gillund, G., and Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychol. Rev.* 91, 1–67. doi: 10.1037/0033-295X.91.1.1

Greene, S. B., Gerrig, R. J., McKoon, G., and Ratcliff, R. (1994). Unheralded pronouns and management by common ground. *J. Mem. Lang.* 33, 511–526. doi: 10.1006/jmla.1994.1024

Greene, S. B., McKoon, G., and Ratcliff, R. (1992). Pronoun resolution and discourse models. *J. Exp. Psychol. Learn. Mem. Cogn.* 18, 266–283. doi: 10.1037/0278-7393.18.2.266

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychol. Rev.* 93, 411–428. doi: 10.1037/0033-295X.93.4.411

Horton, W. S., and Gerrig, R. J. (2005). Conversational common ground and memory processes in language production. *Discourse Process.* 40, 1–35. doi: 10.1207/s15326950dp4001_1

Just, M. A., Carpenter, P. A., and Woolley, J. D. (1982). Paradigms and processing in reading comprehension. *J. Exp. Psychol. Gen.* 111, 228–238. doi: 10.1037/0096-3445.111.2.228

Kennison, S. M. (2003). Comprehending the pronouns *her, him*, and *his*: implications for theories of referential processing. *J. Mem. Lang.* 49, 335–352. doi: 10.1016/S0749-596X(03)00071-8

Kintsch, W. (1988). The role of knowledge in discourse comprehension: a construction-integration model. *Psychol. Rev.* 95, 163–182. doi: 10.1037/0033-295X.95.2.163

Klin, C. M., Guzmán, A. E., Weingartner, K. M., and Ralano, A. S. (2006). When anaphor resolution fails: partial encoding of anaphoric inferences. *J. Mem. Lang.* 54, 131–143. doi: 10.1016/j.jml.2005.09.001

Klin, C. M., Weingartner, K. M., Guzman, A. E., and Levine, W. H. (2004). Readers' sensitivity to linguistic cues in narratives: how salience influences anaphor resolution. *Mem. Cognit.* 32, 511–522. doi: 10.3758/BF03195843

Levine, W. H., Guzmán, A. E., and Klin, C. M. (2000). When anaphor resolution fails. *J. Mem. Lang.* 43, 594–617. doi: 10.1006/jmla.2000.2719

Levine, W. H., and Hagaman, J. A. (2008). Negated concepts interfere with anaphor resolution. *Intercult. Pragmatics* 5, 471–500. doi: 10.1515/IPRG.2008.023

Lewis, R. L. (1996). Interference in short-term memory: the magical number two (or three) in sentence processing. *J. Psycholinguist. Res.* 25, 93–113. doi: 10.1007/BF01708421

Lewis, R. L., and Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cogn. Sci.* 29, 375–419. doi: 10.1207/s15516709cog0000_25

Loftus, G. R., and Masson, M. E. J. (1994). Using confidence intervals in within-subject designs. *Psychono. Bull. Rev.* 1, 476–490. doi: 10.3758/BF03210951

Lorch, R. F., and Myers, J. L. (1990). Regression analyses of repeated measures data in cognitive research. *J. Exp. Psychol. Learn. Mem. Cogn.* 16, 149–157. doi: 10.1037/0278-7393.16.1.149

Love, J., and McKoon, G. (2011). Rules of engagement: Incomplete and complete pronoun resolution. *J. Exp. Psychol. Learn. Mem. Cogn.* 37, 874–887. doi: 10.1037/a0022932

Lund, K., and Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behav. Res. Methods Instrum. Comput.* 28, 203–208. doi: 10.3758/BF03204766

Mason, R. A. (1997). *The Role of Multiple Antecedents in the Time Course of Anaphor Resolution*. Unpublished Master's thesis, University of Massachusetts, Amherst.

McElree, B. (2000). Sentence comprehension is mediated by content-addressable memory structures. *J. Psycholinguist. Res.* 29, 111–123. doi: 10.1023/A:1005184709695

Myers, J. L., and O'Brien, E. J. (1998). Accessing the discourse representation during reading. *Discourse Process.* 26, 131–157.

Myers, J. L., O'Brien, E. J., Balota, D. A., and Toyofuku, M. L. (1984). Memory search without interference: The role of integration. *Cogn. Psychol.* 16, 217–242. doi: 10.1016/0010-0285(84)90008-2

Nairne, J. (1990). A feature model of immediate memory. *Mem. Cognit.* 18, 251–269. doi: 10.3758/BF03213879

O'Brien, E. J., Plewes, P. S., and Albrecht, J. E. (1990). Antecedent retrieval processes. *J. Exp. Psychol. Learn. Mem. Cogn.* 16, 241–249.

O'Brien, E. J. (1987). Antecedent search processes and the structure of text. *J. Exp. Psychol. Learn. Mem. Cogn.* 13, 278–290. doi: 10.1037/0278-7393.13.2.278

Radvansky, G. A. (1998). The organization of information retrieved from situation models. *Psychon. Bull. Rev.* 5, 283–289. doi: 10.3758/BF03212952

Radvansky, G. A., Zwaan, R. A., Federico, T., and Franklin, N. (1998). Retrieval from temporally organized situations models. *J. Exp. Psychol. Learn. Mem. Cogn.* 24, 1224–1237. doi: 10.1037/0278-7393.24.5.1224

Ratcliff, R. (1978). A theory of memory retrieval. *Psychol. Rev.* 85, 59–108. doi: 10.1037/0033-295X.85.2.59

Reinhart, T. (1983). Coreference and bound anaphora: a restatement of the anaphora questions. *Linguist. Philos.* 6, 47–88. doi: 10.1007/BF00868090

Rohde, D. (2003). *Linger: A Flexible Program for Language Processing Experiments.* Available online at: http://tedlab.mit.edu/~dr/Linger/

Smithson, M. (2003). *Confidence Intervals: Quantitative Applications in the Social Sciences*, *No. 140*. Thousand Oaks, CA: Sage.

Sternberg, S. (1966). High-speed scanning in human memory. *Science* 153, 652–654. doi: 10.1126/science.153.3736.652

Sturt, P. (2003). The time-course of the application of binding constraints in reference resolution. *J. Mem. Lang.* 48, 542–562. doi: 10.1016/S0749-596X(02)00536-3

Townsend, J. T., and Fifíc, M. (2004). Parallel versus serial processing and individual differences in high-speed search in human memory. *Percept. Psychophys.* 66, 953–962. doi: 10.3758/BF03194987

Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

van den Broek, P., Rapp, D. N., and Kendeou, P. (2005). Integrating memory-based and constructionist processes in accounts of reading comprehension. *Discourse Process.* 39, 299–316. doi: 10.1080/0163853X.2005.9651685

Van Dyke, J. A., and McElree, B. (2006). Retrieval interference in sentence comprehension. *J. Mem. Lang.* 55, 157–166. doi: 10.1016/j.jml.2006.03.007

van Gompel, R. P. G., and Majid, A. (2004). Antecedent frequency effects during the processing of pronouns. *Cognition* 90, 255–264. doi: 10.1016/S0010-0277(03)00161-6

Wiley, J., Mason, R. A., and Myers, J. L. (2001). Accessibility of potential referents following categorical anaphors. *J. Exp. Psychol. Learn. Mem. Cogn.* 27, 1238–1249. doi: 10.1037/0278-7393.27.5.1238

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 22 February 2014; accepted: 10 July 2014; published online: 29 July 2014.
Citation: Autry KS and Levine WH (2014) A fan effect in anaphor processing: effects of multiple distractors. Front. Psychol. ***5****:818. doi: 10.3389/fpsyg.2014.00818*
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology.
Copyright © 2014 Autry and Levine. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 23 November 2015
doi: 10.3389/fpsyg.2015.01746

Backward- and Forward-Looking Potential of Anaphors

Petra B. Schumacher, Jana Backhaus and Manuel Dangl*

Department of German Language and Literature I, University of Cologne, Cologne, Germany

Personal pronouns and demonstratives contribute differently to the encoding of information in the mental model and they serve distinct backward- and forward-looking functions. While (unstressed) personal pronouns are the default means to indicate coreference with the most prominent discourse entity (backward-looking function) and typically mark the maintenance of the current topic, demonstratives are used to refer to a less prominent entity and serve the additional forward-looking function of signaling a possible topic shift. In Experiment 1, we present an ERP study that examines the time course of processing personal and d-pronouns in German (*er* vs. *der*) and assesses the impact of two prominence features of the antecedent, thematic role and sentential position, as well as neurophysiological correlates of backward- and forward-looking functions of referential expressions. We tested the comprehension of personal and d-pronouns following context sentences containing two potential antecedents. In addition to the factor pronoun type (*er* vs. *der*), we varied the verb type (active accusative verbs vs. dative experiencer verbs) and the thematic role order (canonical vs. non-canonical) in the context sentences to vary the antecedent's prominence. Time-locked to pronoun-onset, the ERPs revealed a general biphasic N400-Late Positivity for d-pronouns over personal pronouns with further subtle interactions of the prominence-lending cues in the early time window. The findings indicate that the calculation of the referential candidates' prominence (backward-looking function) is guided by thematic role and positional information. Thematic role information, in combination with initial position, thus represents a central predictor during referential processing. Coreference with a less prominent entity (assumed for d-pronouns) results in processing costs (N400). The additional topic shift signaled by d-pronouns (forward-looking function) results in attentional reorienting (Late Positivity). This is further supported by Experiment 2, a story continuation study, which showed that personal pronouns trigger topic maintenance, while d-pronouns yield topic shifts.

Keywords: pronoun resolution, prominence, agentivity, position, ERP, N400, Late Positivity, topic shift

OPEN ACCESS

Edited by:
Claudia Felser,
University of Potsdam, Germany

Reviewed by:
Katharina Spalek,
Humboldt-Universität zu Berlin, Germany
Hannah Rohde,
University of Edinburgh, UK

****Correspondence:***
Petra B. Schumacher
petra.schumacher@uni-koeln.de

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: *11 August 2015*
Accepted: *30 October 2015*
Published: *23 November 2015*

Citation:
Schumacher PB, Backhaus J and Dangl M (2015) Backward- and Forward-Looking Potential of Anaphors. Front. Psychol. 6:1746. doi: 10.3389/fpsyg.2015.01746

INTRODUCTION

When a language makes available different forms to refer to entities in the world, these forms typically indicate discrete cognitive states in the mental representation of the interlocutors (cf. Gundel et al., 1993). Accordingly, personal pronouns, demonstrative pronouns, definite noun phrases (NPs) or indefinite NPs serve distinct discourse pragmatic functions. In the following, we will focus on the contribution of personal and demonstrative pronouns to reference tracking.

While (unstressed) personal pronouns are the default means to indicate coreference with the most prominent entity in the current discourse, demonstrative pronouns are used to refer to a less prominent entity or exclude the most prominent entity (cf. Comrie, 1997). We refer to this as the "backward-looking function" of referential expressions. In addition, personal pronouns signal the maintenance of the current topic, while demonstratives suggest that the respective referent is likely to be promoted to topic status in subsequent discourse and thus indicate a topic shift (cf. e.g., Abraham, 2002). This is what we call the "forward-looking function."

Demonstratives come in pronominal (*this, that*) or adnominal form (*this teacher, that book*) and represent deictic expressions that mark the relative distance of the respective referent to the speaker, the hearer or both. Languages vary with regard to how many distance contrasts they encode and whether they only consider the speaker as the deictic center or allow for perspectival centers associated with other protagonists as well; for example English distinguishes the near *this* and the distant *that*, Spanish has a three-way contrast (proximal: *este*, medial: *ese*, distal: *aquel*), Hausa a four-way contrast (near speaker: *nân,* near hearer: *nan*, away from speaker and hearer: *cân*, far away from speaker and hearer: *can*), and some systems encode even more contrasts (e.g., Navajo, Malagasy; Diessel, 2013). German, the language under investigation in this study, employs the demonstrative pronouns *dieser, diese, dieses* (masculine, feminine, neuter) and the d-pronoun *der, die, das*. The former is more restricted in its referential choice and is claimed to prefer the last mentioned entity as its referential candidate, while the d-pronoun does not have such a local restriction (cf. e.g., Zifonun et al., 1997). A less commonly used form to mark distance is *jener, jene, jenes*, but German more frequently uses a modifying adverbial (*hier* "here," *da* "there") to mark distance contrasts.

In the current investigation, we compare the comprehension of the d-pronoun *der* with that of the personal pronoun *er* in contexts with two potential antecedents. The resolution preferences are generally discussed with reference to the notion of referential prominence, which assumes that referents that are accessible in the mental model are ranked in a particular order (cf. e.g., Grosz et al., 1995). But what is prominence? In the literature on pronoun resolution many different factors have been discussed as prominence-lending cues and in the following we provide a brief overview over possible candidate features assumed in the processing literature.

The most influential accounts that investigated personal and demonstrative pronoun resolution considered syntactic function and topicality to be prominence-lending features. Bosch and colleagues initially proposed that personal pronouns in German show a subject preference, while d-pronouns have an anti-subject preference (Bosch et al., 2003). Based on examples with clear discourse topics, they subsequently suggest that personal pronouns favor topical entities and d-pronouns follow an anti-topic interpretation strategy (Bosch and Umbach, 2007; Hinterwimmer, 2015). These accounts assume complementary interpretation preferences for personal and d-pronouns. By contrast, on the basis of data from Finnish, where the personal pronoun was preferably interpreted to refer to the subject while the demonstrative elicited a last-mention preference, Kaiser proposed a non-complementary form-specific distribution of interpretation preferences (Kaiser and Trueswell, 2008). Research on pronoun resolution has identified numerous other candidate factors, including among others linear order, animacy, focus, coherence relations and verb semantics (Stevenson et al., 1994; Chambers and Smyth, 1998; Järvikivi et al., 2005; Kehler et al., 2008; Ellert, 2010).

An alternative account of pronoun resolution is the Bayesian model which promotes a tight relationship between pronoun interpretation and production (Kehler et al., 2008; Kehler and Rohde, 2013). In this framework, interpretive preferences are not merely a function of the prominence structure of previous discourse but arise from the combination of prior expectations for subsequent mention and the production bias for a particular form. Behavioral research within this framework suggests that grammatical function or topichood influence the production bias while coherence relations impact which referent is expected. This approach thus assumes that prominence-lending cues feed into an intricate system of predictive processing that shapes expectation for a particular referent and considers production biases for a particular form. This line of research is promising, but in the current research we do not tease apart production biases and prior expectation. We assess the mechanisms underlying pronoun processing but future research should follow up on the Bayesian predictions within our experimental design.

The current research asks the question whether thematic function is a high ranked candidate for referential prominence. This is motivated by claims that agentivity is part of core cognitive architecture and shapes our thinking and cognitive development in fundamental ways (Leslie, 1995). According to this view, agents are cognitive attractors that hold certain causal properties, initiate actions, pursue goals, have sentience. This is reminiscent of the feature-based characterization of agentivity in semantic theories that attributes causation, volitionality, sentience, self-propelled movement and independent existence to prototypical agents (Dowty, 1991; Primus, 1999). These theories have proposed thematic role hierarchies on the basis of proto-roles, with the highest thematic role being the "proto-agent" and the lower one the "proto-patient." According to this view, agents are the prototypical exemplar of proto-agent because they hold many of the properties listed above but experiencers also satisfy features of proto-agents. Previous research on pronoun resolution has already pointed to the contribution of thematic role information by looking at verb semantics and animacy, and subject or topic preferences may be explained by agent preferences as well, since these features are often aligned.

To disentangle the effect of thematic role from grammatical function, we investigate reference resolution in the context of antecedent clauses with dative experiencer verbs, which critically cross these two predictors for prominence and have an agentive object (i.e., the experiencer) and a non-agentive subject. Example (2) illustrates this construction. In this example, the boxer is the experiencer and the one who must be sentient

based on possible verbal entailments about the argument; hence the object holds more proto-agent properties than the subject.

(1) Der Feuerwehrmann will den Jungen retten … Aber er/der hat …
The firefighter-NOM wants the boy-ACC rescue … But he/D-Pro has …
"The firefighter wants to rescue the boy … But he has…"

(2) Dem Boxer hat der Musiker imponiert … Aber er/der hat …
The boxer-DAT has the musician-NOM impressed …. But he/D-Pro has …
"The boxer was impressed by the musician … But he has …"

Different prominence-lending features of the referents introduced in the context sentences may be responsible for pronoun resolution preferences in these two examples. Crucially, the context sentences differ with respect to the adherence and alignment of the following prominence cues: (i) agentivity (proto-agent > proto-patient), (ii) grammatical function (subject > object) and (iii) topicality. Note that for topicality we assume that the initial argument of a sentence represents the aboutness-topic (cf. Reinhart, 1981). Thus, rather than considering first vs. second mention effects, we pursue a functional approach according to which first mentioned referents serve as topics. **Table 1** illustrates the prominence-lending features incorporated by the initial argument in the active accusative context (1) and the dative experiencer context (2) for canonical and non-canonical argument order. The possible candidate features agent, subject and topic are fully aligned in the canonical active accusative case. The dative experiencer conditions represent an alignment of two of these features and will help to disentangle the contribution of agentivity and subjecthood. Finally, the non-canonical active accusative condition shows even less alignment at the initial argument. If harmonic alignment at the initial position is a key to pronoun resolution, this condition should yield less clear preferences.

As an alternative to alignment of prominence-lending cues, one feature or a combination of features may affect pronoun resolution. For instance if thematic function is a decisive feature during pronoun resolution, this may be reflected in interpretive preferences irrespective of verb type and canonicity. If two or more features act jointly, fine differences should be observable when testing different verb types and canonicity effects. For example, if agentivity and topicality act together, the pronoun following the canonical dative-experiencer construction should link with its antecedent more easily than that following the non-canonical dative-experiencer context; if subjecthood and topicality collaborate, the non-canonical dative-experiencer antecedent clauses should yield clearer interpretive preferences than the canonical dative-experiencer contexts; etc.

Previous behavioral studies indicate a combination of partial feature alignment and the role of thematic function information. In offline tasks, agentivity has been shown to be a stronger predictor than subjecthood for pronoun resolution in German (Schumacher et al., 2016). Sentence completion and referent identification tasks with stimuli that contained either an antecedent clause with active accusative verbs ["rescue" in (1) where topic, subject and agent are aligned] or dative experiencer verbs ["be impressed" in (2) where the proto-agent, the xperiencer, is the object] revealed a proto-agent bias for the personal pronoun and an anti-agent bias for the d-pronoun in the canonical argument order of (1) and (2). When the argument order in the context clause was reversed, the active accusative verbs still registered an agent (or subject) preference—contra first mention or topic preference accounts of personal pronoun resolution—and an anti-agent (anti-subject) bias for the d-pronoun. Argument reversal of (2) resulted in chance performance for both types of pronouns suggesting that in this case the calculation of the relative ranking of the referential candidates was hampered. These data indicate that in a task in which participants are not under time pressure agentivity outweighs subjecthood when it is aligned with topic and/or subject—i.e., in the canonical accusatives (where all three cues are aligned), the canonical dative experiencers (where agent and topic are aligned), and the non-canonical accusatives (where agent and subject are aligned). This suggests that alignment of certain prominence-lending features is beneficial for pronoun resolution. In the case where the agent is not aligned with either topic or subject (the non-canonical dative experiencer contexts), the relative ranking of the referents seems to be too weak to generate an interpretive preference for either of the referential candidates. This reveals that interpretive preferences are not just a consequence of (partial) alignment of prominence-lending cues but that the weighting of these cues is also of relevance.

In the current research, Experiment 1 was designed to investigate the real-time consequences of the verb type × canonicity manipulation for pronoun resolution through event-related brain potentials (ERPs). We hypothesize that prominence-lending cues are used for the generation of fine-tuned predictions about upcoming entities. Personal pronoun resolution as a potential means to signal topic maintenance may thus proceed relatively effortless but could be encumbered in cases in which prominence cues are difficult to process, for example due to certain types of misalignment (as illustrated in **Table 1** and by the behavioral data). D-pronouns in turn require the exclusion of the most prominent referential candidate, which should result in processing costs. Based on previous ERP research, prediction errors—here assumed to be guided by prominence cues—should be reflected in a negative brain potential (N400; for an overview see Bornkessel-Schlesewsky and Schumacher, 2016). N400 effects have for instance been

TABLE 1 | Prominence features of first argument in context sentence.

	Active accusative verbs	Dative experiencer verbs
Canonical order	Agent and Subject and Topic	Agent and Topic
Non-canonical order	Topic	Subject and Topic

observed for referents of differing degrees of givenness—with given entities being more predictable than inferrables and new entities being the least expected (Burkhardt, 2006)—or as an indicator of the distance between anaphor and antecedent—with effects of first mention and recency across multiple sentences (Streb et al., 2004). Negative deflections have also been reported for referential ambiguity during pronoun resolution, which may indicate that a disambiguating referential form is expected in such cases (Nieuwland and Van Berkum, 2006).

With regard to the forward-looking function of demonstrative pronouns, psycholinguistic investigations have been sparse. It has been claimed that demonstratives have the potential to initiate a topic shift and promote their referent to topic status in later discourse. For example, Abraham (2002) explicitly describes the demonstrative as a topic shifter. Empirical evidence comes from the comparison of indefinite *this* ("this egg") and regular indefinites ("an egg"; cf. e.g., Gernsbacher and Shroyer, 1989; Chiriacescu, 2011). Using text continuation tasks, in which participants were instructed to continue a story with five sentences, these studies found that indefinite *this* elicited more mentions of the referent in the continuations, with less marked forms, and had a higher topic shift potential than the regular indefinite. The function of a demonstrative is thus not only to draw the attention to a less prominent discourse entity but also to signal the comprehender that the respective referent may become more prominent in subsequent discourse. Experiment 2 was conducted to investigate the topic shift potential of d-pronouns (and topic maintenance potential of personal pronouns) using a text continuation task. The assumed shift in attention furthermore is predicted to have consequences for discourse representation. Previous research on Japanese and Chinese, in which the notion of topic is crucial for sentence processing, suggests that topic-marked entities that trigger a shift in the ranking of discourse referents and hence require the updating of discourse representation structure evoke a Late Positivity (Hirotani and Schumacher, 2011; Hung and Schumacher, 2012, 2014; Wang and Schumacher, 2013). We therefore predict a Late Positivity for discourse updating due to the topic shift potential of d-pronouns in Experiment 1.

EXPERIMENT 1

The current experiment was designed to assess the online processing of d-pronouns and personal pronouns with a particular focus on contexts in which subject and agent were not aligned. We therefore tested active accusative and dative experiencer antecedent clauses with canonical and non-canonical argument order (see **Table 2** for sample stimuli). As described above, dative experiencer constructions were chosen because they allow us to disentangle the contribution of thematic and syntactic function to pronoun resolution. These verbs come with a dative experiencer (proto-agent in the frameworks of Dowty, 1991 and Primus, 1999) and a subject that represents the lower ranked role and have already shown robust effects of agentivity in behavioral tasks (Schumacher et al., 2016). Note also that we assume that the canonical argument order for these constructions is object before subject (cf. e.g., Haider, 1993; but see Footnote 1 in the Discussion for an alternative view).

Concerning backward-looking, the core function of a pronoun is to refer to an entity available in the mental representation. Hence upon encountering a pronominal expression, a dependency relation between the pronoun and its antecedent must be established. This is guided by the prominence structure of the referents from prior discourse, resulting in a ranked set of referential candidates. Accessibility theories suggest that the personal pronoun prefers the most prominent entity or the entity in focus, which has been attested by corpus research and psycholinguistic experiments (cf. e.g., Gordon et al., 1993; Gundel et al., 1993). Accordingly, personal pronoun resolution should generally proceed rather effortlessly. By contrast, resolution of the d-pronoun has been described to exclude the highest ranked referential candidate (cf. Comrie, 1997; Abraham, 2002). Such an operation should be resource-consuming. All other things being equal, processing the d-pronoun should thus be more costly than processing the personal pronoun. With respect to ERP signatures, we hypothesize that the backward-looking function is first of all closely tied to this form-function correlation interacting with predictive referential parsing reflected in an N400 effect. For predictive parsing, the d-pronoun as the more marked form should be generally more costly than the personal pronoun because it requires the exclusion of the most prominent referent.

This process may be further affected by the misalignment or weighting of prominence features that may encumber the establishment of a ranked set of referential candidates. The experimental design allows us to investigate the organization of the possible set of prominence-lending features and its impact on real-time processing. We thus predict subtle interactions of the factors verb type (varying the combination of grammatical and thematic roles) and canonicity (assigning different topics) on pronoun resolution. If alignment of topic, subject and/or agent is a key force during online pronoun resolution, the different alignments illustrated in **Table 1** may result in processing effort reflected by the N400 amplitude. Likewise the weighting of the different prominence-lending features may affect the processes underlying the N400.

With regard to the forward-looking function, the literature assumes that d-pronouns are topic shifters, which we argue has consequences for discourse updating. We therefore expect a Late Positivity effect for the d-pronoun relative to the personal pronoun. Previous research has not considered the role of prominence cues on forward-looking processes but misalignment of prominence features may result in failure to rank the referential candidates, which may well encumber forward-oriented processing.

TABLE 2 | Example stimuli for the ERP experiment.

Argument order	Sentence	Stimuli
VERB TYPE: ACCUSATIVE VERB		
Canonical	Context sentence	Der Feuerwehrmann | will | den Jungen | retten, | weil | das Haus | brennt. The firefighter-NOM wants the boy-ACC rescue because the house-NOM burns. *The firefighter wants to rescue the boy, because the house is burning.*
	Target sentence	Aber | **er/der** | ist | viel | zu | aufgeregt. But he/D-Pro is way too nervous. *But he is way too nervous.*
Non-canoncial	Context sentence	Den Jungen | will | der Feuerwehrmann | retten, | weil | das Haus | brennt. The boy-ACC wants the firefighter-NOM rescue because the house burns. *The firefighter wants to rescue the boy, because the house is burning.*
	Target sentence	Aber | **er/der** | ist | viel | zu | aufgeregt. But he/D-Pro is way too nervous. *But he is way too nervous.*
Verification Question	Correct answer "Yes"	Brennt das Haus? *Is the house burning?*
	Correct answer "No"	Wackelt das Haus? *Is the house shaking?*
VERB TYPE: DATIVE EXPERIENCER VERB		
Canonical	Context sentence	Dem Boxer | hat | der Musiker | imponiert, | und | zwar | schon | lange. The boxer-DAT has the musician-NOM impressed, in fact already long. *The boxer was impressed by the musician for a long time.*
	Target sentence	Aber | **er/der** | wollte | das | nicht | wahr | haben. But he/D-Pro wanted that not true have. *But he didn't want to accept it.*
Non-Canonical	Context sentence	Der Musiker | hat | dem Boxer | imponiert, | und | zwar | schon | lange. The musician-NOM has the boxer-DAT impressed, in fact already long. *The boxer was impressed by the musician for a long time.*
	Target sentence	Aber | **er/der** | wollte | das | nicht | wahr | haben. But he/D-Pro wanted that not true have. *But he didn't want to accept it.*
Verification question	Correct answer "Yes"	Imponierte der Musiker dem Boxer? *Has the musician impressed the boxer?*
	Correct answer "No"	Imponierte der Musiker dem Fechter? *Has the musician impressed the fencer?*

Methods

Participants

Twenty-seven right-handed, monolingually raised native speakers of German (14 women; mean age: 22; range 19–32) from the University of Mainz participated in this study after giving written informed consent. Participants had normal or corrected-to-normal vision and had no history of neurological or psychiatric disorders. The study was performed in accordance with the Declaration of Helsinki and with the national and institutional recommendations of the Neurolinguistics Lab at the Johannes Gutenberg-University Mainz. Data from three candidates were excluded from the ERP analysis due to excessive artifacts.

Materials

Sample stimuli for the eight conditions can be found in **Table 1**. The first sentence included two NPs that were masculine, animate and definite. In the accusative contexts, the canonical argument order was subject–object, and in the dative experiencer contexts, it was object–subject. Each of the context sentences was followed by a subordinate clause, which contained at most one gender-incongruent referent, to ensure that there was a proper distance between the NPs and the critical pronoun. The target sentence was always introduced by "but," followed by either the personal pronoun "er" or the d-pronoun "der." Sentence completions were kept referentially ambiguous. The material consisted of 60 accusative sets and 60 dative experiencer sets. Additionally, 60 filler sentence pairs were constructed, which included a masculine and feminine antecedent thus eliminating the ambiguity of the pronoun. Each participant was presented with 300 quasi-randomized test items: 240 critical items, consisting of 120 sentences with accusative verb and 120 with dative-experiencer verb, and all 60 fillers. Comprehension questions for each item served to assure that participants were paying attention to the stimuli. Correct and incorrect responses were evenly distributed across the stimuli. The incorrect comprehension questions targeted either an NP from the main clause, the action of the main clause or an element in the subordinate clause of the context sentence. For the filler items, the questions also referred to the content of the target sentence. See **Table 1** for example comprehension questions.

Procedure

During the experiment, each participant was seated in a dimly lit, sound-proof booth. Stimuli were presented visually on a computer screen placed about 100 cm in front of the participant with yellow letters against a dark blue background. Each trial began with a fixation star that was displayed for 500 ms in the center of the screen and followed by a blank screen for 150 ms. Each stimulus was presented in segments as indicated by the horizontal bars in **Table 1**. Single word segments were presented for a duration of 350 ms; phrases containing two or three words were presented for 400 or 450 ms, respectively. An interstimulus interval (ISI) of 150 ms was applied between segments. To verify that the participants had read and understood the sentences, each stimulus was followed by a yes/no verification question. After a blank screen of 150 ms, three question marks occurred for 500 ms, followed by the verification question which was presented in its entirety for 4000 ms. Participants were required to respond as quickly and accurately as possible by pressing a "yes" or "no" button on a gamepad. The assignment of the left and right response buttons was counterbalanced across participants. After the question, a blank screen was presented for 400 ms, followed by the next trial. Prior to the experimental run, participants completed a brief practice session to get acquainted with the experimental procedure.

EEG Recording and Preprocessing

The electroencephalogram (EEG) was recorded from 24 Ag/AgCl scalp electrodes and mounted in an elastic cap (*Easycap*, Munich, Germany). Electrode placement adhered to the international 10–20 system. The ground electrode was positioned at AFz. Electrodes were referenced to the left mastoid and re-referenced offline to linked mastoids. To account for artifacts resulting from eye movements, horizontal, and vertical eye movements were monitored by means of two sets of electrode pairs placed at the outer side of each eye for the horizontal electrooculogram (EOG) and above and below the participant's right eye for the vertical EOG. Electrode impedances were kept below 4 kΩ. All EEG and EOG channels were amplified with a *BrainAmp DC amplifier* (Munich, Germany) and digitized with a rate of 500 Hz.

Before averaging, the EEG data were band pass filtered offline with 0.3–20 Hz to remove unsystematic pre-stimulus differences caused by slow signal drifts. This filter has been identified as an appropriate filter for language-related research that overcomes certain drawbacks arising from baseline correction and has been applied by a number of research groups in previous years (e.g., Wolff et al., 2008; Schumacher and Hung, 2012; Kulakova et al., 2014). Next, automatic (set to ±40 μV for the EOG rejection criterion) and manual rejections were performed to exclude trials containing ocular, amplifier saturation, and other artifacts. Trials with incorrect answers or time-outs to the comprehension question were also excluded from the ERP data analysis. The application of all of these rejection criteria amounted to the exclusion of 12.55% of the data points. Average ERPs were time-locked to the onset of the critical pronoun in the target sentence.

Data Analysis

Statistical analyses were carried out by means of repeated measures analyses of variance (ANOVAs) and were performed with the factors PRONOUN (personal vs. d-pronoun), VERB (TYPE) (active accusative vs. dative experiencer) and CANON(ICITY) (canonical vs. non-canonical). Additionally, REGION OF INTEREST (ROI) entered the analysis as a factor. The analysis was carried out separately for midline and lateral electrode sites. The lateral electrodes were grouped by topographical ROIs which entered the analysis with four levels: left anterior (F3, F7, FC1, FC5), left posterior (CP1, CP5, P3, P7), right anterior (F4, F8, FC2, FC6), right posterior (CP2, CP6, P4, P8). The midline analysis included the six midline sites as levels (Fz, FCz, Cz, CPz, Pz, POz). All statistical analyses were based on the mean amplitude value per condition and were carried out in a hierarchical order. Huynh–Feldt adjustment was applied when the analysis involved factors with more than one degree of freedom in the numerator. The analyses were performed using the ez-package (Lawrence, 2013) in R (R Core Team, 2015).

Results

Figure 1 shows ERPs time-locked to the onset of the personal pronoun (in red) and the d-pronoun (in blue) collapsed over conditions. The plot reveals a negative maximum for the d-pronoun peaking around 300 ms after pronoun onset and a subsequent positive deflection for the d-pronoun between 450 and 600 ms. In addition, there were fine-grained differences arising from the contextual manipulation of verb type and canonicity. This is illustrated by **Figure 2** which shows that while the main effect of pronoun—i.e., a negativity around 300 ms followed by a positivity around 500 ms—is found for the two canonical conditions (top row), the non-canonical conditions (bottom row) diverge from the general picture. The non-canonical active accusatives shows no negativity for the d-pronoun over the personal pronoun and the non-canonical dative experiencer contexts seem to have evoked no positivity difference. With the exception of the last comparison, these observations were supported by statistical analyses. After visual inspection, two time windows were determined for the statistical analysis: 275–400 ms for the negativity effect and 450–600 ms for the positivity.

The statistical analysis for the 275–400 ms time window registered a main effect for PRONOUN over lateral electrode sites [$F_{(1,\ 23)} = 20.65, p < 0.001$] as well as over the midline electrodes [$F_{(1,\ 23)} = 211.43, p < 0.001$] and a four-way interaction for PRONOUN × VERB × CANON × ROI [lateral regions: $F_{(3,\ 69)} = 3.64, p < 0.05$; midline electrodes: $F_{(5,115)} = 3.86, p < 0.05$], reflecting the more pronounced negative deflection for the d-pronoun in comparison to the personal pronoun. Separate resolutions of these interactions for lateral and midline regions by region registered no topographical difference for the midline electrodes (and only main effects of PRONOUN over all midline electrodes) but the lateral ROI analysis indicated that the interaction was strongest over right anterior electrode sites [$F_{(1,\ 23)} = 4.60, p < 0.05$]. Subsequent resolution by the factor VERB within this ROI produced the following pattern: for the accusative verbs there was an interaction of PRONOUN ×

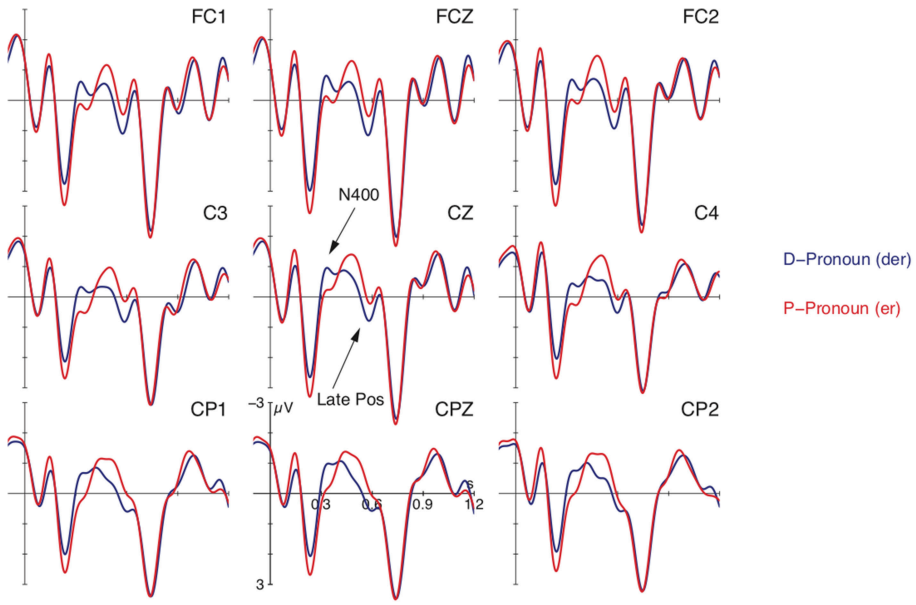


FIGURE 1 | Grand average ERPs at selected electrodes time-locked to the pronoun for the d-pronoun (blue) and the personal pronoun (red) averaged over both verb types and argument orders. Pronoun onset is at vertical bar. Negativity is plotted upwards.

CANON [$F_{(1, 23)} = 7.31$, $p < 0.01$] reflected in an effect of PRONOUN for the canonical subject-before-object order [$F_{(1, 23)} = 9.24$, $p < 0.01$] and no difference between the two types of pronouns in the non-canonical object-before-subject order [$F_{(1,23)} < 0.41$]. The dative experiencer verbs showed a main effect of PRONOUN [$F_{(1, 23)} = 4.87$, $p < 0.05$] and no interaction of PRONOUN × CANON [$F_{(1,23)} < 0.58$]. These patterns are illustrated by the pairwise comparisons in **Figure 2**.

For the time window between 450 and 600 ms, the analyses showed main effects for PRONOUN [lateral sites: $F_{(1, 23)} = 31.28$, $p < 0.001$; midline electrodes: $F_{(1,23)} = 23.87$, $p < 0.001$] and an interaction of PRONOUN × CANON [lateral: $F_{(1, 23)} = 8.06$, $p < 0.01$; midline: $F_{(1, 23)} = 4.69$, $p < 0.05$]. Resolution of this interaction by CANON showed an effect of PRONOUN for the canonical orders [lateral: $F_{(1, 23)} = 37.36$, $p < 0.001$; midline: $F_{(1, 23)} = 28.61$, $p < 0.001$] and a weaker effect of PRONOUN for the non-canonical orders [$F_{(1, 23)} = 7.65$, $p < 0.05$; midline: $F_{(1, 23)} = 5.67$, $p < 0.05$]. The effects reflect the more enhanced positivity for the d-pronoun over the personal pronoun.

Before turning to the discussion of how these findings inform pronoun resolution, we would like to address one further issue. We want to show that the observed effects for d-pronouns are not due to the ambiguity between the pronoun and the definite determiner in German. According to this, the processing costs registered for the demonstrative could also be caused by the ambiguity between the d-pronoun and the masculine definite determiner (both "der" in German). If costs were due to form ambiguity or anticipation of a noun following the determiner (for a sustained negativity for definite vs. indefinite determiners in German see Schumacher, 2009), additional (reanalysis) processes should be observable in the segment following the critical region. **Figure 3** spans until 2000 ms after pronoun onset and illustrates that no effects occurred in spill-over regions. The segment-wise presentation mode chosen in the current investigation may have been conducive to this as well because NPs were always presented in their entirety. We thus exclude form ambiguity as a potential explanation for the observed differences between personal and d-pronouns.

Discussion

In the discussion of the ERP study, we first focus on the general effects of pronoun in the two time windows before looking at the subtle interactions with the other factors in more detail.

Main Effect of Pronoun

Averaged over canonicity and verb type, the d-pronoun in comparison to the personal pronoun displayed a biphasic pattern with a more pronounced negativity in the early time window between 275 and 400 ms and an enhanced positivity in the later time window between 450 and 600 ms (see **Figure 1**). We propose that these two effects reflect backward- and forward-looking operations respectively.

The backward-looking function represents a core characteristic of a pronoun, which is referentially deficient and depends on an antecedent. We take the observed negativity (N400) for the d-pronoun as an indication for the more demanding processing of such a dependency relation on the basis of the instruction to exclude the most prominent referential candidate. The N400 for the d-pronoun patterns well with other findings from reference resolution that indicate that more computationally demanding anaphor-antecedent relations engender a negativity, including surface distance, semantic distance or referential ambiguity to name a few (Streb et al., 2004; Burkhardt, 2006; Nieuwland and Van Berkum, 2006).

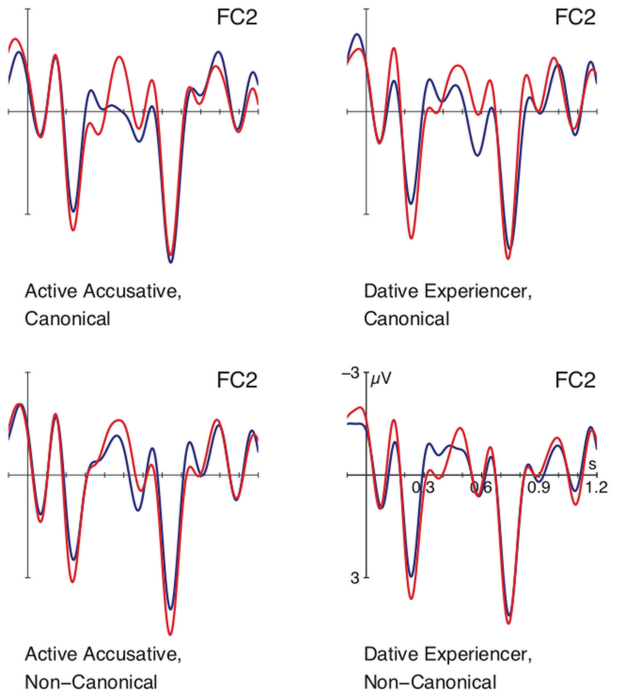


FIGURE 2 | Grand average ERPs for paired comparisons between the d-pronoun (blue) and the personal pronoun (red) at a selected right anterior electrode site (in which the four-way interaction was resolved). Pronoun onset is at vertical bar. Negativity is plotted up.

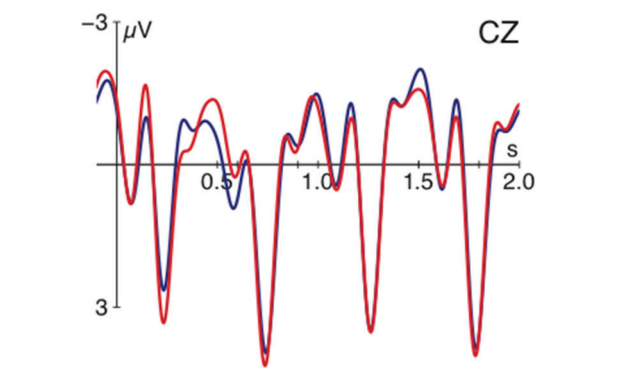


FIGURE 3 | Grand average ERPs at a selected electrode time-locked to pronoun-onset (at vertical bar) and spanning until 2000 ms later for the d-pronoun (blue) and personal pronoun (red) averaged over the factors verb type and argument order. Negativity is plotted upwards.

The current data with an enhanced N400 for d-pronouns over personal pronouns add to this view.

An alternative account for the observed cost would be that the d-pronoun is less expected than a personal pronoun (as a result of the information structural topic maintenance preference) and counters the particular prediction for an upcoming referent formed on the basis of prominence structure. Along these lines, the N400 has more generally been described as an expectation-driven process that is enlarged whenever a processing expectation is not met. However, the next section demonstrates that there are subtle interactions of the different prominence-lending cues manipulated in this study. Such findings indicate to us that the N400 for the d-pronoun reflects aspects associated with the prominence structure underlying the set of referential candidates (i.e., backward-looking operations). We assume that coreference relations depend on certain prominence features that govern the ranked set of referential candidates in the mental representation. Coreference with a less prominent entity (assumed for d-pronouns) results in processing costs.

The subsequent positivity (Late Positivity) for the d-pronoun over the personal pronoun is taken to reflect mental model updating costs. While a personal pronoun typically indicates the continuation of the current discourse topic, a d-pronoun signals a possible shift in attention toward a non-topical referent and therefore has a forward-oriented potential in providing cues about the changing (prominence) structure of the upcoming discourse (cf. e.g., Abraham, 2002). The d-pronoun further occurs in the topic position of the target sentence marking an interruption of the referential coherence. The processing of such forward-directed information exerts costs associated with the organization of discourse referents and the maintenance of the mental representation. Previous research on information structural influences on referential processing reported a Late Positivity for topic shift as well as contrastive focus (e.g., Hirotani and Schumacher, 2011; Wang and Schumacher, 2013; Hung and Schumacher, 2014). These information structural phenomena have in common that they can promote the cognitive status of their referents and direct the addressee's attention to a previously less attended referent. Behavioral data substantiate this role of topic and focus constituents (cf. Almor, 1999; Kaiser and Trueswell, 2004; Cowles et al., 2007). For the mental representation this implies that the prominence level of referents may shift dynamically and that any change may result in discourse updating costs. To substantiate these claims and assess whether d-pronouns affect the topic structure of subsequent discourse, we carried out Experiment 2 below.

Prominence Cues

When we look at the interaction of pronoun type with the two verb types and canonicity, subtle differences occur in particular with respect to processes in the N400 time window. Resolution of the four-way interaction revealed a more pronounced negativity for the d-pronoun over the personal pronoun in all conditions but the non-canonical active accusative antecedent contexts (see **Figure 2**). We take this to reflect processing differences associated with the computation of prominence, which seems to be most severely encumbered in the latter condition. This is best explained by the alignment based hypothesis (see **Table 1**): The four antecedent contexts differ with respect to their alignment of a number of potential prominence features, as illustrated by **Table 1**: (i) proto-agent > proto-patient, (ii) subject > object, and (iii) topic > non-topic (which we take to be a matter of sentence position). In the two canonical argument order cases, in which the proto-agent precedes the proto-patient, the underlying processes look much alike. As **Table 1** illustrates, all three prominence-lending cues are aligned to the first argument in the

canonical accusative contexts. The canonical dative experiencer contexts differ in that the initial topical argument is the agent but not the subject. This suggests that in this case of partial alignment, the absence of subjecthood does not have a negative impact on computation. And it indicates—in line with previous behavioral data (Schumacher et al., 2016)—that thematic role information represents a more highly ranked constraint during pronoun processing than grammatical function. Yet, grammatical function information still seems to contribute to pronoun comprehension to a certain extent because an N400 difference between personal and d-pronouns is still observed following the non-canonical dative experiencer contexts. In this case, the subject is aligned with the first position. Critically, in the non-canonical conditions, the active accusative condition diverges, which is the condition in which neither thematic role nor grammatical function information is aligned with the initial argument. This constellation apparently has real-time consequences for both personal and d-pronoun comprehension since the N400-morphology of both pronouns following the non-canonical active accusative contexts looks rather different from the other contexts. This suggests to us that the prominence-lending features made available by this particular context are not powerful enough to feed into prominence computation, encumbering coreference dependencies at this point in time.

Prominence computation—i.e., the calculation of a ranked set of referential candidates—thus seems to rely on the combination of weighted constraints over referential candidates. When agent or subject arguments occur in sentence-initial position, the resolution instruction ("corefer with the most prominent referential candidate" for the personal pronoun and "exclude the most prominent referential candidate" for the d-pronoun) can be executed, reflected in more computational demands for the exclusion of a referential candidate in the case of demonstratives. In situations in which the initial position is not aligned with either agent or subject (i.e., the non-canonical active accusative case), processing is hampered for both resolution instructions. This indicates that agents in first position are ideal candidates for referential prominence, regardless of grammatical function. When the first argument does not carry the highest thematic role, subjecthood of this argument enhances its referential status. This also indicates that initial position is one of the crucial cues contributing to referential processing (see e.g., Gernsbacher and Hargreaves, 1988). The first position typically hosts information structurally prominent entities, such as topics in German, which has led to proposals for topic and anti-topic pronominal resolution strategies, adding this feature to the prominence candidate set (Bosch et al., 2007; Hinterwimmer, 2015). One caveat arises from the unlicensed non-canonical argument order utilized in the current context sentences, which might well benefit from a richer context with an established discourse topic that paves the way for a marked argument linearization. In contextually enriched cases, prominence computation in non-canonical active accusative contexts may then be eased after all (cf. the research on information structural influences on argument linearization, e.g., Kaiser and Trueswell, 2004; Schumacher and Hung, 2012; Burmester et al., 2014).

Following our claim that the N400 reflects initial processes of executing the pronoun-specific linking instruction, one might ask how these data connect with the interpretive preferences obtained in previous offline studies (Schumacher et al., 2016). Similar to previous offline data that also tested the factors verb type and canonicity, the statistical analyses indicate more pronounced patterns for the canonical argument order (proto-agent > proto-patient) than for the non-canonical order. Yet the ERP data also differ partially from previous offline data in that the offline measures registered more interpretive insecurity in the non-canonical dative experiencer constructions, while the N400 patterns suggest that the non-canonical active accusative constructions are hampered. Certainly offline preferences may be influenced by additional factors and reflect more conscious and controlled operations. However, the differences between online and offline measures may also point out that the observed N400 effect reflects a more automatic process of prominence computation, which is calculated prior to referent selection[1]. A close look at **Figure 2** may even suggest a link between the Late Positivity and the offline data, where the non-canonical dative experiencers showed no positivity for the d-pronoun between 450 and 600 ms. This may be reflected by the Pronoun × Canonicity interaction in this time window, which yielded weaker effects for the two non-canoncial vs. the two canonical orders. However, the hierarchical analysis of the ERP data that we adopted does not allow us to test the non-canonical dative experiencer constructions in isolation. Since the coreference process is a discourse-internal operation, final resolution may well occur within the discourse-updating stage (cf. the two phases of bonding and resolution in e.g., Sanford and Garrod, 1989; Garrod and Terras, 2000). Coreference of personal pronouns is resolved effortlessly because the most prominent entity is maintained, while d-pronouns are more computationally demanding. Misalignments in the earlier prominence computation stage may then result in disruptive processing during discourse updating.

EXPERIMENT 2

In this study we wanted to test, whether d-pronouns have the capacity to initiate a topic shift, which would strengthen our account of the Late Positivity in Experiment 1. We employed a text continuation study, in which participants are provided with context-target sentence pairs and are asked to continue the story by writing six additional sentences. We then determined the topic constituent of each continuation sentence and calculated the topic shift potential of each pronoun, i.e., is the topic of the initial sentence maintained in the story sentences or is the other

[1]The ERP data may also be informative for a debate in the theoretical literature about the status of the dative experiencer linearizations. While we followed Haider (1993) among others who takes the dative-nominative order to be canonical, Barddal et al. (2014) argue that the two available argument orders alternate because both arguments carry certain subject features. This claim is supported by patterns of subject-verb inversion, covert realizations in control infinitives or reflexivization. While previous behavioral data (with uncertainty in the case of the nominative-dative order) did not strengthen this latter view, the N400 data show no order difference for the dative experiencer constructions.

referent promoted to topic status in subsequent discourse. This ties in with research that previously attested a larger amount of topic shifts for indefinite *this* relative to a regular indefinite NP (cf. Gernsbacher and Shroyer, 1989; Chiriacescu, 2011). Based on the claim that demonstratives are topic shifters (Abraham, 2002), we predict that the d-pronoun should show a higher capacity of topic shifting as the story unfolds, while the personal pronoun should encourage topic maintenance (cf. Grosz et al., 1995 for topic continuity expressed by the personal pronoun). Such a main effect of pronoun would substantiate the claim that the Late Positivity is associated with additional demands due to topic shifting, and based on the findings from Experiment 1 as well as the research literature, we predict more topic shift potential for all d-pronoun conditions irrespective of verb type and canonicity. Note however that there was a pronoun × canonicity interaction in the Late Positivity window in the ERP experiment which resulted from more pronounced effects in the canonical compared to the non-canonical conditions. Accordingly, non-canonical antecedent clauses—and in particular the non-canonical accusative contexts—which show misalignment of topic and agent may impede the dynamic updating of the discourse representation structure.

Methods

In this survey, participants were presented with context-target sentence pairs and were asked to continue the story by writing down six additional sentences.

Participants

Thirty-two native speakers of German (16 women; mean age: 25; range: 18–33 years), all monolingual, from the University of Cologne participated in this online survey. The investigation was performed in accordance with the Declaration of Helsinki and with the national and institutional recommendations of the Empirical Linguistics Lab at the University of Cologne.

Materials and Procedure

Four active accusative and four dative experiencer constructions were selected from Experiment 1 and each was presented in the four (canonicity × pronoun) versions. To reduce the number of given referents, only the main clause was used from the context sentence, followed by a target sentence with either a personal or a d-pronoun. The 32 critical items were distributed across 16 lists, so that each participant finished two items. Pilot research had shown that presenting more than two continuations is not recommendable.

Data Analysis

We wanted to find out which referent served as sentence topic in the continuation sentences. To this end we assume that the sentence-initial position holds the sentence topic (cf. aboutness-topic, Reinhart, 1981) and therefore determined whether the initial argument of each continuation reflected a shift or maintenance relative to the story-initial topic.

Each sentence of a continuation was coded with respect to whether it referred to the first or second NP in the context sentence or to another (new) referent that was introduced as part of the continuation. We only analyzed the first five (out of six) continuations, since in this task the last sentence often encourages a summary or wrap-up of the story line. Since we are interested in how the two referents from the initial sentence are picked up in subsequent sentences, reference to newly introduced entities were discarded prior to the analyses. Reference to the initial argument was coded as topic maintenance and reference to the second argument as topic shift. We first calculated the absolute frequency of topic shift and topic maintenance for the eight conditions. We further ran regression analyses with the predictors PRONOUN (personal pronoun; d-pronoun), VERB type (active accusative; dative experiencer) and CANON(ICITY) (canonical; non-canonical).

Results

Figure 4 depicts the difference scores determined from subtracting tokens of topic maintenance from tokens of topic shift. It is based on the cumulative absolute frequency of topic maintenance and topic shifts for the eight conditions. Positive values indicate more topic shifts, negative values reflect more topic maintenance. The figure illustrates that personal pronouns (in red) are more likely to maintain the sentence-initial topic—with the exception of the non-canonical active accusative condition—while d-pronouns (in blue) show a small but stable tendency for topic shift.

The regression analysis produced a final model that retained the entire set of effects and interactions. A test of this full model against a model reduced of interactions was statistically significant [likelihood ratio: $\chi^{2(4)} = 20.37$, $p < 0.001$]. As predicted the d-pronoun triggered more topic shifts than the personal pronoun. The analysis also showed that non-canonical constructions triggered more topic shifts than their canonical counterparts. As **Figure 4** indicates this effect of canonicity as well as the two-way interactions involving

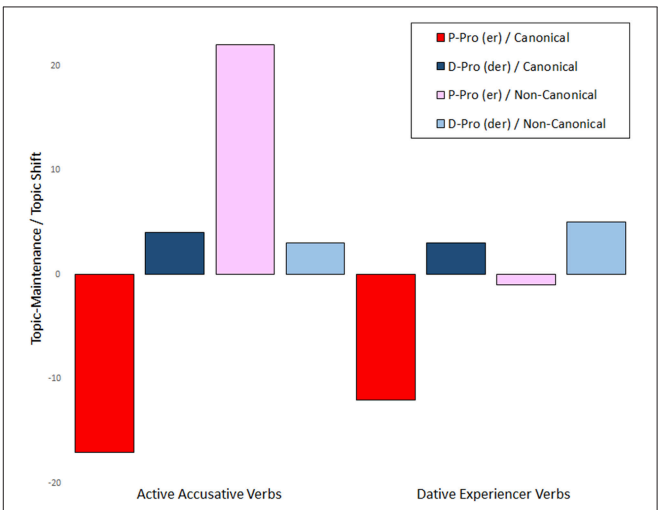


FIGURE 4 | Forward-directed potential of personal and d-pronouns in the eight conditions. Preference for topic shift is indicated by positive values (upwards) and for topic maintenance by negative values.

TABLE 3 | Regression analysis of Experiment 2.

Predictor	Beta	SE	*z*	*p*
Pronoun	1.47	0.55	2.66	0.008
Verb	0.31	0.60	0.53	0.598
Canonicity	3.10	0.69	4.51	0.000
Pronoun * Verb	−0.37	0.78	−0.48	0.633
Pronoun * Canonicity	−3.15	0.85	−3.70	0.000
Canonicity * Verb	−2.26	0.89	−2.55	0.011
Pronoun * Verb * Canonicity	2.46	1.14	2.16	0.031

canonicity (CANON × PRONOUN and CANON × VERB) and the three-way interaction CANON × VERB × PRONOUN are mainly driven by the unexpected pattern registered for the personal pronoun following the non-canonical active accusative condition. These interactions are reflected by the following patterns: While the d-pronouns show robust topic shift across conditions, personal pronouns in non-canonical antecedent clauses diverge from the topic maintenance observed in the canonical contexts. Active accusative contexts diverge immensely in this regard and even show a large amount of topic shift, while personal pronouns in non-canonical dative experiencer contexts registered only the smallest number of topic maintenance. **Table 3** reports the respective coefficients for the topic shift potential with the reference levels "er" for the factor pronoun, "accusative" for verb type and "canonical" for canonicity.

Discussion

The findings of this text continuation experiment confirm that the different pronouns serve discrete forward-looking functions. They show that the d-pronoun triggers more topic shifts in subsequent discourse than the personal pronoun. This supports previous research on the forward potential of indefinite demonstratives in English and German (cf. Gernsbacher and Shroyer, 1989; Chiriacescu, 2011). The personal pronoun in turn typically prompts topic continuations. The topic shift preference of the d-pronoun corroborates our proposal that the Late Positivity observed in the ERP study is associated with forward-directed signals that are encoded in discourse representation.

Based on these forward-oriented functions, the results for the personal pronoun in the non-canonical antecedent clauses suggest an interplay of prominence computation and discourse updating potential. In particular the pattern observed for the personal pronoun in the non-canonical active accusative constructions is surprising but it also emulates the exceptional role of this condition in Experiment 1, where we argued that the fact that neither proto-agent nor subject are aligned with the first position interferes with prominence computation. This seems to have far reaching consequences for subsequent discourse, where speakers possibly opt for an alternative strategy or even reset their mental representation and pick up the last mentioned referent making this the most prominent one (which results in topic shifts in Experiment 2).

GENERAL DISCUSSION

This research supports a dissociation of backward- and forward-looking functions for pronouns and reveals discrete patterns for personal and d-pronouns. The ERP data indicate a discrete time-course of the two functions and the text continuation data strengthen the account that d-pronouns are more likely to initiate a topic shift, while personal pronouns support topic maintenance.

Backward-looking Function

Overall, the current findings call for a resolution algorithm that considers multiple weighted prominence cues. Centering Theory (CT; Grosz et al., 1995) has served as a solid basis for numerous investigations of pronoun resolution. It assumes that certain referents of an utterance are more central than others, which, in turn, affects the processing of the subsequent utterance. Furthermore, personal pronouns are claimed to be preferably resolved toward the most central referential entity, which is understood as a means to establish coherence (Abraham, 2002). Within the CT framework, every utterance may contain several entities that have the potential to establish coherence with the following utterance. These referential expressions are called "Forward-looking Centers" (Cfs) and are ranked according to prominence features, whereby the highest ranked Cf of an utterance is referred to as "Preferred Center" (Cp). To determine if and how coherent two subsequent utterances are, CT offers an algorithm based on two parameters: the cognitive state of the "Backward-looking Center" (Cb), that is the element that picks up the highest ranked Cf from the previous utterance—ideally the Cp—and the current Cb's relation to the Cb of the previous utterance: either the Cb remains the same (Continue or Retain relations) or the Cb changes across two utterances (Smooth or Rough Shift relations; Brennan et al., 1987). Based on pronoun resolution in English, the ranking of the Cfs has been framed according to grammatical function (subject > object > other). Cross-linguistic comparisons however indicate that the setup of prominence cues is subject to language-specific constraints. Research on Japanese and German suggests that information structural notions contribute to the centering algorithm as well which has led to expansion of the grammatical function hierarchy (e.g., for Japanese: topic > empathy > subject > object > other; Kameyama, 1985; Walker et al., 1994, 1998; Di Eugenio, 1998; Abraham, 2002; Speyer, 2007).

While the application of the modified hierarchy may to a certain extent account for utterances with accusative verbs, it does not predict the proto-agent-preference observed for the dative experiencer verbs. We therefore propose to include proto-agentivity as a high-ranking constraint for the Cf ordering in German (proto-agent > proto-recipient > proto-patient; cf. e.g., Dowty, 1991; Primus, 1999). This shift from the grammatical function to the thematic role hierarchy does not affect the results for the canonical sentences with accusative verbs since the highest Cf is also the subject, but it serves to explain the preferences observed for the dative experiencer verbs in which subject and agent are assigned to distinct referents. Due to the non-canonical linearizations, we further

suggest to consider information structural notions as suggested previously on the basis of data from German and Japanese (cf. e.g., Walker et al., 1994; Abraham, 2002; Speyer, 2007). In particular, positional cues in the antecedent clause mark additional information status, with initial entities signaling topic status or contrast. In our case this information structural function may be weakened by the contextually unmotivated placement of a discourse-new object in initial position of the context sentence. But nevertheless first position in combination with other prominence-lending cues provides important information for prominence computation. The current data thus suggest an intricate interaction of agentivity, information structure, and subjecthood, which needs to be tested in more elaborate discourse contexts in future research. Furthermore, CT typically considers only the set of Cfs from the previous utterance; yet, larger discourse structure should be incorporated into CT algorithms. To summarize the backward-looking processes, the data indicate that the thematic role cue is tied to positional information, i.e., agents in initial position are the best candidates for prominence in the current study. In cases, where agents are not aligned with the initial position, grammatical function information collaborates with positional information to boost referential prominence.

Finally, a CT-like algorithm should also account for the resolution of demonstratives. In particular, resolution processes should exclude the Cp as a potential antecedent for the d-pronoun. In the absence of evidence to the contrary, this assumes that personal and d-pronouns in German make use of the same constraints over prominence structure (contra Kaiser's claims of form-specific constraints in Finnish, see Kaiser and Trueswell, 2008). In this regard, the Cp holds an important function within the referential space, which singles it out from the set of referential candidates.

Forward-looking Function

While demonstratives have been described as topic shifters, this forward-directed potential of referential expressions has been neglected in the research literature to a large extent (with the notable exceptions of Gernsbacher and Shroyer, 1989; Chiriacescu, 2011). To our knowledge, the continuation data from Experiment 2 represent the first test of the predictive potential of d-pronouns. They show that personal and d-pronouns influence the structure of subsequent discourse in different ways, yielding more topic maintenance and more topic shift respectively. This forward function of the pronouns can be regarded as a signal-driven cue whereby the d-pronoun promotes attention reorienting toward a new topic.

This finding thus strengthens our account of the Late Positivity in Experiment 1 as a marker of mental model updating triggered by the d-pronoun's inherent instruction to change the overall topic structure. Based on previous ERP research, we predicted a positive deflection for topic shift and attention orienting more generally, which is supported by the main effect of pronoun in the later time window with a more pronounced positive deflection for the d-pronoun relative to the personal pronoun. This suggests that the forward-looking function has real-time consequences during processing.

As far as the difficulties with the non-canonical active accusative contexts are concerned, the behavioral and ERP data converge. While the online data show no difference between the two pronouns in the N400 window—in contrast to all the other conditions—which we attributed to weak cues for prominence computation, this condition also diverges for the discourse continuation behavior by showing a surprising topic shift preference. This suggests that forward-oriented processing may be affected by the prominence structure of the preceding discourse. In the case where neither agent nor subject align with the first position, the relevant ranking of the referents seems to be destabilized hampering the typical forward potential of the personal pronoun.

CONCLUSION

The current investigation revealed differences in the time course of the resolution of personal and d-pronouns, reflected by a biphasic N400—Late Positivity pattern. We suggest that the N400 effect manifests an automatic operation of prominence computation that feeds into the pronoun-specific resolution instruction ("corefer with vs. exclude the most prominent referential candidate"). This early process is further influenced by verb specific information and word order, where the cooccurrence of agentivity and initial position yields an ideal candidate for referential prominence in German but prominence calculation may also be aggravated when particular prominence-lending cues are not aligned. The Late Positivity displays a discourse-internal updating process that provides cues for the possible change in prominence structure of the upcoming discourse, which is also supported by the story continuation task.

ACKNOWLEDGMENTS

We would like to thank Anika Joedicke and Flora Bastian for their assistance during data collection as well as Yu-Chen Hung and Hanna Weiland-Breckle for their support during data analysis.

REFERENCES

Abraham, W. (2002). Pronomina im Diskurs: Deutsche Personal- und Demonstrativpronomina unter 'Zentrierungsperspektive.' Grammatische Überlegungen zu einer Teiltheorie der Textkohärenz. *Sprachwissenschaft* 27, 447–491.

Almor, A. (1999). Noun-phrase anaphora and focus: the informational load hypothesis. *Psychol. Rev.* 106, 748–765. doi: 10.1037/0033-295X.106.4.748

Barddal, J., Eythorsson, T., and Dewey, T. K. (2014). Alternating predicates in Icelandic and German: a sign-based construction grammar account. *Work. Pap. Scand. Syntax* 93, 50–101. Available online at: http://hdl.handle.net/1854/LU-5787362

Bornkessel-Schlesewsky, I., and Schumacher, P. B. (2016). "Towards a neurobiology of information structure," in *The Oxford Handbook of Information Structure,* eds C. Féry and S. Ishihara (Oxford: Oxford University Press).

Bosch, P., Katz, G., and Umbach, C. (2007). "The non-subject bias of German demonstrative pronouns," in *Anaphora in Text: Cognitive, Formal and Applied Approaches To Anaphoric Reference,* eds M. Schwarz-Friesel, M. Consten, and M. Knees (Amsterdam; Philadelphia, PA: John Benjamins), 145–164. doi: 10.1075/slcs.86.13bos

Bosch, P., Rozario, T., and Zhao, Y. (2003). "Demonstrative pronouns and personal pronouns. German der vs. er," in *Proceedings of the EACL 2003* (Budapest).

Bosch, P., and Umbach, C. (2007). Reference determination for demonstrative pronouns. *ZAS Pap. Linguist.* 48, 39–51. Available online at: http://www.zas.gwz-berlin.de/185.html

Brennan, S. E., Friedman, M. W., and Pollard, C. J. (1987). "A centering approach to pronouns," in *Proceedings of the 25th Annual Meeting on Association for Computational Linguistics* (Stroudsburg, PA). doi: 10.3115/981175.981197

Burkhardt, P. (2006). Inferential bridging relations reveal distinct neural mechanisms: evidence from event-related brain potentials. *Brain Lang.* 98, 159–168. doi: 10.1016/j.bandl.2006.04.005

Burmester, J., Spalek, K., and Wartenburger, I. (2014). Context updating during sentence comprehension: the effect of aboutness topic. *Brain Lang.* 137, 62–76. doi: 10.1016/j.bandl.2014.08.001

Chambers, C. G., and Smyth, R. (1998). Structural parallelism and discourse coherence: a test of centering theory. *J. Mem. Lang.* 39, 593–608. doi: 10.1006/jmla.1998.2575

Chiriacescu, S. (2011). "Effects of reference form on frequency of mention and rate of pronominalization," in *Anaphora Processing and Application,* eds I. Hendrickx, S. Lalitah Devi, A. Branco, and R. Mitkov (Heidelberg: Springer), 132–143. doi: 10.1007/978-3-642-25917-3_12

Comrie, B. (1997). "Pragmatic binding demonstratives as anaphors in Dutch," in *Proceedings of the 23rd Annual Meeting of the Berkeley Linguistics Society: General Session and Parasession on Pragmatics and Grammatical Structure,* Vol. 23, eds M. L. Juge and J. L. Moxley (Berkeley, CA: Berkeley Linguistics Society), 50–61.

Cowles, H. W., Walenski, M., and Kluender, R. (2007). Linguistic and cognitive prominence in anaphor resolution: topic, contrastive focus and pronouns. *Topoi Int. Rev. Philos.* 26, 3–18. doi: 10.1007/s11245-006-9004-6

Diessel, H. (2013). "Distance contrasts in demonstratives," in *The World Atlas of Language Structures Online,* eds M. S. Dryer and M. Haspelmath (Leipzig: Max Planck Institute for Evolutionary Anthropology). Available online at: http://wals.info/chapter/41

Di Eugenio, B. (1998). "Centering in Italian," in *Centering Theory in Discourse,* eds M. A. Walker, A. K. Joshi, and E. F. Prince (Oxford: Clarendon Press), 115–138.

Dowty, D. (1991). Thematic proto-roles and argument selection. *Language* 67, 547–619. doi: 10.1353/lan.1991.0021

Ellert, M. (2010). *Ambiguous Pronoun Resolution in L1 and L2 Dutch*. Ph.D. dissertation, Radboud University Nijmegen.

Garrod, S., and Terras, M. (2000). The contribution of lexical and situational knowledge to resolving discourse roles: bonding and resolution. *J. Mem. Lang.* 42, 526–544. doi: 10.1006/jmla.1999.2694

Gernsbacher, M. A., and Hargreaves, D. J. (1988). Accessing sentence participants: the advantage of first mention. *J. Mem. Lang.* 27, 699–717. doi: 10.1016/0749-596X(88)90016-2

Gernsbacher, M. A., and Shroyer, S. (1989). The cataphoric use of the indefinite this in spoken narratives. *Mem. Cognit.* 17, 536–540. doi: 10.3758/BF03197076

Gordon, P. C., Grosz, B. J., and Gilliom, L. A. (1993). Pronouns, names, and the centering of attention in discourse. *Cogn. Sci.* 17, 311–347. doi: 10.1207/s15516709cog1703_1

Grosz, B. J., Joshi, A. K., and Weinstein, S. (1995). Centering: a framework for modeling the local coherence of discourse. *Comput. Linguist.* 21, 203–225.

Gundel, J. K., Hedberg, N., and Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language* 69, 274–307. doi: 10.2307/416535

Haider, H. (1993). *Deutsche Syntax Generativ*. Tübingen: Narr.

Hinterwimmer, S. (2015). "A unified account of the properties of German demonstrative pronouns," in *Proceedings from the Workshop on Pronominal Semantics at NELS* (Amherst, MA: University of Massachusetts).

Hirotani, M., and Schumacher, P. B. (2011). Context and topic marking affect distinct processes during discourse comprehension in Japanese. *J. Neurolinguist.* 24, 276–292. doi: 10.1016/j.jneuroling.2010.09.007

Hung, Y.-C., and Schumacher, P. B. (2012). Topicality matters: position-specific demands on Chinese discourse processing. *Neurosci. Lett.* 511, 59–64. doi: 10.1016/j.neulet.2012.01.013

Hung, Y.-C., and Schumacher, P. B. (2014). Animacy matters: ERP evidence for the multi-dimensionality of topic-worthiness in Chinese. *Brain Res.* 1555, 36–47. doi: 10.1016/j.brainres.2014.01.046

Järvikivi, J., van Gompel, R. P. G., Hyönä, J., and Bertram, R. (2005). Ambiguous pronoun resolution contrasting the first-mention and subject-preference accounts. *Psychol. Sci.* 16, 260–264. doi: 10.1111/j.0956-7976.2005.01525.x

Kaiser, E., and Trueswell, J. C. (2004). The role of discourse context in the processing of a flexible word-order language. *Cognition* 94, 113–147. doi: 10.1016/j.cognition.2004.01.002

Kaiser, E., and Trueswell, J. C. (2008). Interpreting pronouns and demonstratives in Finnish: evidence for a form-specific approach to reference resolution. *Lang. Cogn. Process.* 23, 709–748. doi: 10.1080/01690960701771220

Kameyama, M. (1985). *Zero Anaphora: The Case of Japanese.* Ph.D. dissertation, Stanford University.

Kehler, A., Kertz, L., Rohde, H., and Elman, J. L. (2008). Coherence and coreference revisited. *J. Semant.* 25, 1–44. doi: 10.1093/jos/ffm018

Kehler, A., and Rohde, H. (2013). A probabilistic reconciliation of coherence-driven and centering-driven theories of pronoun interpretation. *Theor. Linguist.* 39, 1–37. doi: 10.1515/tl-2013-0001

Kulakova, E., Freunberger, D., and Roehm, D. (2014). Marking the counterfactual: ERP evidence for pragmatic processing of German subjunctives. *Front. Hum. Neurosci.* 8:548. doi: 10.3389/fnhum.2014.00548

Lawrence, M. A. (2013). *ez: Easy Analysis and Visualization of Factorial Experiments*. R package version 4.2-2. Available online at: http://CRAN.R-project.org/package=ez

Leslie, A. M. (1995). "A theory of agency," in *Causal Cognition. A Multidisciplinary Debate*, eds D. Sperber, D. Premack, and A. J. Premack (Oxford: Clarendon Press), 121–141.

Nieuwland, M. S., and van Berkum, J. J. A. (2006). Individual differences and contextual bias in pronoun resolution: evidence from ERPs. *Brain Res.* 1118, 155–167. doi: 10.1016/j.brainres.2006.08.022

Primus, B. (1999). *Cases and Thematic Roles: Ergative, Accusative and Active*. Tübingen: Niemeyer. doi: 10.1515/9783110912463

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Reinhart, T. (1981). Pragmatics and linguistics: an analysis of sentence topics. *Philosophica* 27, 53–94.

Sanford, A. J., and Garrod, S. C. (1989). What, when, and how? Questions of immediacy in anaphoric reference resolution. *Lang. Cogn. Process.* 4, 235–263. doi: 10.1080/01690968908406369

Schumacher, P. B. (2009). Definiteness marking shows late effects during discourse processing: evidence from ERPs. *Lect. Notes Artif. Intell.* 5847, 91–106. doi: 10.1007/978-3-642-04975-0_8

Schumacher, P. B., Dangl, M., and Uzun, E. (2016). "Thematic role as prominence cue during pronoun resolution in German," in *Empirical Perspectives on Anaphora Resolution: Information Structural Evidence in the Race for Salience,* eds A. Holler, C. Göb, and K. Suckow (Berlin: de Gruyter).

Schumacher, P. B., and Hung, Y.-C. (2012). Positional influences on information packaging: insights from topological fields in German. *J. Mem. Lang.* 67, 295–310. doi: 10.1016/j.jml.2012.05.006

Speyer, A. (2007). Die Bedeutung der Centering Theory für Fragen der Vorfeldbesetzung im Deutschen. *Z. Sprachwiss.* 26, 83–115. doi: 10.1515/ZFS.2007.011

Stevenson, R. J., Crawnley, R. A., and Kleinman, D. (1994). Thematic roles, focus and the representation of events. *Lang. Cogn. Process.* 9, 519–548. doi: 10.1080/01690969408402130

Streb, J., Hennighausen, E., and Rösler, F. (2004). Different anaphoric expressions are investigated by event-related brain potentials. *J. Psycholinguist. Res.* 33, 175–201. doi: 10.1023/B:JOPR.0000027961.12577.d8

Walker, M. A., Iida, M., and Cote, S. (1994). Japanese discourse and the process of centering. *Comput. Linguist.* 20, 193–232.

Walker, M. A., Joshi, A. K., and Prince, E. F. (1998). *Centering Theory in Discourse*. Oxford: Clarendon Press.

Wang, L., and Schumacher, P. B. (2013). New is not always costly: evidence from online processing of topic and contrast in Japanese. *Front. Psychol.* 4:363. doi: 10.3389/fpsyg.2013.00363

Wolff, S., Schlesewsky, M., Hirotani, M., and Bornkessel-Schlesewsky, I. (2008). The neural mechanisms of word order processing revisited: electrophysiological evidence from Japanese. *Brain Lang.* 107, 133–157. doi: 10.1016/j.bandl.2008.06.003

Zifonun, G., Hoffmann, L., Strecker, B., Ballweg, J., Brauße, U., Breindl, E., et al. (1997). *Grammatik der deutschen Sprache*. Berlin: Walter de Gruyter.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2015 Schumacher, Backhaus and Dangl. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 27 June 2014
doi: 10.3389/fpsyg.2014.00630

Immediate sensitivity to structural constraints in pronoun resolution

***Wing-Yee Chow*[1,2]*†, *Shevaun Lewis*[1,3†] *and Colin Phillips*[1,4]**

[1] *Department of Linguistics, University of Maryland, College Park, MD, USA*
[2] *Basque Center on Cognition, Brain and Language, Donostia – San Sebastián, Spain*
[3] *Department of Cognitive Science, Johns Hopkins University, Baltimore, MD, USA*
[4] *Program in Neuroscience and Cognitive Science, University of Maryland, College Park, MD, USA*

Edited by:
Claudia Felser, University of Potsdam, Germany

Reviewed by:
Charles Jr. Clifton, University of Massachusetts Amherst, USA
Clare Patterson, Universität Potsdam, Germany

****Correspondence:***
Wing-Yee Chow, Basque Center on Cognition, Brain and Language, Paseo Mikeletegi 69, 2nd Floor, Donostia – San Sebastián, 20009, Spain
e-mail: wingyeechow.zoey@gmail.com

† These authors have contributed equally to this work.

Real-time interpretation of pronouns is sometimes sensitive to the presence of grammatically-illicit antecedents and sometimes not. This occasional sensitivity has been taken as evidence that structural constraints do not immediately impact the initial antecedent retrieval for pronoun interpretation. We argue that it is important to separate effects that reflect the initial antecedent retrieval process from those that reflect later processes. We present results from five reading comprehension experiments. Both the current results and previous evidence support the hypothesis that agreement features and structural constraints immediately constrain the antecedent retrieval process for pronoun interpretation. Occasional sensitivity to grammatically-illicit antecedents may be due to repair processes triggered when the initial retrieval fails to return a grammatical antecedent.

Keywords: pronoun resolution, Principle B, memory retrieval, self-paced reading, eye-tracking

INTRODUCTION

This paper is concerned with how different kinds of linguistic constraints are used in memory retrieval processes in the course of real-time comprehension. We focus on third-person pronouns for two reasons. First, the interpretation of such pronouns almost always requires the identification of an antecedent from the previous discourse, so they reliably trigger memory retrieval processes in comprehension. Second, the dependency between a pronoun and its antecedent is subject to several kinds of linguistic constraints. Thus, the outcome of the antecedent retrieval process is potentially quite informative about whether the memory system is able to take advantage of different kinds of linguistic constraints to aid sentence processing.

We consider two broad types of constraints on pronominal dependencies. *Agreement* constraints require that the pronoun and its antecedent share certain features, such as number, person, gender, and animacy. For example, in (1), 'Mary' cannot be the antecedent for 'him' because it mismatches the pronoun in gender. *Structural* constraints require that the antecedent bear certain relations to the pronoun in the syntactic and discourse representations. We focus on the structural constraint known as Binding Principle B (Chomsky, 1981): roughly, a pronoun cannot be bound by an antecedent within its local clause. In (1), 'Peter' cannot be the antecedent for 'him' because it would bind the pronoun from within the local clause.

(1) Bill explained to Mary that Peter had deceived him.

Note that for the purposes of this paper, it is sufficient to understand Principle B as a descriptive generalization. We are only concerned with which elements are potential antecedents, and which are not. This approach therefore abstracts away from questions about how the distribution and interpretation of pronouns should be explained at the syntactic, semantic, and pragmatic level (Reinhart, 1983; Grodzinsky and Reinhart, 1993). We are also restricting ourselves to pronouns with intrasentential antecedents, and thus will not be considering the role of various discourse-level structural constraints on extrasentential antecedents.

The combination of agreement and structural constraints substantially narrows the field of potential antecedents for a pronoun. In (1), for example, three entities are mentioned before 'him,' but only 'Bill' is a possible antecedent. Thus, it seems that an efficient comprehension system would take advantage of all available constraints as soon as possible—in the initial retrieval of an antecedent. However, antecedent retrieval relies on memory processes, and it is by no means guaranteed that the human memory system is capable of using any and all linguistic constraints to restrict retrieval. It is therefore an open empirical question which kinds of constraints are used immediately in the initial retrieval, and which have their effect later, as "filters" on the results of the retrieval.

Previous research has demonstrated that comprehenders are sensitive to agreement constraints very early in the process of pronoun interpretation. They rapidly and accurately identify feature-matching entities in the discourse (Arnold et al., 2000)

and detect feature mismatches between a pronoun and its supposed antecedent (Osterhout and Mobley, 1995; Carreiras et al., 1996; Van Berkum et al., 2004). Based on these findings, we consider it uncontroversial that the initial antecedent retrieval process takes advantage of agreement features to restrict the set of candidates.

On the other hand, while many previous studies have examined the role of Principle B in real-time pronoun interpretation, their results have led to divergent conclusions (see Nicol and Swinney, 2003 and Sturt, 2013 for reviews). While existing accounts differ on many aspects, one critical point of contention among them concerns whether and how structural constraints impact the initial antecedent-retrieval processes. Some have argued that only structurally acceptable potential antecedents are considered during the early stages of processing (e.g., Nicol and Swinney, 1989; Clifton et al., 1997, 1999; Lee and Williams, 2008; Patterson et al., 2014), while others have contended that structurally unacceptable candidates can be retrieved initially if they match the pronoun in features (e.g., Badecker and Straub, 2002; Kennison, 2003; Runner et al., 2006). We hope to clarify the kind of evidence that would be necessary to support each of these alternatives by distinguishing different forms of sensitivity to structurally unacceptable potential antecedents. We argue that the apparently conflicting results from previous studies, as well as the results from our own studies, are all consistent with the simultaneous use of agreement and structural constraints in the initial antecedent retrieval process.

We abstract away from other differences among existing accounts and consider two competing hypotheses that differ minimally. The *Agreement First* hypothesis is that the initial retrieval process uses agreement features, but not Principle B, to restrict the set of potential antecedents for a pronoun. Under the *Simultaneous Constraints* hypothesis, the initial retrieval process uses both agreement and structural constraints simultaneously. Here we adopt the operational definition that, when the retrieval process "uses" or "implements" a constraint, the initial set of candidate antecedents does not contain any elements that would be ruled out by that constraint. The retrieval either returns all elements allowed by the constraint, or returns any one of those elements with equal probability. Let's return to our example in (1). Under the Agreement First hypothesis, the initial set of candidate antecedents would contain both 'Bill' and 'Peter' (or either one with some probability), since both match 'him' in agreement features. The structurally unacceptable 'Peter' would have to be ruled out later. Under the Simultaneous Constraints hypothesis, the initial candidate set would contain only 'Bill.'

This particular comparison reflects a slight change of perspective since the earlier studies on pronouns and the hypotheses considered here do not map directly onto any existing account. In particular, while earlier proposals asked whether structural constraints could preempt agreement features as an initial filter on the set of candidate antecedents (e.g., Nicol and Swinney, 1989; Clifton et al., 1997), we now take for granted that agreement features are used in the earliest stages of pronoun processing. The early sensitivity to agreement features fits naturally with models of sentence processing that incorporate cue-based retrieval in a content-addressable memory system (e.g., Lewis and Vasishth, 2005; Martin and McElree, 2008). What remains to be determined is when and how structural constraints play their role.

The Simultaneous Constraints hypothesis should be distinguished from previous accounts involving multiple weighted constraints, in which different constraints can be weighted and applied probabilistically (e.g., Badecker and Straub, 2002; Runner et al., 2006). For example, under Badecker and Straub's (2002) "interactive parallel constraints" hypothesis, a structural constraint can be outweighed by a discourse prominence constraint. As such, even evidence for the retrieval of a structurally unacceptable potential antecedent can be fully compatible with the immediate application of a structural constraint. In contrast, in the current formulation the simultaneously applied constraints are deterministic, so that the retrieval process cannot return an element that is ruled out by any one of them. Thus, we can falsify this hypothesis when we obtain evidence that an element that is ruled out by one (or more) of the constraints is retrieved initially.

To examine the effects of structural constraints on real-time pronoun interpretation, reading studies have generally used feature-mismatch paradigms (e.g., Clifton et al., 1999; Badecker and Straub, 2002; Lee and Williams, 2008). For example, the paradigm illustrated in (2) orthogonally manipulates the gender match between the pronoun and two potential antecedents: the structurally acceptable main clause subject ('John'/'Jane'), and the structurally unacceptable embedded clause subject ('Bill'/'Mary'). Reading times at and following the pronoun are considered indicative of the relative difficulty of resolving the reference of the pronoun (e.g., Carreiras et al., 1996; Clifton et al., 1997).

(2) a. John thought that Bill liked him a lot.
b. John thought that Mary liked him a lot.
c.Jane thought that Bill liked him a lot.
d. Jane thought that Mary liked him a lot.

In (2), the main clause subject is the only structurally acceptable potential antecedent for the pronoun 'him.' When this subject mismatches the pronoun in gender, as in (c) and (d), the sentences are considered ungrammatical. (Although an antecedent outside the sentence is technically possible, such sentences are initially perceived as ungrammatical when no context is provided. Presumably it takes some time to accommodate the lack of antecedent by inventing a potential context with an appropriate antecedent). Thus, sensitivity to the features of the structurally acceptable candidate in reading times at or following the pronoun is termed a *grammaticality effect*. On the other hand, since Principle B rules out the embedded clause subject as a potential antecedent for the pronoun, its features are irrelevant to the acceptability of the sentence. Thus (a) and (b) are equally acceptable, and (c) and (d) are equally unacceptable. Any sensitivity to the features of structurally unacceptable potential antecedents is broadly termed an *interference effect*.

Our predictions for each hypothesis depend on the assumption that some cost will be incurred if the context does not contain a potential antecedent that satisfies the constraints on the initial antecedent retrieval process. This assumption is compatible with most popular conceptions of memory retrieval mechanisms. On

the one hand, if retrieval is deterministic and exhaustive, returning all and only the candidates that satisfy all the constraints, then a lack of satisfactory potential antecedents in the context will result in retrieval failure. Since a pronoun cannot be interpreted without identifying its antecedent, retrieval failure would trigger an error signal or a repair or reanalysis process, either of which would be observable as increased reading times. On the other hand, if retrieval is probabilistic, returning a single candidate with greater or lesser likelihood depending on how fully it satisfies the constraints, then a lack of satisfactory potential antecedents in the context will result in the retrieval of partial matches. In this case, the additional cost arises from the need to rule out partial matches after they have been retrieved. Note that this "filter" would be separate from and prior to a filter based on constraints that were not active in the initial retrieval process.

Let us now consider the predictions of each of our alternative hypotheses. Under the Simultaneous Constraints hypothesis, agreement and structural constraints are both applied during the initial retrieval process, resulting in a set of feature-matching and structurally acceptable candidate antecedents. In (c) and (d), where the structurally acceptable noun phrase mismatches the pronoun in gender, the retrieval process will either—depending on one's preferred model of memory retrieval—fail to return any candidates, or return partial matches that must be ruled out. Either option would be costly. The Simultaneous Constraints hypothesis would therefore predict a grammaticality effect: longer reading times after the pronoun in (c) and (d) compared to (a) and (b). In fact, previous studies have consistently reported grammaticality effects (e.g., Clifton et al., 1999; Badecker and Straub, 2002; Lee and Williams, 2008).

Under the Agreement First hypothesis, structural constraints are not applied in initial antecedent retrieval, resulting in a set of feature-matching candidates that may or may not be structurally acceptable. In (2), (a) to (c) all contain at least one male name, matching the pronoun 'him' in features. The retrieval should only encounter difficulties in (d), which contains no feature-matching names. Thus, reading times after the pronoun in (a) to (c) should pattern together, contrasting with longer reading times in (d). Since reading times would differ between (c) and (d) based solely on the features of a structurally unacceptable potential antecedent, this is a type of interference effect. We will refer to this pattern as a *facilitative interference effect*, since the presence of a feature-matching but structurally unacceptable potential antecedent reduces reading times relative to (d).

Only a facilitative interference effect constitutes sufficient evidence to support the Agreement First hypothesis and rule out the Simultaneous Constraints hypothesis. It is therefore essential to emphasize that a facilitative interference effect has never been reported for pronouns. No existing evidence rules out the possibility that structural constraints restrict the initial antecedent retrieval process.

However, previous studies have occasionally reported *inhibitory interference effects*. These are of two types, distinguished by whether they arise in grammatical or ungrammatical sentences. Badecker and Straub (2002) observed a *multiple match effect*: when the structurally acceptable candidate matched the pronoun in gender (i.e., in grammatical sentences), reading times were longer when the structurally unacceptable candidate also matched, as in (2a), compared to when it did not (2b). Badecker and Straub interpreted this result as an effect of competition between the two feature-matching candidates. Crucially, however, since they also observed a grammaticality effect, not a facilitative interference effect, the results cannot be taken as evidence for the Agreement First hypothesis. At most, this pattern might support weakening the Simultaneous Constraints hypothesis, so that structural constraints interact with agreement constraints probabilistically, rather than deterministically restricting the initial set of candidate antecedents. To our knowledge, Badecker and Straub (2002) are the only authors to report a multiple match effect in a reading study. It is also worth noting that Clackson et al. (2011) found a similar effect in a visual world eye-tracking study.

A second type of inhibitory interference effect has also been reported. Kennison (2003) examined reading times in sentences like (3) where no structurally acceptable antecedent was available within the sentence. Reading times were longer when a structurally unacceptable potential antecedent matched the pronoun in features ('Carl'), compared to when it did not ('Susan'). We will call this type of pattern an *ungrammatical match effect*. Based on this finding, Kennison argued that structurally unacceptable potential antecedents are included in the initial set of candidate antecedents. Their presence delays the point when the comprehender can terminate the search for an antecedent and assume that the intended antecedent is an unmentioned discourse entity.

(3) {Carl/Susan} watched him yesterday during the open rehearsals of the school play.

Sturt (2003) observed a similar ungrammatical match effect in late processing measures in an eye-tracking study on reflexives. However, this study, unlike Kennison's, included a manipulation of the gender match of a structurally acceptable potential antecedent as well as the unacceptable one. There were early grammaticality effects: first-fixation and first-pass reading times on the reflexive were faster in sentences like (4), where the structurally acceptable potential antecedent ('the surgeon') matched the reflexive in stereotypical gender, compared to sentences like (5) where it mismatched. The ungrammatical match effect emerged later: in sentences like (5), second pass reading times on the pronoun were longer when the structurally unacceptable potential antecedent matched the reflexive in features ('She') than when it did not ('He'). Based on the combination of early grammaticality effects and late inhibitory interference, Sturt proposed that the initial set of candidate antecedents is structurally constrained (by Principle A, in the case of reflexives), but structurally unacceptable potential antecedents may be considered at a later stage if no acceptable candidates are retrieved initially.

(4) . . . {She/He} remembered that the surgeon had pricked himself with a used syringe needle. . . .
(5) . . . {She/He} remembered that the surgeon had pricked herself with a used syringe needle. . . .

In summary, different forms of sensitivity to structurally acceptable potential antecedents warrant distinct interpretations (cf. Sturt, 2013). Facilitative interference provides the only clear evidence for the Agreement First hypothesis. Other forms of interference are consistent with the Simultaneous Constraints hypothesis: they may reflect other properties of the processing system or later stages of processing. Thus, given our stricter interpretation of interference effects, the previous literature provides no positive evidence for the Agreement First hypothesis and is consistent with the Simultaneous Constraints hypothesis.

We had two goals for our experiments: to probe further for facilitative interference, and to investigate the causes of the other attested forms of interference. In Experiment 1 we show that comprehenders are immediately sensitive to the structural constraints on pronoun interpretation, regardless of the similarity between the candidate antecedents and their linear distance from the pronoun. We found robust effects of grammaticality, but no interference effects of any kind. In Experiment 2 we attempted to reconcile the discrepancy between our results and Badecker and Straub's (2002) findings in Experiment 2 by directly reproducing their experiment. We replicated our findings from Experiment 1, observing a clear effect of grammaticality but no interference effects. In three additional experiments, we never observed a multiple match effect. Thus, our results support the stronger version of the Simultaneous Constraints hypothesis: structural constraints immediately restrict the initial antecedent retrieval process.

EXPERIMENT 1

Experiment 1 had two goals. First, we wanted to investigate whether structural constraints immediately restrict the set of candidate antecedents for a pronoun. Second, we wanted to explore the possibility that superficial differences in experimental materials may have caused the discrepancies among previous findings.

The first goal of Experiment 1 was to examine whether structural constraints immediately restrict the set of candidate antecedents. We used the feature mismatch paradigm, manipulating the gender match between the object pronoun 'him' and two candidate antecedents: the structurally acceptable main clause subject and the structurally unacceptable embedded clause subject. According to the Simultaneous Constraints hypothesis, the initial retrieval process only returns candidate antecedents that satisfy both feature-match and structural constraints. We should observe only a main effect of grammaticality—gender match of the structurally acceptable candidate—in reading times at the pronoun and subsequent words. According to the Agreement First hypothesis, the initial retrieval process relies on feature matching alone and structural constraints only impact later stages of processing. Under this hypothesis, we should observe an interaction between the two factors in a pattern of facilitative interference. Specifically, reading times should be longer when neither potential antecedent matches the pronoun in gender, compared to the other three conditions where at least one of the potential antecedents matches the pronoun in features. The Agreement First hypothesis would also be consistent with a concurrent or subsequent multiple match effect, if retrieving multiple feature-matching candidates leads to competition-related costs.

The second goal of Experiment 1 was to explore the possibility that superficial differences in experimental materials may have caused the discrepancies among previous findings. We focused on two properties of the materials: (1) similarity between the structurally acceptable and unacceptable candidate antecedents; and (2) linear distance between the pronoun and the structurally acceptable candidate. Even if structural constraints can immediately restrict the set of candidate antecedents during the initial retrieval process (*Simultaneous Constraints* hypothesis), similarity-based interference (e.g., Gordon et al., 2001) and memory decay (Keppel and Underwood, 1962) may make it more difficult for comprehenders to distinguish between structurally acceptable and unacceptable potential antecedents during retrieval. As a result, when the potential antecedents are more similar or when the distance between the pronoun and the structurally acceptable potential antecedent is greater, comprehenders may be more likely to retrieve a feature-matching but structurally unacceptable candidate from a noisy memory representation and show facilitative interference. To explore the effects of these factors, we manipulated the properties of the embedded subject (the structurally unacceptable potential antecedent), as illustrated in **Table 1**.

Previous studies vary in the similarity between the potential antecedents in the sentence. For example, Badecker and Straub (2002) used sentences where both potential antecedents were proper names, and observed a multiple match effect. By contrast, Lee and Williams (2008) used a common noun as the structurally acceptable candidate and a proper name as the unacceptable candidate, and did not observe any interference effects. In our experiment, the main clause subject was always an unambiguously gendered proper name (e.g., 'Ethan' or 'Paige'). We manipulated the similarity between the structurally acceptable and unacceptable candidates by using either another unambiguously gendered proper name (e.g., 'Ronald,' 'Marissa') or a gender-biased common noun (e.g., 'the analyst,' 'the receptionist') as the embedded subject, as shown in **Table 1**.

Previous studies also vary in the distance between the pronoun and potential antecedents. A previous eye-tracking study by Ehrlich and Rayner (1983) found that reading times following a pronoun are longer when its antecedent is further away (cf. Walker et al., 1983). In our experiment, we increased the linear distance between the pronoun and the structurally acceptable candidate by modifying the common noun embedded subject with a subject relative clause or a prepositional phrase (e.g., 'the analyst who attended the office party'), as shown in **Table 1**.

METHODS

Participants

Thirty-six students (26 female, mean age = 20 years, range between 18 and 28) from the University of Maryland, College Park participated in this experiment. All participants were native speakers of English and had normal or corrected-to-normal vision. All participants gave informed consent and received course credit for their participation. Procedures for this experiment as well as Experiments 2–5 were approved by the Internal Review Board of the University of Maryland, College Park.

Table 1 | Experimental conditions and sample materials in Experiment 1.

	Embedded match	Embedded mismatch
COMMON NOUN		
Main clause match (grammatical)	Ethan discovered that the analyst had mocked him mercilessly for singing karaoke after drinking...	Ethan discovered that the receptionist had mocked him mercilessly for singing karaoke after drinking...
Main clause mismatch (ungrammatical)	Paige discovered that the analyst had mocked him mercilessly for singing karaoke after drinking...	Paige discovered that the receptionist had mocked him mercilessly for singing karaoke after drinking...
MODIFIED COMMON NOUN		
Main clause match (grammatical)	Ethan discovered that the analyst who attended the office party had mocked him mercilessly for singing karaoke after drinking...	Ethan discovered that the receptionist who attended the office party had mocked him mercilessly for singing karaoke after drinking...
Main clause mismatch (ungrammatical)	Paige discovered that the analyst who attended the office party had mocked him mercilessly for singing karaoke after drinking...	Paige discovered that the receptionist who attended the office party had mocked him mercilessly for singing karaoke after drinking...
PROPER NAME		
Main clause match (grammatical)	Ethan discovered that Ronald had mocked him mercilessly for singing karaoke after drinking...	Ethan discovered that Marissa had mocked him mercilessly for singing karaoke after drinking...
Main clause mismatch (ungrammatical)	Paige discovered that Ronald had mocked him mercilessly for singing karaoke after drinking...	Paige discovered that Marissa had mocked him mercilessly for singing karaoke after drinking...

Design and Materials

We crossed two levels of *main clause subject gender* (match/mismatch) and *embedded subject gender* (match/mismatch) with three levels of *embedded subject type* (proper name/common noun/modified common noun) to result in a $2 \times 2 \times 3$ within-participant design. The pronoun was always 'him,' since the feminine pronoun 'her' is ambiguous between an object pronoun and a possessive pronoun. We created 60 sets of experimental sentences. Each set included twelve variants, one in each condition. A sample set is shown in **Table 1**. A complete set of experimental stimuli are available in the Supplementary Materials.

The main clause subject was always an unambiguously gendered proper name. The embedded subject was either an unambiguously gendered proper name (e.g., 'Ronald,' 'Marissa'), a gender-biased common noun (e.g., 'the analyst,' 'the receptionist'), or a gender-biased common noun modified with a subject relative clause or a prepositional phrase (e.g., 'the analyst who attended the office party').

The gender-biased nouns were selected based on norming data from Kennison and Trofe (2003) and the intuitions of a native speaker. We collected gender bias ratings for all gender-biased nouns used in this experiment and Experiments 3–5 using Amazon Mechanical Turk. Twenty participants (9 female, mean age = 25 years, range between 21 and 28) rated each noun on a scale from 1 (most likely female) to 7 (most likely male). Overall the results support our choice of nouns. In Experiment 1, the female-biased nouns had an average rating of 2.5 (all of which had an average rating below 4) and the male-biased nouns had an average rating of 5.3 (57 out of 60 had an average rating above 4). The median rating difference between the female-biased and male-biased nouns within the same item was 2.6; 58 of the 60 pairs had mean differences of at least 1.

The 60 item sets were divided into 12 lists, such that each list contained exactly one version of each item and 5 items in each condition. Each list also contained 60 filler sentences, which varied in length and syntactic complexity and contained other referential expressions (e.g., proper names and gender-neutral nouns) and anaphors (e.g., feminine pronouns and reflexives). A third of the experimental and filler sentences were followed by a yes/no comprehension question to ensure that participants were attending to the stimuli. The comprehension questions never referred to the referential dependency between the pronoun and its antecedent. The order of experimental and filler sentences was randomized across participants.

Procedure

The experiment was conducted in a quiet room on a desktop PC. Participants read the sentences in a word-by-word, self-paced moving window task (Just et al., 1982) implemented with the Linger software package (Rohde, 2003). Each trial began with the sentence masked by underscores (___), with the words separated by spaces. Participants began a trial by pressing the spacebar, upon which the first word of the sentence appeared. They continued to press the spacebar to read each successive word. As each word appeared, the previous word was remasked. Participants were instructed to read at a natural pace and to make sure they understood what they were reading so that they could respond to comprehension questions accurately. Reaction times (RTs) were measured for each word from the time it appeared on the screen until the spacebar was pressed for the next word. In a third of the items, a comprehension question appeared after the last word in the sentence was read. Participants responded by pressing the "F" key for "Yes" and the "J" key for "No," and could then proceed to the next trial by pressing the spacebar. The experimental session was preceded by 6 practice trials to familiarize the participant

with the procedure. Testing sessions lasted approximately 35 minutes.

Analysis

Details of data analysis were consistent across all self-paced reading experiments (Experiments 1–4) and are presented for Experiment 1 only. In each experiment, only data from participants with at least 75% accuracy on the comprehension questions (and on the probe identification task in Experiment 2) were used in the analyses. No participants were excluded due to poor accuracy in Experiment 1. Trials containing RTs greater than 2000 ms were excluded from the analysis. This affected 3.4% of the data for Experiment 1.

Average reading times were compared across conditions in the following regions of interest: the pronoun itself (*pronoun*) and the two words immediately following the pronoun (*pronoun+1* and *pronoun+2*). Data for each of the regions of interest were entered into a $2 \times 2 \times 3$ repeated measures ANOVA with *main clause match*, *embedded match*, and *embedded subject type* as within-participant and within-item factors. ANOVAs were computed on the participant means collapsing over items (F1), and on the item means collapsing over participants (F2). Below we report comparisons that revealed a statistically significant difference in at least one of the by-participant and by-item analyses. Since the manipulation of embedded subject type resulted in superficial differences in the materials (e.g., sentence length), effects of embedded subject type are not interpretable unless they interact with the effects of main clause match and/or embedded match. Therefore, only effects involving main clause match or embedded match are discussed. Further, a 2×2 repeated measures ANOVA with *main clause match* and *embedded match* were conducted on each level of *embedded subject type* when it interacted with one or both of the other factors.

RESULTS

Participants answered the comprehension questions with an average of 87.8% accuracy.

Table 2 shows average reading times and standard errors in each region of interest (ROI) across all conditions. **Figure 1** shows average reading times starting from the word preceding the pronoun (the embedded verb) to one word following *pronoun+2* across conditions in each level of embedded subject type. The three-way repeated measures ANOVA in the *pronoun* region revealed a significant main effect of *embedded match* in the by-participant analysis [$F_{1(1,\ 35)} = 5.84$, $p < 0.05$; $F_{2(1,\ 59)} = 3.33$, $p = 0.07$]: reading times were longer when the embedded subject mismatched the pronoun in gender. A significant main effect of *main clause match* was observed in both the *pronoun+1* [$F_{1(1,\ 35)} = 16.12$, $p < 0.001$; $F_{2(1,\ 59)} = 30.57$, $p < 0.001$] and *pronoun+2* [$F_{1(1,\ 35)} = 12.24$, $p < 0.01$; $F_{2(1,\ 59)} = 21.35$, $p < 0.001$] regions: reading times were significantly longer when the main clause subject mismatched the pronoun in gender (i.e., a grammaticality effect). This main effect was accompanied by a significant interaction between main clause match and embedded subject type in the *pronoun+2* region [$F_{1(1,\ 35)} = 16.49$, $p < 0.001$; $F_{2(2,\ 118)} = 3.57$, $p < 0.05$].

Table 2 | Grand average reading times (with standard deviations) in each ROI across all conditions in Experiment 1.

	Pronoun	Pronoun+1	Pronoun+2
COMMON NOUN			
Main clause match, embedded match	372 (15)	406 (23)	374 (14)
Main clause match, embedded mismatch	391 (17)	401 (18)	399 (18)
Main clause mismatch, embedded match	377 (18)	465 (28)	412 (18)
Main clause mismatch, embedded mismatch	418 (20)	479 (26)	421 (21)
MODIFIED COMMON NOUN			
Main clause match, embedded match	348 (11)	354 (13)	376 (15)
Main clause match, embedded mismatch	359 (17)	380 (19)	376 (16)
Main clause mismatch, embedded match	341 (15)	405 (32)	414 (19)
Main clause mismatch, embedded mismatch	364 (20)	438 (26)	374 (16)
PROPER NAME			
Main clause match, embedded match	387 (16)	420 (23)	382 (13)
Main clause match, embedded mismatch	386 (15)	410 (22)	388 (16)
Main clause mismatch, embedded match	393 (23)	467 (33)	438 (25)
Main clause mismatch, embedded mismatch	407 (20)	497 (34)	471 (26)

To better understand the interaction involving embedded subject type, a two-way repeated measures ANOVA with *main clause match* and *embedded match* was conducted on each level of *embedded subject type* in the *pronoun+2* region. When the embedded subject was a common noun (e.g., 'the analyst'; **Figure 1A**), there was a significant main effect of *main clause match* in the by-item analysis [$F_{1(1,\ 35)} = 2.60$, $p > 0.1$; $F_{2(1,\ 59)} = 4.08$, $p < 0.05$]. When the embedded subject was a modified common noun (e.g., 'the analyst who attended the office party'; **Figure 1B**), the main effect of *main clause match* was not significant ($p > 0.1$), but there was an interaction between *main clause match* and *embedded match* that is significant in the by-participant analysis [$F_{1(1,\ 35)} = 5.59$, $p < 0.05$; $F_{2(1,\ 59)} = 1.91$, $p > 0.1$]: gender mismatch between the pronoun and the main clause subject led to longer reading times in the *pronoun+2* region only when the embedded subject matched the pronoun (i.e., an ungrammatical match effect). When the embedded subject was a proper name (e.g., 'Ronald'; **Figure 1C**), there was a significant main effect of *main clause match* [$F_{1(1,\ 35)} = 14.83$, $p < 0.001$; $F_{2(1,\ 59)} = 13.59$, $p < 0.001$]: reading times were significantly longer when there was a gender mismatch between the pronoun and the main clause subject (i.e., a grammaticality effect).

DISCUSSION

Experiment 1 had two main findings. First, we observed a robust grammaticality effect: reading times after the pronoun were

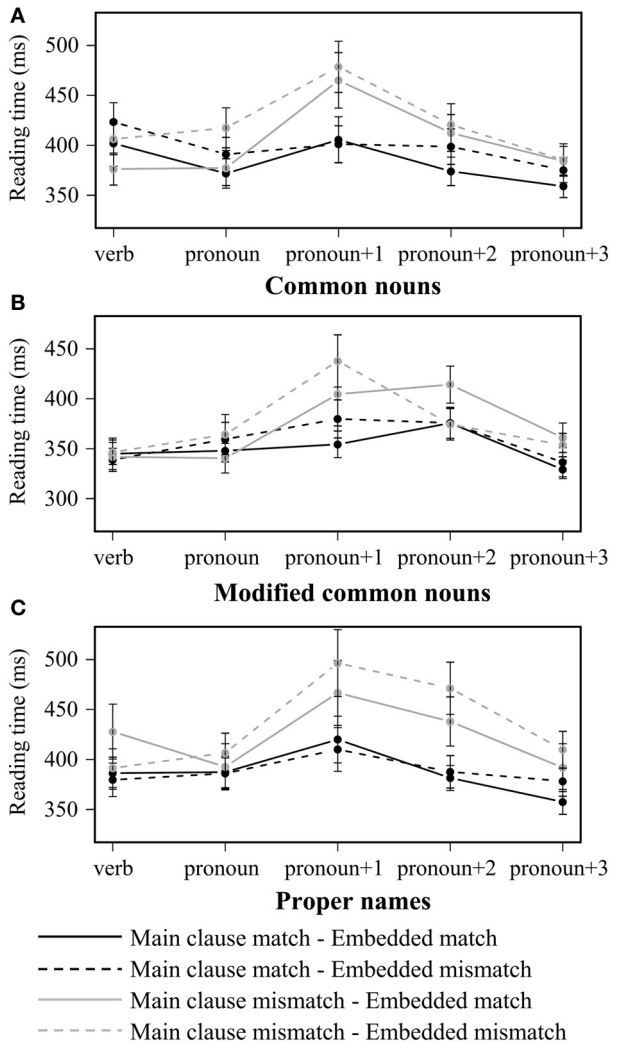


FIGURE 1 | Word-by-word reading times for (A) the common noun, (B) modified common noun and (C) proper name conditions in Experiment 1. Error bars indicate standard error of the participant mean.

significantly longer when the only structurally acceptable potential antecedent mismatched the pronoun in gender (a main effect of *main clause match* in both post-pronoun regions). This grammaticality effect was modulated by embedded subject type in the *pronoun+2* region: the effect of grammaticality was largest when both the main clause and embedded subjects were proper names.

Second, we never observed a facilitative interference effect. Overall three-way ANOVAs did not reveal a significant *main clause match* × *embedded match* interaction in any of the ROIs. Although a significant interaction was observed in the two-way ANOVA in the *pronoun+2* region in the modified common noun condition, it showed the opposite pattern: the presence of a feature-matching structurally unacceptable potential antecedent led to longer, rather than shorter, reading times in the absence of an acceptable antecedent. Therefore, neither similarity between the structurally acceptable and unacceptable candidates nor increased distance between the pronoun and the acceptable antecedent resulted in more retrievals of structurally unacceptable potential antecedents. Taken together, the robust sensitivity to the gender of a structurally acceptable potential antecedent and the absence of facilitative inference effects support the Simultaneous Constraints hypothesis. Structural criteria can immediately restrict the set of candidate antecedents during the initial memory retrieval processes.

An unexpected finding of this experiment is the observation of a significant main effect of *embedded match* in the *pronoun* region in the by-participant analysis. This effect did not interact with *embedded subject type* and was not predicted by either hypothesis. A closer inspection of the data suggests that it was mainly carried by the difference in the common noun condition (embedded mismatch: 404 ms vs. embedded match: 374 ms; **Figure 1A**), which showed a similar difference in reading times in the preceding region (415 ms vs. 389 ms). Therefore, this effect may be spurious and unrelated to pronoun processing.

Although the manipulation of *embedded subject type* never resulted in any facilitative interference effects, it did lead to an interesting interaction between *main clause match* and *embedded subject type* in the *pronoun+2* region. In particular, while main clause mismatch led to significantly longer reading times across all embedded subject types in the *pronoun+1* region, this effect continued to be observed in the *pronoun+2* region only in the proper name condition. That is, when the structurally acceptable and unacceptable potential antecedents were more similar to each other (both were proper names), the grammaticality effect lasted longer. This unexpected pattern could reflect either a more sustained processing disruption or greater variability in the onset time of the disruption.

When the linear distance between the pronoun and the structurally acceptable potential antecedent was lengthened (in the modified common noun condition), we observed a late-emerging ungrammatical match effect. Following the main effect of *main clause match* in the *pronoun+1* region, main clause mismatch led to longer reading times in the *pronoun+2* region only when the embedded subject matched the pronoun. Following Sturt's (2003) proposal for the processing of reflexives, we propose that this inhibitory interference reflects a repair process triggered by an initial failure to retrieve a feature-matching and structurally acceptable antecedent for the pronoun. We take the increased reading times in the *main clause mismatch, embedded match* condition to suggest that a feature-matching antecedent from a structurally unacceptable position may be retrieved when the initial retrieval fails. The observation of an ungrammatical match effect in the modified common noun conditions, but not in the common noun and proper name conditions, suggests that an initial retrieval failure may be more likely to trigger a repair process when the memory representation of the structurally acceptable potential antecedent is less activated due to decay over time. Note, however, that the embedded subject NP was heavier (and more complex) in the modified common noun condition than in the other two conditions. Since a heavier NP may require a more detailed memory representation, the heaviness (or complexity) of the structurally unacceptable potential antecedent may also impact the likelihood of triggering a repair process. More research will be needed to explore the effects of the heaviness of an NP

on memory representation and how it might impact memory encoding and retrieval more generally.

Finally, we never observed a multiple match effect: reading times were never longer when both subjects matched the pronoun compared to when only the main clause subject matched the pronoun. As shown in **Figure 1**, reading times in the multiple match conditions were short in every region, across all three levels of embedded subject type. We aimed to resolve this discrepancy between the current results and Badecker and Straub's (2002) findings in Experiment 2.

EXPERIMENT 2

In Experiment 1 we never observed any sensitivity to the presence of multiple feature-matching candidate antecedents. This contrasts with Badecker and Straub's (2002) repeated observations of a multiple match effect—longer reading times when both candidate antecedents matched the pronoun. We reasoned that, even though the proper name condition in Experiment 1 mirrored Badecker and Straub's (2002; hereafter B&S) Experiment 1, other differences between the experimental materials and procedures might have given rise to the discrepancy in the results. Thus, in Experiment 2, we attempted to directly replicate B&S's Experiment 1, using identical experimental materials and procedures.

We identified three main differences between the materials and procedures used in our Experiment 1 and B&S's Experiment 1. First, while we only used the masculine object pronoun 'him,' in order to avoid the ambiguity of the pronoun, B&S used the ambiguous feminine object pronoun 'her' in half of the sentences, and analyzed the results from sentences with feminine and masculine object pronoun together. Second, while our participants answered yes/no comprehension questions after a third of the items, B&S's participants performed a probe recognition task after each sentence and answered a yes/no comprehension only after a quarter of the sentences. Furthermore, B&S's participants received auditory feedback on their accuracy for both secondary tasks. Finally, while we presented sentences in a moving-window paradigm, B&S presented each word serially in the center of the screen. Since any of these differences may have contributed to differences in the results, we decided to begin our investigation by adopting all of the methods from B&S, in an attempt to replicate their original findings.

METHODS

Participants

Twenty-six students (25 female, mean age = 20 years, range between 18 and 22) from the University of Maryland, College Park participated in this experiment. All participants were native speakers of English and had normal or corrected-to-normal vision. All participants gave informed consent and received course credit for their participation. Data from two additional participants were excluded: one because accuracy on comprehension questions was too low (71%); the other because too many experimental items (25%) contained RTs greater than 2000 ms.

Design and Materials

This experiment had a 2 × 2 within-participant design in which *main clause match* and *embedded match* were fully crossed. We used the original 24 sets of sentences in B&S's Experiment 1. These materials contained an unambiguously gendered proper name in both the main clause and embedded subject positions and therefore resembled the materials used in the proper name condition in Experiment 1 (see **Table 3** for an example). In the original study half of the sentences used the feminine object pronoun 'her,' but sentences with feminine and masculine object pronoun were analyzed together. In order to increase the statistical power for examining the effects of the gender of the pronoun, we created 24 additional sets of sentences modeled after B&S's items, half of which used the feminine object pronoun. A complete set of experimental stimuli are available in the Supplementary Materials.

The 48 item sets were divided into four presentation lists, such that each list contained exactly one version of each item and 6 items in each condition. Each list also contained 100 filler sentences, which varied in length and syntactic complexity and contained other referential expressions (e.g., proper names and gender-neutral nouns) and anaphors (e.g., feminine pronouns and reflexives). Following B&S, a single word probe was selected for each experimental and filler item set. For half of the items, the probe word was selected from among the content words of the sentence—never the pronoun or either of the proper names.

Table 3 | Experimental conditions and sample materials in Experiment 2.

	Embedded match	Embedded mismatch
MASCULINE OBJECT PRONOUN 'HIM'		
Main clause match (grammatical)	Arthur believed that Ben owed him a second chance to solve the problem.	Arthur believed that Meg owed him a second chance to solve the problem.
Main clause mismatch (ungrammatical)	Sheila believed that Ben owed him a second chance to solve the problem.	Sheila believed that Meg owed him a second chance to solve the problem.
FEMININE OBJECT PRONOUN 'HER'		
Main clause match (grammatical)	Sheila believed that Meg owed her a second chance to solve the problem.	Sheila believed that Ben owed her a second chance to solve the problem.
Main clause mismatch (ungrammatical)	Arthur believed that Meg owed her a second chance to solve the problem.	Arthur believed that Ben owed her a second chance to solve the problem.

The location of the probe in the sentence (initial, medial, or final) was counterbalanced across items. For the other half of the items, words that did not occur in the sentence(s) were selected. Among these "no" probes, one third were semantic associates to a content word in the sentence (e.g., beach—ocean), one third were morphologically related to a word in the sentence (e.g., accepted—acceptance), and one third were neither semantically nor morphologically related to any content words in the sentence. Following B&S, comprehension questions were presented on one quarter of the trials. As in Experiment 1, responses to comprehension questions never required successful pronoun resolution. Finally, five additional complete trials were constructed to serve as practice trials.

Procedure

The procedure was similar to that of Experiment 1 with three critical differences. First, words were presented at the center of the screen. Second, at the end of each sentence, a probe word appeared at the center of the screen and the participant used the keyboard to indicate whether that probe word had occurred in the sentence. Finally, auditory feedback was provided to indicate accuracy on both of the secondary tasks. Testing sessions lasted approximately 30 min.

Analysis

As in Experiment 1, trials containing RTs greater than 2000 ms were excluded from the analysis. This affected 2.2% of the data. Initial statistical analyses were performed on data from all items, collapsing across pronoun gender. Data for each of the regions of interest were entered into a 2 × 2 repeated measures ANOVA with main clause match and embedded match as within-participant factors. We conducted two follow-up analyses to further examine potential differences between the present results and B&S's findings. We performed the same 2 × 2 repeated measures ANOVA on the subset of items taken from B&S to determine whether those items would show a different pattern. Finally, to examine the role of the pronoun gender, we added Gender as an additional factor to analyze all the items together. Data for each of the regions of interest were entered into a 2 × 2 × 2 repeated measures ANOVA with main clause match, embedded match and pronoun gender as within- participant factors.

RESULTS

Participants answered the comprehension questions and performed the probe recognition task with an average of 86.2 and 93.9% accuracy respectively.

All items

Main clause mismatch led to longer reading times across several regions (see **Figure 2**). The main effect of *main clause match* was significant in the by-items analysis in the *pronoun* region [422 ms vs. 400 ms; $F_{1(1,\ 25)} = 3.57$, $p = 0.07$; $F_{2(1,\ 47)} = 4.87$, $p < 0.05$], and in both analyses in the *pronoun+1* [447 ms vs. 392 ms; $F_{1(1,\ 25)} = 12.0$, $p < 0.01$; $F_{2(1,\ 47)} = 16.9$, $p < 0.001$] and *pronoun+2* [414 ms vs. 377 ms; $F_{1(1,\ 25)} = 8.82$, $p < 0.01$; $F_{2(1,\ 47)} = 16.8$, $p < 0.001$] regions. No other comparisons revealed a statistically significant difference (p's > 0.05).

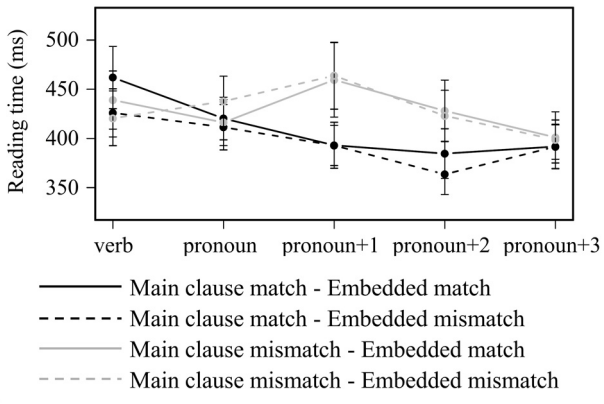


FIGURE 2 | Word-by-word reading times for all items in Experiment 2. Error bars indicate standard error of the participant mean.

Badecker and Straub's (2002) items

A similar pattern of results was observed in a separate analysis of just the subset of items from B&S. Main clause mismatch led to longer reading times at the pronoun and the three subsequent regions. this main effect of *main clause match* was statistically significant in the *pronoun* region [418 ms vs. 386 ms; $F_{1(1,\ 25)} = 5.26$, $p < 0.05$; $F_{2(1,\ 23)} = 7.25$, $p < 0.05$], and in the by-items analysis in the *pronoun+1* region [438 ms vs. 392 ms; $F_{1(1,\ 25)} = 3.89$, $p = 0.06$; $F_{2(1,\ 23)} = 4.81$, $p < 0.05$]. In the *pronoun+2* region, there was an interaction between *main clause match* and *embedded match* which was significant in the by-items analysis [$F_{1(1,\ 25)} = 3.57, p = 0.07; F_{2(1,\ 23)} = 5.72, p < 0.05$]. This interaction had the pattern of an ungrammatical match effect: when the main clause subject mismatched the pronoun, RTs were longer when the embedded subject matched (431 ms vs. 383 ms). No other comparisons revealed a statistically significant difference (p's > 0.05).

Pronoun gender

In a follow-up analysis which included pronoun gender as an additional factor, we continued to observe a main effect of *main clause match* across the *pronoun* [$F_{1(1,\ 25)} = 3.63$, $p = 0.07$; $F_{2(1,\ 47)} = 6.18$, $p < 0.05$], *pronoun+1* [$F_{1(1,\ 25)} = 12.1$, $p < 0.01$; $F_{2(1,\ 47)} = 15.8$, $p < 0.001$], and *pronoun+2* [$F_{1(1,\ 25)} = 8.89$, $p < 0.01$; $F_{2(1,\ 47)} = 14.3$, $p < 0.001$] regions. In addition, reading times were significantly longer for 'him' than for 'her' in the *pronoun* [426 ms vs. 396 ms, $F_{1(1,\ 25)} = 12.0$, $p < 0.01$; $F_{2(1,\ 47)} = 9.46$, $p < 0.01$] and *pronoun+1* [430 ms vs. 408 ms, $F_{1(1,\ 25)} = 7.70, p < 0.05$; $F_{2(1,\ 47)} = 3.27, p = 0.08$] regions. No other comparisons revealed a statistically significant difference (p's > 0.05).

DISCUSSION

In this experiment we adopted the experimental materials and procedures used in B&S's Experiment 1 in an attempt to replicate their observation of a multiple match effect. This attempt was unsuccessful, as we once again failed to observe any sensitivity to the presence of multiple feature-matching candidate antecedents. Instead we replicated the key findings from our

Experiment 1: a robust effect of grammaticality—longer reading times when the main clause subject mismatched the pronoun—and no facilitative interference effect or multiple match effect. When we looked at the subset of items taken from B&S's original study, we observed only a late ungrammatical match effect, similar to that observed in the modified common noun condition in Experiment 1. Across all cases, reading times were never modulated by the presence/absence of a feature-matching embedded subject when the main clause subject matched the pronoun in features.

EXPERIMENTS 3–5

Here we present three further attempts to explore the potential cause of comprehenders' sensitivity (or lack thereof) to the presence of multiple candidate antecedents when processing a pronoun. The design of these experiments was different from that of Experiments 1 and 2. To focus on potential multiple match effects, we removed the manipulation of the main clause subject—it always matched the pronoun in gender. We added a new manipulation of the *pronoun type*: object ('him') vs. possessive ('his'). In the possessive condition, both the main clause and embedded subjects are structurally acceptable, so the pronoun is referentially ambiguous. Thus, if the multiple match effect is possible, we should certainly expect to see it in the possessive condition.

We also added a manipulation of the referential status of the embedded subject: referential (e.g., 'the consultant') vs. quantified (e.g., 'every consultant'). This manipulation was originally motivated by the hypothesis that the multiple match effects observed by B&S could be related to the fact that local antecedents are acceptable in certain pragmatic contexts (Evans, 1980). Such effects might not be expected for quantified NPs (Reinhart, 1983). However, since we never observed a multiple match effect in any of our experiments, a full explanation of the theoretical motivation for this manipulation is beyond the scope of this paper. Here it serves only as a further test of the more basic questions about the structure sensitivity of the initial antecedent retrieval process across a wider range of sentences.

In Experiments 3 and 4 we used a moving-window self-paced reading paradigm. In Experiment 5 we used eye-tracking to examine comprehenders' eye movements while reading. In self-paced reading paradigms, reading must proceed in one direction, while in eye-tracking paradigms participants are free to skip or re-read parts of the sentence that they have previously read (or skipped). Thus, eye-tracking may be able to detect differences that only emerge in more naturalistic reading.

To preview, we never observed a multiple match effect in any of these experiments. In fact, comprehenders did not show increased reading times to a pronoun and its subsequent words, even in cases of genuine referential ambiguity, where multiple structurally acceptable and feature-matching candidate antecedents were available (in the possessive condition). Since the same design was used in all three experiments and they yielded minimally different results, below we report the methods and results of the three experiments together.

DESIGN AND MATERIALS

The same experimental design was used across Experiments 3–5. We manipulated *pronoun type* (object vs. possessive pronoun), the embedded subject's *referential status* (referential vs. quantified) and the *embedded subject gender match* (match vs. mismatch) in a $2 \times 2 \times 2$ within-participant design. A sample item set from Experiment 3 is shown in **Table 4**. The *pronoun type* determined the structural acceptability of the embedded subject as an antecedent for the pronoun: in the object pronoun condition ('him'), only the main clause subject is structurally acceptable as an antecedent, while in the possessive pronoun condition ('his'), both subjects are structurally acceptable. We only used singular masculine pronouns, as in Experiment 1, to avoid the lexical ambiguity of 'her.' The embedded subject, a stereotypically gender-biased common noun, was either quantified (e.g., 'every consultant') or referential (e.g., 'the consultant'). The main clause subject always matched the pronoun in gender, but the embedded subject was manipulated to either match or mismatch the gender of the pronoun.

There were minimal differences in the experimental materials across the three experiments. First, the main clause subjects were stereotypically male common nouns in Experiment 3, and unambiguously male proper names in Experiments 4 and 5. Second, to make the sentences more felicitous in the quantified condition, the embedded subject was modified by a relative clause or prepositional phrase in Experiment 3. In Experiment 4, the embedded subject was not modified; instead the experimental sentence was preceded by a context sentence. Finally, in Experiments 4 and 5, we added longer words (e.g., adverbs) immediately after the pronoun to reduce the likelihood of floor effects on reading times in the critical regions. The differences between the materials are illustrated in (6) and (7).

(6) *A sample item from Experiment 3:*
The lawyer believed that the stock broker who reported the fraud had deceived him about the extent of the illegal activity.

Table 4 | Experimental conditions and sample materials for Experiment 3.

	Embedded match (multiple match)	Embedded mismatch (single match)
Referential	The lawyer believed that the stock broker who reported the fraud had deceived him/his *boss* about the extent of the illegal activity.	The lawyer believed that the secretary who reported the fraud had deceived him/his *boss* about the extent of the illegal activity.
Quantified	The lawyer believed that every stock broker who reported the fraud had deceived him/his *boss* about the extent of the illegal activity.	The lawyer believed that every secretary who reported the fraud had deceived him/his *boss* about the extent of the illegal activity.

(7) *A sample item from Experiments 4 and 5:*
There appeared to be widespread fraud in the management of the hedge fund. Brian believed that the stock broker had deceived him repeatedly about the extent of the illegal activity.

A total of 80 sets of experimental sentences were used in Experiment 3; 64 sets were adapted and used in Experiments 4 and 5. A complete set of experimental stimuli are available in the Supplementary Materials. Gender bias of the common nouns was determined in an offline rating study (see Experiment 1, Design and Materials). On a scale from 1 (most likely female) to 7 (most likely male), female-biased nouns had an average rating 2.5 in Experiment 3 and 2.4 in Experiments 4 and 5. All female-biased nouns had an average rating below 4 (more likely to be female). Male-biased nouns had an average rating of 5.2 in Experiment 3 and 5.3 in Experiments 4 and 5. Most of them (76 of 80 main clause subjects and 75 of 80 embedded subjects in Experiment 3; 61 of 64 in Experiments 4 and 5) had an average rating above 4 (more likely to be male). The median rating difference between the female-biased and male-biased nouns within the same item was 2.6 points in both sets of stimuli.

In each experiment, experimental sentences were divided into 8 lists, each containing exactly one version of each item and the same number of items in each condition. A total of 80, 64, and 104 filler sentences of comparable length and structural complexity were used in Experiments 3, 4, and 5 respectively. Filler sentences contained other referential expressions (e.g., proper names and gender-neutral nouns) and anaphors (e.g., feminine pronouns and reflexives). In Experiments 3 and 5 every experimental and filler sentence was followed by a yes/no comprehension question; in Experiment 4 a yes/no comprehension question appeared following approximately one third of the trials (22 of 64 experimental and filler sentences respectively). The comprehension questions never referred to the referential dependency between the pronoun and its antecedent(s). The order of experimental and filler sentences was randomized across participants.

EXPERIMENT 3

Participants

Twenty-six students (15 female, mean age = 22 years) from the University of Maryland, College Park participated in this experiment. All gave informed consent and were paid $10 per hour for their participation. Data from two additional participants were excluded: one because accuracy on comprehension questions was too low (<70%); the other because too many experimental items (>20%) contained RTs greater than 2000 ms.

Procedure

The procedure was identical to that of Experiment 1. Testing sessions lasted approximately 45 min.

Analysis

Data for different pronoun types were analyzed separately. Data from each region of interest were entered into a 2 × 2 repeated measures ANOVA with *referential status* and *embedded match* as within-participant factors. As in Experiments 1 and 2, trials containing RTs greater than 2000 ms were excluded from the analysis. This affected 1.2% of the data.

Results

Participants answered the comprehension questions with an average of 86.7% accuracy.

Grand average reading times in each ROI across all conditions are presented in **Table 5**.

No significant differences were observed in the *pronoun* and the *pronoun+1* region in either object pronoun or possessive pronoun condition. In the *pronoun+2* region, there was a significant main effect of *referential status* in the object pronoun condition [$F_{1(1,\ 25)} = 5.05$, $p < 0.05$; $F_{2(1,\ 79)} = 4.12$, $p < 0.05$]: reading times were shorter when the embedded subject was quantified (300 ms) compared to when it was referential (312 ms). A reversed pattern was observed in the possessive pronoun condition [main effect of *referential status*: $F_{1(1,\ 25)} = 4.11$, $p = 0.05$; $F_{2(1,\ 79)} = 6.16$, $p < 0.05$]: reading times were longer when the embedded subject was quantified (347 ms) compared to when it was referential (326 ms). No other comparisons revealed a statistically significant difference (p's > 0.1).

EXPERIMENT 4

Participants

Thirty-eight students (30 female, mean age = 22 years) from the University of Maryland, College Park participated in this experiment. All gave informed consent and received course credit or $10 per hour for their participation. Data from one additional participant were excluded due to low accuracy on the comprehension questions (<70%).

Procedure

The procedure was identical to that of Experiments 1 and 3.

Analysis

The analysis method was identical to that of Experiment 3. Outlier rejection (RTs > 2000 ms) affected 2.5% of the data.

Table 5 | Grand average reading times in each ROI across all conditions in Experiment 3.

	Object pronoun (him)			Possessive pronoun (his)		
	Pronoun	**Pronoun+1**	**Pronoun+2**	**Pronoun**	**Pronoun+1**	**Pronoun+2**
Quantified, multiple match	331 (13)	319 (10)	300 (9)	325 (10)	328 (14)	343 (15)
Quantified, single match	330 (12)	329 (12)	301 (11)	322 (11)	335 (12)	351 (14)
Referential, multiple match	332 (14)	320 (12)	315 (13)	330 (14)	324 (13)	327 (12)
Referential, single match	332 (12)	322 (11)	309 (12)	327 (12)	330 (13)	325 (11)

Results

Participants answered the comprehension questions with an average of 90.0% accuracy.

Grand average reading times in each ROI across all conditions are presented in **Table 6**.

No significant differences were observed in any of the regions of interest in either of the pronoun conditions (all p's > 0.1).

EXPERIMENT 5

Participants

Twenty-four students (13 female, mean age = 22 years) from the University of Maryland, College Park participated in this experiment. All gave informed consent and received course credit or $10 per hour for their participation. Data collected from five additional participants were excluded due to problems with calibration.

Procedure

Participants were tested individually in a quiet room in one session lasting 45–60 min. Eye movements were recorded using an EyeLink 1000 eye-tracker (SR Research, Toronto, Ontario, Canada) interfaced with a PC computer. Participants were seated with their chin and forehead stabilized by the eye-tracker apparatus, 32 inches from an LCD monitor which displayed the stimuli. At this distance, 4.6 characters were displayed per degree of visual arc. The eye-tracker has an angular resolution of 0.25–0.5°. Viewing was binocular, but only the right eye was recorded. The sampling rate for recordings was 1000 Hz. Stimulus presentation and interface with the eye-tracker was implemented with the EyeTrack software suite (University of Massachusetts, Amherst).

Sentences were presented in 12-point fixed-width Courier font in two lines. The line break was located after the first word occurring at least 100 characters from the beginning of the line. Depending on the length of the first sentence, the line break generally fell around the fourth or fifth word of the second sentence—for example, between 'the' and 'consultant' in the sample item above. This location for the line break ensured that the pronoun and its following word appeared near the center of the second line. A calibration procedure was performed before the experiment, and re-calibration was carried out between trials as needed. Before the experiment began, each participant was instructed to read for comprehension as naturally as possible. Each trial began with only a gray square on the left edge of the display. The participant triggered the appearance of the sentences by fixating on the square, and pressed a button when they had finished reading to end the display of the item and trigger the presentation of the comprehension question.

Analysis

The initial stage of data analysis was carried out using EyeDoctor (UMass Amherst, http://www.psych.umass.edu/eyelab/software/). Trials with major tracker losses were excluded from the analyses. This resulted in the exclusion of 2.3% of all trials. Each trial was visually inspected to correct for small vertical drifts. Fixations of less than 80 ms in duration and within one character of the previous or following fixation were incorporated into this neighboring fixation. All remaining fixations shorter than 80 ms were excluded. Following Rayner and Pollatsek (1989), we assume that readers do not extract much information during such short fixations. We also excluded fixations longer than 800 ms.

We analyzed three regions, which corresponded to (i) the *pronoun* region, which included the pronoun and its immediately preceding word (i.e., the embedded verb), (ii) the *pronoun+1* region, which included the word immediately following the pronoun, and (iii) the *pre-final* region, which consisted of all words between the *pronoun+1* region and the sentence-final word (exclusive). Spaces between regions were included in the following region. Regions are indicated by brackets in the sample in (8).

(8) The international firm was to hold a press conference in the coming week. Patrick said that the consultant had [prepared him][sufficiently][to make a statement at the] meeting.

Standard eye-tracking measures (Rayner, 1998) were calculated for each region. We report three eye-tracking measures that are representative of early and late measures. *First-pass time* is the sum of all fixation times starting with the first fixation inside a region until the first fixation outside the region (either to the left or right) provided that the reader has not fixated subsequent text. For regions consisting of a single word, first-pass time corresponds to gaze duration (Rayner and Duffy, 1986). *Regression-path time* (e.g., Brysbaert and Mitchell, 1996) is the sum of all fixation times starting with the first fixation inside the region until the first fixation to the right of the region, again provided that the reader has not fixated subsequent text. Finally, *total time* is the sum of all fixations in a region. For all reading time measures, the data for a particular region were excluded if the reading time measure for that region was zero.

As in Experiments 3 and 4, data for different pronoun types were analyzed separately. Data from each region of interest were entered into a 2 × 2 repeated measures ANOVA with *referential status* and *embedded match* as within-participant and within-item

Table 6 | Grand average reading times in each ROI across all conditions in Experiment 4.

	Object pronoun (him)			Possessive pronoun (his)		
	Pronoun	**Pronoun+1**	**Pronoun+2**	**Pronoun**	**Pronoun+1**	**Pronoun+2**
Quantified, multiple match	380 (17)	373 (13)	371 (12)	374 (15)	424 (22)	385 (15)
Quantified, single match	383 (17)	385 (15)	383 (13)	357 (15)	406 (19)	408 (16)
Referential, multiple match	386 (18)	381 (15)	360 (11)	362 (16)	400 (19)	394 (14)
Referential, single match	406 (24)	370 (11)	375 (13)	367 (18)	404 (18)	382 (15)

factors. Below we report F1 and F2 statistics for data in the object pronoun condition and only F1 statistics for data in the possessive pronoun condition due to missing data in a small set of items in one of the regions or measures.

Results

Participants answered the comprehension questions with an average of 91.0% accuracy.

Grand average first pass time, regression path time, and total reading times in each ROI across all conditions are presented in **Table 7**.

Object pronoun condition (him). Repeated measures ANOVA revealed a significant main effect of *referential status* on first-pass time in the pronoun region [$F_{1(1,\ 23)} = 4.81$, $p < 0.05$; $F_{2(1,\ 63)} = 3.41$, $p < 0.1$]: reading times were longer when the embedded subject was quantified than when it was referential. No other comparisons revealed a statistically significant difference (p's > 0.05).

Possessive pronoun condition (his). In the *pronoun+1* region, there was a significant main effect of *referential status* on regression path time [$F_{1(1,\ 23)} = 7.77$, $p < 0.05$]: reading times were longer when the embedded subject was quantified than when it was referential. This effect was reversed in the *pre-final* region, in which regression path time was significantly shorter in the quantified conditions than in the referential conditions [$F_{1(1,\ 23)} = 6.27$, $p < 0.05$]. No other comparisons revealed a statistically significant difference (p's > 0.05).

DISCUSSION

In Experiments 3–5 we examined whether comprehenders are sensitive to the presence of multiple feature-matching candidate antecedents, in cases where both candidates are structurally acceptable ('his') and in cases where only one is ('him').

The results were largely the same across all three experiments, and consistent with the findings of Experiments 1 and 2. Comprehenders were not sensitive to the gender match of the embedded subject, regardless of its referential status. Surprisingly, this also held in the 'his' condition, where we expected to observe a multiple match effect due to the referential ambiguity. This suggests that resolving this referential ambiguity did not lead to any observable processing cost, or that comprehenders did not in fact resolve it online. We will return to discuss this in more detail in the General Discussion.

GENERAL DISCUSSION

Our goal in this paper was to investigate the role of structural constraints in the early stages of pronoun resolution—specifically, the initial retrieval of potential antecedents. We considered two hypotheses. Under the Simultaneous Constraints hypothesis, the initial retrieval would return a set of candidate antecedents constrained by both structural and agreement criteria. Under the Agreement First hypothesis, the initial retrieval would be constrained only by agreement features, while structural constraints would come into play later. Across all our experiments, the results supported the Simultaneous Constraint hypothesis.

In Experiments 1 and 2, we found that comprehenders are sensitive to structural constraints on antecedents as early as agreement constraints. Across all five experiments, we never observed any facilitative interference from the structurally unacceptable potential antecedent. Evidence for inhibitory interference was sparse: there were no instances of multiple match effects, and only one instance of an ungrammatical match effect, which emerged later than the initial sensitivity to structural constraints. Thus we have strong, consistent evidence for structure sensitivity in the earliest stages of pronoun resolution.

NO FACILITATIVE INTERFERENCE

The consistent lack of facilitative interference effects speaks against the Agreement First hypothesis. If the initial stages of pronoun resolution used only agreement features to identify a set of candidate antecedents, then reading times immediately

Table 7 | Grand average first pass time, regression path time, and total reading times in each ROI across all conditions in Experiment 5.

	Object pronoun (him)			Possessive pronoun (his)		
Measure	**Pronoun**	**Pronoun+1**	**Pre-final**	**Pronoun**	**Pronoun+1**	**Pre-final**
FIRST-PASS TIME						
Quantified, multiple match	324 (23)	285 (12)	790 (50)	334 (22)	297 (18)	1034 (62)
Quantified, single match	338 (24)	290 (16)	882 (56)	302 (22)	276 (10)	1011 (72)
Referential, multiple match	309 (15)	275 (12)	827 (50)	324 (24)	290 (16)	1005 (56)
Referential, single match	289 (16)	310 (16)	841 (64)	328 (19)	270 (14)	1063 (68)
REGRESSION PATH TIME						
Quantified, multiple match	471 (36)	394 (24)	1365 (165)	473 (38)	405 (30)	1518 (103)
Quantified, single match	436 (36)	422 (39)	1486 (150)	519 (68)	443 (36)	1777 (149)
Referential, multiple match	477 (33)	406 (43)	1320 (142)	476 (33)	354 (24)	1668 (154)
Referential, single match	456 (33)	423 (27)	1492 (147)	475 (35)	342 (25)	1923 (168)
TOTAL TIME						
Quantified, multiple match	549 (45)	420 (35)	1205 (110)	545 (45)	402 (33)	1533 (122)
Quantified, single match	544 (42)	417 (34)	1268 (107)	566 (38)	409 (20)	1567 (121)
Referential, multiple match	552 (44)	411 (28)	1218 (107)	535 (48)	407 (31)	1472 (124)
Referential, single match	533 (38)	428 (29)	1226 (99)	569 (45)	359 (24)	1591 (139)

following a pronoun should be sensitive only to the features of potential antecedents, not their structural position. The presence of a feature-matching (albeit structurally unacceptable) candidate would facilitate processing in sentences with no grammatical antecedent. We never observed such a pattern: reading times were always longer when the main clause subject mismatched the pronoun in gender, regardless of whether the embedded subject matched the pronoun in gender.

Some researchers have argued that studies may fail to observe interference effects due to a lack of power. If the predicted pattern of facilitative interference occurred with extremely small effect sizes, we could have failed to detect it even with multiple studies. We think this is unlikely, based on comparison with a case where facilitative interference is observed readily, without large numbers of participants and items: subject-verb agreement (production: Bock and Miller, 1991; comprehension: Staub, 2009, 2010; Wagers et al., 2009). For example, in ungrammatical sentences like (9), reading times on the verb 'praise' are shorter when a plural NP ('the musicians') is present in the context (as in 9b), compared to when only singular NPs are present (as in 9a).

(9) a. *The musician who the reviewer praise so highly will probably win a Grammy.
b. *The musicians who the reviewer praise so highly will probably win a Grammy.

There is good evidence that facilitative interference in subject-verb agreement arises because the retrieval of the subject triggered by the verb is guided primarily by agreement features, not structure (Wagers et al., 2009). If so, under the Agreement First hypothesis, we would expect facilitative interference effects for subject-verb agreement and pronoun resolution to look the same, all things being equal. Of course, all things are not equal. However, the sentences we tested do favor the possibility of facilitative interference: the structurally unacceptable potential antecedent (the embedded subject) is closer to the pronoun both linearly and structurally, so it should be more highly activated in memory than the structurally acceptable potential antecedent at the point when the pronoun triggers the retrieval process. Thus, if antecedent retrieval for pronouns were unconstrained by structure, like subject retrieval for verb agreement, we would expect effect sizes at least as large as those observed in studies of subject-verb agreement, which should therefore be observable in experiments with the same power.

The lack of facilitative interference effects aligns pronouns with reflexives, which also resist interference from structurally unacceptable potential antecedents. Thus, there seems to be a broad division between the processing of agreement dependencies, which show the hallmarks of Agreement First retrieval, and the processing of referential dependencies like pronouns and reflexive (Dillon et al., 2013; but see Parker et al., 2012). Future research will need to establish ways in which the processing of referential and agreement dependencies differ (or not). This will likely provide insights into how different kinds of linguistic information are represented and accessed in memory.

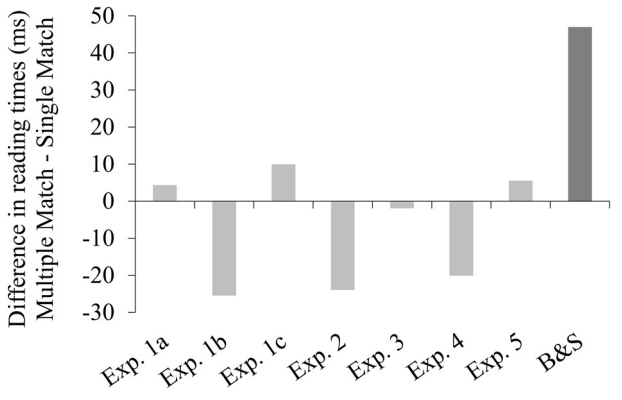


FIGURE 3 | Difference in reading times between the multiple match (main clause match/embedded match) and single match (main clause match/embedded mismatch) conditions in the *pronoun+1* region across Experiments 1–5, compared to Badecker and Straub's (2002) Experiment 1.

NO MULTIPLE MATCH OR REFERENTIAL AMBIGUITY EFFECTS

Another important finding of the current study is that we never observed the multiple match effect reported by Badecker and Straub (2002). In this case, we need not worry about a lack of power to detect the effect: all of our experiments had more participants and items than Badecker and Straub's, resulting in 1.5–5 times as many relevant data points in each experiment. **Figure 3** compares the lack of multiple match effect in the *pronoun+1* region across our five experiments to the rather sizeable effect observed in Badecker and Straub's Experiment 1.

In fact, the multiple match effect seems to be quite rare in the literature. Several other studies include the relevant comparison (Clifton et al., 1999, Experiment 3; Lee and Williams, 2008, Experiments 1 and 2; Patterson et al., 2014), but the effect has only been reported in one (Nicol, 1997; cited in Nicol and Swinney, 2003). In that study, the effect was driven by trials where the participant failed to identify the correct referent for the pronoun in a comprehension question. Nicol and Swinney (2003) therefore suggest that the presence of a multiple match effect depends on the participants' mode of reading.

We note, however, that the availability of more than one potential antecedent—even when they are all grammatically acceptable—does not necessarily lead to increased processing costs. In the possessive conditions of Experiments 3–5, the possessive pronoun 'his' was referentially ambiguous when it matched both the main clause and embedded subjects in features (e.g., 'The executive insisted that the consultant who worked on the project should prepare his client for the weekly press meeting'). This referential ambiguity was not associated with any observable processing cost: embedded subject match never impacted comprehenders' reading time profiles in the possessive pronoun condition.

Although the lack of cost for ambiguity may seem surprising, such effects are often absent in studies comparing ambiguous and unambiguous pronouns in reading comprehension (e.g., Caramazza et al., 1977; Lee and Williams, 2008; Cunnings and

Sturt, 2012; cf. MacDonald and MacWhinney, 1990; Garnham et al., 1995; Arnold et al., 2000; Nieuwland et al., 2007). The pronoun may be resolved using discourse constraints or heuristic strategies (e.g., first-mention bias: Corbett and Chang, 1983; implicit verb causality: Caramazza et al., 1977). Further, effects of referential ambiguity can also be modulated by other factors such as individual differences in working memory span (Nieuwland and Van Berkum, 2006), depth of processing (Stewart et al., 2007) and task demands (Yee and Heller, 2012). These factors might encompass the "mode of reading" idea suggested by Nicol and Swinney (2003).

Thus, even though both the main clause and embedded subjects were plausible antecedents for the possessive pronoun in the present experimental materials, various factors may have contributed to the lack of an ambiguity effect. Future work will be needed to determine how task and individual differences may explain the variation across studies. What is clear is that multiple match effects are far from being the dominant pattern in cases of multiple feature-matching intrasentential antecedents.

LIMITED UNGRAMMATICAL MATCH EFFECTS

We observed the "ungrammatical match" type of inhibitory interference in two cases: the *modified common noun* condition of Experiment 1, and the items in Experiment 2 drawn from Badecker and Straub's (2002) study. In these cases, the presence of a feature-matching but structurally unacceptable potential antecedent led to longer reading times when no grammatical antecedent was available.

Following Sturt's (2003) proposal, we suggest that initial failure to retrieve an acceptable antecedent for a reflexive or pronoun may trigger reanalysis processes leading to increased processing time when a structurally unacceptable potential antecedent matches the pronoun in features. Specifically, to recover an antecedent for the pronoun or reflexive, a feature-matching antecedent in a structurally unacceptable position may be considered. This consideration leads to increased processing time compared to the case when there are no feature-matching candidates at all to be considered. This account makes two predictions. First, sensitivity to a structurally unacceptable potential antecedent should be present only when no grammatical antecedents are available. Second, the effect should be delayed relative to the effect of grammaticality. Both of these predictions are compatible with the evidence available thus far. For instance, while Sturt (2003) observed an effect of grammaticality in first pass reading times, inhibitory interference in ungrammatical sentences was present only in second pass reading times. Correspondingly, in our Experiment 1, while there was an effect of grammaticality in the *pronoun+1* region, the inhibitory interference in ungrammatical sentences was observed only in the *pronoun+2* region.

Note, however, that this effect has only been observed in a subset of the existing studies that allowed the relevant comparison. It emerged in the modified common noun condition of our Experiment 1, but neither of the other embedded subject types, and in only half the items in Experiment 2. Other studies with similar designs have also failed to find any inhibitory interference in ungrammatical sentences (e.g., Clifton et al., 1999; Badecker and Straub, 2002; Lee and Williams, 2008). We take the inconsistency of the effect to suggest that initial failures to retrieve a structurally acceptable and feature-matching antecedent do not always trigger additional reanalysis processes, even when a feature-matching and structurally unacceptable potential antecedent is available. Future research will be needed to explore whether and how this effect may be modulated by factors such as task demands and the memory representation of the potential antecedents.

CONCLUSION

In the current study we examined whether structural constraints (Binding Principle B) impact the initial memory retrieval process alongside agreement constraints during pronoun interpretation. We argue that both the current results and previous evidence support the hypothesis that agreement features and structural constraints are used simultaneously in the process of pronoun interpretation.

ACKNOWLEDGMENTS

Earlier versions of this work have been presented at the CUNY Human Sentence Processing Conference in 2010 and at a GLOW workshop in 2012. We thank Bill Badecker for sharing experimental materials and for discussion. We would like to thank Sunyoung Lee-Ellis, Alex Drummond and Shayne Sloggett for their valuable help in carrying out these studies. We thank Chuck Clifton and Clare Patterson for helpful comments on earlier versions of this paper. This work was supported in part by NSF grant BCS-0848554 to Colin Phillips and NSF IGERT DGE-0801465 to the University of Maryland.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/Language_Sciences/10.3389/fpsyg.2014.00630/abstract

REFERENCES

Arnold, J. E., Eisenband, J. G., Brown-Schmidt, S., and Trueswell, J. C. (2000). The rapid use of gender information: evidence of the time course of pronoun resolution from eyetracking. *Cognition* 76, B13–B26. doi: 10.1016/S0010-0277(00)00073-1

Badecker, W., and Straub, K. (2002). The processing role of structural constraints on the interpretation of pronouns and anaphora. *J. Exp. Psychol. Learn. Mem. Cogn.* 28, 748–769. doi: 10.1037/0278-7393.28.4.748

Bock, K., and Miller, C. A. (1991). Broken agreement. *Cogn. Psychol.* 23, 45–93. doi: 10.1016/0010-0285(91)90003-7

Brysbaert, M., and Mitchell, D. C. (1996). Modifier attachment in sentence parsing: evidence from Dutch. *Q. J. Exp. Psychol.* 49A, 664–695. doi: 10.1080/713755636

Caramazza, A., Grober, E., Garvey, C., and Yates, J. (1977). Comprehension of anaphoric pronouns. *J. Verbal Learn. Verbal Behav.* 16, 601–609. doi: 10.1016/S0022-5371(77)80022-4

Carreiras, M., Garnham, A., Oakhill, J., and Cain, K. (1996). The use of stereotypical gender information in constructing a mental model: evidence from English and Spanish. *Q. J. Exp. Psychol.* 49, 639–663. doi: 10.1080/713755647

Chomsky, N. (1981). *Lectures On Government and Binding*. Berlin: Mouton de Gruyter.

Clackson, K., Felser, C., and Clahsen, H. (2011). Children's processing of reflexives and pronouns in English: evidence from eye-movements during listening. *J. Mem. Lang.* 65, 128–144. doi: 10.1016/j.jml.2011.04.007

Clifton, C., Frazier, L., and Deevy, P. (1999). Feature manipulation in sentence comprehension. *Rivisita Ling.* 11, 11–39.

Clifton, C., Kennison, S., and Albrecht, J. (1997). Reading the words 'her', 'him' and 'his': Implications for parsing principles based on frequency and on structure. *J. Mem. Lang.* 36, 276–292. doi: 10.1006/jmla.1996.2499

Corbett, A. T., and Chang, F. R. (1983). Pronoun disambiguation: accessing potential antecedents. *Mem. Cogn.* 11, 283–294. doi: 10.3758/BF03196975

Cunnings, I., and Sturt, P. (2012). "The time-course of reference resolution in picture noun phrases: evidence from eye-movements during reading." Poster, 25th Annual CUNY Human Sentence Processing Conference (New York, NY).

Dillon, B., Mishler, A., Sloggett, S., and Phillips, C. (2013). Contrasting interference profiles for agreement and anaphora: experimental and modeling evidence. *J. Mem. Lang.* 69, 85–103. doi: 10.1016/j.jml.2013.04.003

Ehrlich, K., and Rayner, K. (1983). Pronoun assignment and semantic integration during reading: eye movements and immediacy of processing. *J. Verbal Learn. Verbal Behav.* 22, 75–87. doi: 10.1016/S0022-5371(83)80007-3

Evans, G. (1980). Pronouns. *Ling. Inq.* 11, 337–362.

Garnham, A., Oakhill, J., Ehrlich, M.-F., and Carreiras, M. (1995). Representations and processes in the interpretation of pronouns: new evidence from Spanish and French. *J. Mem. Lang.* 34, 41–62. doi: 10.1006/jmla.1995.1003

Gordon, P. C., Hendrick, R., and Johnson, M. (2001). Memory interference during language processing. *J. Exp. Psychol. Learn. Mem. Cogn.* 27, 1411–1423. doi: 10.1037/0278-7393.27.6.1411

Grodzinsky, Y., and Reinhart, T. (1993). The innateness of binding and coreference. *Ling. Inq.* 24, 69–101.

Just, M. A., Carpenter, P. A., and Wooley, J. D. (1982). Paradigms and processes in reading comprehension. *J. Exp. Psychol. Gen.* 111, 228–238. doi: 10.1037/0096-3445.111.2.228

Kennison, S. M. (2003). Comprehending the pronouns *her*, *him*, and *his*: Implications for theories of referential processing. *J. Mem. Lang.* 49, 335–352. doi: 10.1016/S0749-596X(03)00071-8

Kennison, S. M., and Trofe, J. L. (2003). Comprehending pronouns: a role for word-specific gender stereotype information. *J. Psycholinguist. Res.* 32, 355–378. doi: 10.1023/A:1023599719948

Keppel, G., and Underwood, B. J. (1962). Proactive-inhibition in short-term retention of single items. *J. Verbal Learn. Verbal Behav.* 1, 153–161. doi: 10.1016/S0022-5371(62)80023-1

Lee, M.-W., and Williams, J. N. (2008). *The Role of Grammatical Constraints in Intra-Sentential Pronoun Resolution*. London: London Metropolitan University/Cambridge University manuscript.

Lewis, R., and Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cogn. Sci.* 29, 375–419. doi: 10.1207/s15516709cog0000_25

MacDonald, M., and MacWhinney, B. (1990). Measuring inhibition and facilitation from pronouns. *J. Mem. Lang.* 29, 469–492. doi: 10.1016/0749-596X(90)90067-A

Martin, A. E., and McElree, B. (2008). A content-addressable pointer mechanism underlies comprehension of verb-phrase ellipsis. *J. Mem. Lang.* 58, 879–906. doi: 10.1016/j.jml.2007.06.010

Nicol, J. (1997). *Processing Pronouns in Text: Effects of Feature Congruence of Potential Antecedents*. Tucson, AZ: Manuscript, University of Arizona.

Nicol, J., and Swinney, D. (2003). "The psycholinguistics of anaphora," in *Anaphora: A Reference Guide*, ed A. Barss (Malden, MA: Blackwell), 72–104.

Nicol, J., and Swinney, D. A. (1989). The role of structure in coreference assignment during sentence comprehension. *J. Psycholinguist. Res.* 18, 5–19. doi: 10.1007/BF01069043

Nieuwland, M. S., Otten, M., and Van Berkum, J. J. A. (2007). Who are you talking about? Tracking discourse-level referential processing with event-related brain potentials. *J. Cogn. Neurosci.* 19, 228–236. doi: 10.1162/jocn.2007.19.2.228

Nieuwland, M. S., and Van Berkum, J. J. A. (2006). Individual differences and contextual bias in pronoun resolution: evidence from ERPs. *Brain Res.* 1118, 155–167. doi: 10.1016/j.brainres.2006.08.022

Osterhout, L., and Mobley, L. A. (1995). Event-related brain potentials elicited by failure to agree. *J. Mem. Lang.* 34, 739–773. doi: 10.1006/jmla.1995.1033

Parker, D., Lago, S., and Phillips, C. (2012). *Retrieval Interference in the Resolution of Anaphoric PRO. Talk given at the Timing of Grammar Workshop at the 35th GLOW Colloquium*, Potsdam.

Patterson, C., Trompelt, H., and Felser, C. (2014). The online application of binding condition B in native and non-native pronoun resolution. *Front. Psychol.* 5:147. doi: 10.3389/fpsyg.2014.00147

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychol. Bull.* 124, 372–422. doi: 10.1037/0033-2909.124.3.372

Rayner, K., and Duffy, S. A. (1986). Lexical complexity and fixation times in reading: effects of word frequency, verb complexity, and lexical ambiguity. *Mem. Cogn.* 14, 191–201. doi: 10.3758/BF03197692

Rayner, K., and Pollatsek, A. (1989). *The Psychology of Reading.* Englewood Cliffs, NJ: Prentice Hall.

Reinhart, T. (1983). *Anaphora and Semantic Interpretation.* London: Croom Helm.

Rohde, D. (2003). *Linger (Version 2.94)* [*Software*]. Available online at: http://tedlab.mit.edu/∼dr/Linger/

Runner, J. T., Sussman, R. S., and Tanenhaus, M. K. (2006). Processing reflexives and pronouns in picture noun phrase. *Cogn. Sci.* 30, 193–241. doi: 10.1207/s15516709cog0000_58

Staub, A. (2009). On the interpretation of the number attraction effect: response time evidence. *J. Mem. Lang.* 60, 308–327. doi: 10.1016/j.jml.2008.11.002

Staub, A. (2010). Response time distributional evidence for distinct varieties of number attraction. *Cognition* 114, 447–454. doi: 10.1016/j.cognition.2009.11.003

Stewart, A. J., Holler, J., and Kidd, E. (2007). Shallow processing of ambiguous pronouns: evidence for delay. *Q. J. Exp. Psychol.* 60, 1680–1696. doi: 10.1080/17470210601160807

Sturt, P. (2003). The time-course of the application of binding constraints in reference resolution. *J. Mem. Lang.* 48, 542–562. doi: 10.1016/S0749-596X(02)00536-3

Sturt, P. (2013). "Syntactic constraints on referential processing," in *Sentence Processing*, ed R. P. G. van Gompel (East Sussex: Psychology Press), 136–159.

Van Berkum, J. J. A., Zwitserlood, P., Bastiaansen, M. C. M., Brown, C. M., and Hagoort, P. (2004). "So who's "he" anyway? Differential ERP and ERSP effects of referential success, ambiguity and failure during spoken language comprehension," in *Annual Meeting of the Cognitive Neuroscience Society (CNS-2004)* (San Francisco, CA).

Wagers, M., Lau, E., and Phillips, C. (2009). Agreement attraction in comprehension: representations and processes. *J. Mem. Lang.* 61, 206–237. doi: 10.1016/j.jml.2009.04.002

Walker, N., Jones, J. P., and Mar, H. H. (1983). Encoding processes and the recall of text. *Mem. Cogn.* 11, 275–282. doi: 10.3758/BF03196974

Yee, E., and Heller, D. (2012). "Lookingmore when you know less: goal-dependent eye movements during reference resolution," in *Poster Presented at the Annual Meeting of the Psychonomic Society*, (Minneapolis, MN).

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 08 April 2014; accepted: 03 June 2014; published online: 27 June 2014.
*Citation: Chow W-Y, Lewis S and Phillips C (2014) Immediate sensitivity to structural constraints in pronoun resolution. Front. Psychol. **5**:630. doi: 10.3389/fpsyg.2014.00630*
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology.
Copyright © 2014 Chow, Lewis and Phillips. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 25 February 2014
doi: 10.3389/fpsyg.2014.00147

The online application of binding condition B in native and non-native pronoun resolution

Clare Patterson, Helena Trompelt and Claudia Felser*

Potsdam Research Institute for Multilingualism, Faculty of Human Sciences, University of Potsdam, Potsdam, Germany

Edited by:
Matthew Wagers, University of California, Santa Cruz, USA

Reviewed by:
Brian Dillon, University of Massachusetts Amherst, USA
Barbara Hemforth, Centre National de la Recherche, France

****Correspondence:***
Claudia Felser, Potsdam Research Institute for Multilingualism, Faculty of Human Sciences, University of Potsdam, Haus 2, Karl-Liebknecht-Straße 24-25, 14476 Potsdam, Germany
e-mail: felser@uni-potsdam.de

Previous research has shown that anaphor resolution in a non-native language may be more vulnerable to interference from structurally inappropriate antecedents compared to native anaphor resolution. To test whether previous findings on reflexive anaphors generalize to non-reflexive pronouns, we carried out an eye-movement monitoring study investigating the application of binding condition B during native and non-native sentence processing. In two online reading experiments we examined when during processing local and/or non-local antecedents for pronouns were considered in different types of syntactic environment. Our results demonstrate that both native English speakers and native German-speaking learners of English showed online sensitivity to binding condition B in that they did not consider syntactically inappropriate antecedents. For pronouns thought to be exempt from condition B (so-called "short-distance pronouns"), the native readers showed a weak preference for the local antecedent during processing. The non-native readers, on the other hand, showed a preference for the matrix subject even where local coreference was permitted, and despite demonstrating awareness of short-distance pronouns' referential ambiguity in a complementary offline task. This indicates that non-native comprehenders are less sensitive during processing to structural cues that render pronouns exempt from condition B, and prefer to link a pronoun to a salient subject antecedent instead.

Keywords: pronoun resolution, binding, sentence processing, eye-movement monitoring, bilingualism, English

INTRODUCTION

During language comprehension linguistic structure must be encoded, and rapid decisions about dependency formation such as pronominal reference need to be made. Whilst it is generally agreed that processing a pronoun involves the retrieval or reactivation of an antecedent (either explicit or understood from the context), there is no clear consensus on the precise role that structural constraints play in this retrieval process.

Much of the recent debate in this area has been around the memory processes involved in long-distance dependencies, with particular reference to reflexive processing and subject-verb agreement (see Dillon, 2011, for an overview). One view is that reflexive processing in particular involves a structure-sensitive search, so that the target of the retrieval is identified through its position in the linguistic structure (Dillon, 2011; Dillon et al., 2013). An opposing view is that retrieval for reflexives exploits the cues carried on prior representations, so that, for example, a singular, masculine reflexive triggers a search for representations carrying the features *singular* and *masculine*. Importantly, this second approach predicts that retrieval interference is possible from antecedents that are not structurally licensed (e.g., Patil, 2012).

As far as pronouns[1] are concerned, structure alone is not sufficient to uniquely identify a referent, and the interpretation of pronouns is subject not only to structural constraints but also a range of discourse constraints, distinguishing it from reflexive interpretation. Despite this, there is debate around the primacy of the structure-sensitive constraint known as condition B of the Binding Theory (Chomsky, 1981). Condition B restricts the interpretation of pronouns such that a pronoun cannot refer to a c-commanding antecedent within its local binding domain[2]. For example in (1), the direct object pronoun *him* cannot refer to *David* but it can refer to *Nick*. The embedded subject *David* is "inaccessible" as a binder for *him* because the two are coarguments of the same predicate.

(1) Nick$_i$ thinks that David$_k$ likes him$_{i,*k}$

Whether or not condition B can be defined in purely structural terms, though, is debatable. Binding Theory assumes an exclusion on the basis of structural position, but other views involve excluding the inaccessible antecedent on mainly pragmatic grounds (Huang, 1994) or by comparing two alternative semantic sentence representations (Reinhart, 1983; Reuland, 2001, 2011). In this paper, the term "condition B" will henceforth be used as a general

[1]For simplicity, reflexive pronouns will henceforth be referred to as reflexives and non-reflexive pronouns will be referred to as pronouns.

[2]The term "c-command" refers to a particular structural relationship between constituents that is defined in terms of hierarchical dominance (Reinhart, 1983). If, for example, a pronoun is contained within a category that is dominated by the same branching node that immediately dominates another NP, that NP is said to c-command the pronoun.

term to express the exclusion of inaccessible antecedents for pronouns, rather than endorsing a particular theoretical approach.

According to the *binding as initial filter* (BAIF) hypothesis by Nicol and Swinney (1989), condition B is used to exclude inaccessible antecedents from an early stage of processing. In the case of canonical condition B environments exemplified in (1), the local (inaccessible) antecedent would be immediately ruled out and would not influence the parse at any point. That is, condition B should prevent consideration of inaccessible antecedents even when they carry number or gender features that match those of the pronoun. Evidence for this hypothesis came from several cross-modal priming studies which found antecedent reactivation effects only for accessible but not for inaccessible antecedents (Nicol and Swinney, 1989). Further support for this hypothesis mainly comes from negative evidence in self-paced reading studies, i.e., a lack of a demonstrable effect from manipulating the gender or number features of an inaccessible antecedent. When no effect is found, the assumption is that the inaccessible antecedent is not being considered. Negative evidence of this kind has been found by Clifton et al. (1997, 1999).

A variant of the BAIF hypothesis is the idea that binding constraints may act as defeasible filters, with inaccessible antecedents potentially being considered at later processing stages. Evidence in support of this comes from an eye-movement study on English reflexives reported by Sturt (2003).

An alternative to both the BAIF and the defeasible filter hypotheses was put forward by Badecker and Straub (2002). They suggested that multiple cues or constraints that are relevant for pronoun processing (including structural constraints) all contribute in parallel, positively or negatively, to an antecedent's activation. Thus, positive activation from one constraint may be canceled out by inhibition from another. Due to this parallel activation/inhibition, the feature match or mismatch of an inaccessible antecedent will have an influence on processing, in direct contrast to the BAIF hypothesis. Badecker and Straub found that the reading times in regions following a pronoun were longer when both the accessible and inaccessible antecedents matched in gender with the pronoun, compared to when only the accessible antecedent matched. They suggested that all feature-matching referents, whether accessible or inaccessible according to Binding Theory, are evaluated. Further evidence that the inaccessible antecedent is not immediately excluded from consideration comes from Clackson et al.'s (2011) eyetracking-during-listening study. Adult participants' eye gaze patterns revealed that they experienced interference from a gender-matching but structurally inaccessible antecedent after encountering a pronoun. Such evidence can be characterized as supporting a feature-based antecedent search as proposed by Badecker and Straub.

Thus the current evidence bearing on the BAIF with respect to pronouns appears to point in two directions, and there is as yet no clear consensus on whether or not condition B gates access to certain potential antecedents during processing.

In order to establish a broader picture of the mechanisms behind pronoun processing, environments which are exempt from condition B should also be considered. While there are, of course, many syntactic environments in which condition B plays no role (because there is no inaccessible antecedent to exclude) the use of specific exceptions to condition B is more informative. In these cases, condition B *should* apply to rule out a local antecedent, but it does not. The exception that is made use of in the current study is the case of so-called "short distance pronouns" (SDPs). In certain structures such as (2) below, a local c-commanding noun phrase (NP) *can* be interpreted as the antecedent for the pronoun, and it seems that both reflexives and pronouns can appear in these positions (Lees and Klima, 1963, among others).

(2) Nick$_i$ saw David$_k$ put the cat beside him$_{i/k}$.

Possible reasons as to why SDPs seem exempt from condition B include proposals to the effect that prepositional phrases such as *beside him* in (2), or certain kinds of (verb phrase internal) aspectual phrases, can be binding domains (Hestvik, 1991; Tenny, 2004). Under this view, the local subject *David* in (2) is outside the pronoun's binding domain and is thus allowed to bind it without condition B being violated. More widely accepted is the proposal that the scope of condition B should be restricted to anaphoric dependencies involving coarguments (e.g., Reinhart and Reuland, 1993). This also allows for the pronoun *him* in (2) to enter into a referential dependency with the local subject *David* because the two are not in fact arguments of the same predicate. Alternatively, Rooryck and Vanden Wyngaerd (2011) have proposed that rather than being bound by the local subject NP, SDPs are variable-bound by a covert operator located at the left clausal periphery. Regardless of which of the above theoretical accounts is ultimately deemed preferable, recognizing syntactic environments in which local coreference is permitted requires sensitivity to the relevant structural differences between standard condition B environments such as (1) above and SDP environments such as (2).

Exceptions such as SDPs, then, make a good comparison point with canonical condition B environments because their structure is quite similar, but they can reveal how pronoun processing unfolds when condition B appears not to apply. This may, for example, shed further light on possible feature-driven processes, or reveal an underlying sensitivity to the linear ordering of antecedents, as has been found in certain syntactic environments (Cunnings et al., 2014). The online processing of pronouns in SDP environments has rarely been investigated. Experimental evidence for the referential ambiguity of SDPs has been reported by Sekerina et al. (2004). Using eyetracking-during-listening, they examined English-speaking children and adults' processing of questions such as (3) below.

(3) Which picture shows that the boy has placed the box behind himself/him?

Participants had to choose between two alternative pictures, one of which showed the box being located behind a boy (= the sentence-internal referent) and one in which it was located behind an adult male character (= the sentence-external referent). Participants' eye-gaze patterns showed a reduced proportion of looks to the picture corresponding to sentence-internal

reference resolution in the pronoun compared to the reflexive condition, suggesting that the alternative, sentence-external antecedent was more likely to be considered in the pronoun than in the reflexive condition. In a corresponding offline task, the adult participants showed a strong across-the-board preference for sentence-internal antecedents. The focus of Sekerina et al.'s study was on sentence internal vs. external antecedents, and possible differences between antecedent preferences for reflexives vs. pronouns. It does not give a broader picture of pronoun processing in environments with two potential sentence-internal antecedents, although it is interesting to note that pronouns appear to be more flexible in their interpretation than reflexives. In our current study, we use SDP environments such as (2) as a contrast to condition B environments. The crucial factor here is that both antecedents are thought to be accessible to the pronoun.

There are other environments which appear to be exempt from condition B; so-called "picture noun phrases" are a well-studied example (Runner et al., 2003; Kaiser et al., 2009, among others)[3] . The main finding from these studies regarding pronouns is that non-structural factors such as semantic role information are important. Most relevant to the current study, however, is that previous studies have shown that native English-speaking comprehenders are aware of the referential ambiguity of binding-theory exempt pronouns during processing.

NON-NATIVE PROCESSING OF PRONOMINAL ANAPHORS

It is not only exceptions to condition B that can provide a broader picture about the processing of pronouns. The processing profiles of different populations, in this case non-native speakers, can also be informative. Models of parsing, particularly those that are closely tied to aspects of general cognition, should be able to account not only for native language processing but also for processing in a non-native language. Additionally, non-native speakers have been shown in previous studies to take a more discourse-driven strategy than native speakers during the processing of, for example, reflexives (Felser and Cunnings, 2012), findings which appear to challenge the universal validity of serial or syntax-first models that were proposed on the basis of monolingual processing data.

Most previous research on non-native anaphor resolution has examined learners' knowledge of binding using offline judgment or antecedent choice tasks. Unlike the developmental delay of condition B that has been reported in the child language acquisition literature (e.g., Chien and Wexler, 1990), the application of binding condition B appears to be relatively unproblematic in the post-childhood acquisition of non-native speakers (henceforth L2s). White (1998), for example, reports that even intermediate-level L2 learners of English patterned with English native speakers in a truth-value judgment task in disallowing local antecedents for pronouns. Using a multiple-choice antecedent identification task, Bertenshaw (2009) found that native Japanese-speaking learners of English correctly rejected inaccessible antecedents for pronouns 92.8% of the time, a figure that compares favorably with the native speaker controls' correct rejection rate of 87.5%. Similarly high accuracy rates have been reported by Cook (1990).

Conversely, little is known about whether or when binding constraints are applied during online L2 processing. L2s have been claimed to show reduced sensitivity to syntactic information during processing compared to native speakers (henceforth L1s), and difficulty establishing structurally mediated discontinuous dependencies in a native-like way (Clahsen and Felser, 2006). However, a reduced ability to process syntactically mediated dependencies may affect L2 online interpretation of reflexives more than the ability to interpret pronouns, all other things being equal. This is under the assumption that binding of argument reflexives is contingent on mechanisms of syntactic computation, whereas non-reflexive pronouns can also be linked to an antecedent via discourse-based coreference assignment (e.g., Reuland, 2001, 2011).

While L1 speakers appear to respect condition A of the Binding Theory (which states that reflexives must be locally bound) from the earliest measurable point in processing (Sturt, 2003; Xiang et al., 2009), a different picture emerges in L2 processing. Felser et al. (2009) report evidence from timed grammaticality judgments and eye-movement monitoring showing that native Japanese speakers experienced competition from inaccessible antecedents for English argument reflexives during processing, despite demonstrating native-like knowledge of binding condition A in complementary offline tasks. Felser and Cunnings (2012) further explored the interaction of structural and discourse factors in non-native anaphor resolution by examining native German speakers' processing of English reflexives. Two eye-movement monitoring experiments were carried out using sentences such as (4a) and (4b) in a gender-mismatch paradigm (compare e.g., Sturt, 2003).

(4a) James has worked at the army hospital for years. He noticed that the soldier had wounded himself while on duty in the Far East.

(4b) James has worked at the army hospital for years. The soldier that he treated on the ward wounded himself while on duty in the Far East.

The L2s' reading-time patterns differed from the L1s' in that they initially showed unmodulated main effects of the inaccessible antecedent's gender only. This was the case both for sentences like (4a), in which the inaccessible antecedent (the pronoun *he*) c-commands the reflexive, and for sentences such as (4b), where it does not. Only in later measures and/or sentence regions did the L2 speakers pattern with the L1 controls in showing main effects of the accessible antecedent's gender. Taken together, these results indicate that unlike L1s, L2 speakers do not immediately apply binding condition A during processing but initially try to link argument reflexives to the most discourse-prominent antecedent via coreference assignment instead.

To our knowledge, the timing of binding condition B during L2 pronoun processing has never been investigated. L2 processing studies on pronoun resolution have focused on discourse anaphors rather than bound pronouns. The findings from these studies suggest that L2s can use information-structural cues such

[3] A typical example of a picture noun phrase is "Nick's picture of himself/him," where both the reflexive and the pronoun can be understood as referring to Nick.

as focus to guide pronoun resolution (Ellert, 2010) and may experience more competition than L1s in the presence of more than one feature-matching discourse antecedent (Roberts et al., 2008). Roberts et al. examined the role of contextual information in native Turkish and German speakers' real-time comprehension of ambiguous pronouns in L2 Dutch also using eye-movement monitoring. The two L2 groups patterned together in showing elevated total and second-pass reading times at the pronoun region when two (rather than only one) matching antecedents were present in the sentence-external discourse. The native Dutch controls, on the other hand, were not measurably distracted by the presence of another matching discourse antecedent.

Two experiments are described below which aim to explore the application and timing of condition B during L1 and L2 sentence processing using eye-movement monitoring during reading. To obtain information about participants' ultimate interpretation preferences, the two online reading experiments are complemented by an offline antecedent choice task (Experiment 1). Our first eye-movement experiment (Experiment 2) examines readers' processing of canonical condition B sentences such as (1) above, while Experiment 3 examines online pronoun resolution in SDP environments such as (2). Experiments 2 and 3 were run concurrently during the same experimental session. All experimental sentences contained one pronoun and two potential antecedents, local and non-local.

The following specific questions will be explored:

- Does condition B immediately exclude inaccessible antecedents from the candidate set?
- Does the order/timing of considering the two antecedents differ according to whether or not condition B applies?
- Are there any L1/L2 differences in the application of condition B?

We begin by reporting the results from the offline questionnaire study.

MATERIALS AND METHODS, EXPERIMENT 1

The purpose of Experiment 1, an offline antecedent choice task, was to examine the offline antecedent choices of L1 and L2 participants in the two different syntactic environments under investigation, in the absence of any time pressure. This is especially important for the SDPs because they are thought to be ambiguous.

PARTICIPANTS

The L1 group comprised 83 participants, all of whom reported that they were native speakers of English (33 males, mean age 40 years, range 19–72 years). They were recruited via email and word of mouth to people who were known to be native speakers of English, and through an advertisement on an English-language forum on the internet. The L2 group comprised 35 native German-speaking students at the University of Potsdam (10 males, mean age 22.2, range 19–37 years) who had learned English as their second language at school [4] . All L2 speakers participated in a subpart of the grammar section of the Oxford Placement Test (OPT; Allan, 2004). Their mean score was 39/50 (*proficient*), range 30–48 (*lower intermediate* to *expert user*).

MATERIALS

The materials were ten sentences in which pronoun interpretation was constrained by condition B such as (5) below, and ten sentences containing SDPs such as (6).

(5) The boy remembered that Matthew had bought him a new computer game.
(6) Harry heard William pull the curtain around him in the quiet hospital ward.

The critical sentences all contained a direct object pronoun and two potential antecedents which matched the pronoun in gender. In (5), the local antecedent *Matthew* is ruled out by condition B, whereas in (6), it should be possible for the pronoun to be linked to either the non-local antecedent (*Harry*) or the local one (*William*). Within each experimental condition an equal number of masculine and feminine pronouns was used. We also took care to create scenarios in which the local and the non-local antecedent were equally plausible as antecedents for the pronoun.

The experimental sentences were mixed and pseudo-randomized with 22 filler sentences containing ambiguous or unambiguous pronouns and reflexives in different syntactic environments, yielding a total of 42 items.

PROCEDURE

The questionnaire was administered via the internet using SurveyGizmo (surveygizmo.com). The L1 group completed the questionnaire remotely. The L2 participants completed the questionnaire as part of the experimental session for online Experiments 2 and 3, after they had finished the online element. Because the experimenters had less direct control over the conditions in which the L1 participants did the questionnaire, a larger number of L1 participants were included to increase the reliability of the responses[5] .

All participants were instructed to read each sentence carefully and decide who the pronoun probably referred to. The use of *probably* takes account of the fact that another interpretation is possible, although unlikely. After each sentence the same question appeared: "Who does [pronoun] refer to?" In each case participants were given three choices as in (7) below.

(7) The boy remembered that Matthew had bought him a new computer game. *Who does "him" refer to?*

- The boy
- Matthew
- Either

[4]The reason for the larger number of participants in the L1 group compared to the L2 group is discussed in the "procedure" section.

[5]Additionally, responses of both L1 and L2 participants to unambiguous filler items were checked to ensure that the participants had understood the task. The percentage of correct answers was 98% for the L1 and 93% for the L2 group.

The order of the two antecedent responses was varied throughout the questionnaire, and the *either* option always appeared at the bottom.

RESULTS, EXPERIMENT 1

One item was removed from the analysis of the condition B sentences because it could be construed as being ambiguous. **Figures 1**, **2** show the percentage of responses to the canonical condition B structures and the SDP structures, for each group.

For the canonical condition B structures (**Figure 1**), the preference for the non-local (accessible) antecedent is very clear in both groups; they both chose this option above 90% of the time (L1 98%, L2 91%). A 3 × 2 ANOVA with an appropriate logistic transformation (Agresti, 2002) of the response rates of each type (*non-local*, *local*, and *either*) showed a main effect of antecedent choice [$F_{1(2,\ 232)} = 2110.3$, $p < 0.0001$; $F_{2(2,\ 32)} = 349.2$, $p < 0.0001$] and an interaction between antecedent choice and group [$F_{1(2,232)} = 19.4$, $p < 0.001$; $F_{2(2,\ 32)} = 49.6$, $p < 0.001$]. The L1 group chose the *non-local* response more often than the L2 group [$t_{1(116)} = 5.3$, $p < 0.001$; $t_{2(16)} = 19.8$, $p < 0.001$]. Nevertheless, within-group *t*-tests confirmed that in both groups the percentage of *non-local* responses was significantly higher than that of *local* responses [L1: $t_{1(82)} = 55.8$, $p < 0.001$; $t_{2(8)} = 20.8$, $p < 0.001$; L2: $t_{1(34)} = 18.5$, $p < 0.001$; $t_{2(8)} = 12.2$, $p < 0.001$] and *either* responses [L1: $t_{1(82)} = 68.8$, $p < 0.001$; $t_{2(8)} = 21.4$, $p < 0.001$; L2:$t_{1(34)} = 26.9$, $p < 0.001$; $t_{2(8)} = 9.1$, $p < 0.001$].

Compared to the canonical condition B structures, for the SDP structures (**Figure 2**) there was more variability in the two groups' responses. There was a numerical preference in both groups for choosing the *either* response indicating that the pronoun was ambiguous (L1 60%; L2 43%). A 3 × 2 ANOVA showed a main effect of antecedent choice [$F_{1(2,\ 232)} = 24.0$, $p < 0.0001$; $F_{2(2,\ 36)} = 16.7$, $p < 0.0001$] and an interaction between antecedent choice and group [$F_{1(2,232)} = 6.6$, $p < 0.01$; $F_{2(2,\ 36)} = 4.1$, $p < 0.05$]. For the L1 group the *either* option was chosen significantly more often than both the *local* response [$t_{1(82)} = 7.3$, $p < 0.001$; $t_{2(9)} = 6.1$, $p < 0.001$] and the *non-local* response [$t_{1(82)} = 6.6$, $p < 0.001$; $t_{2(9)} = 4.3$, $p < 0.01$]. For the L2 group the *either* response was chosen significantly more often than the *local* response [$t_{1(34)} = 2.2$, $p < 0.05$; $t_{2(9)} = 3.9$, $p < 0.01$] but not significantly more than the *non-local* response [$t_{1(34)} = 0.4$, $p = 0.6$; $t_{2(9)} = 1.2$, $p < 0.2$]. When the *either* option was not chosen, the L1 group chose the local and non-local antecedent at roughly the same rate (18 and 22% respectively); a *t*-test showed no significant difference between these two response rates [$t_{1(82)} = 2.0$, $p = 0.5$; $t_{2(9)} = 0.3$, $p = 0.7$]. The L2 group, however, chose the non-local antecedent more often than the local antecedent (34 and 21% respectively), a difference which proved (marginally) significant in a *t*-test [$t_{1(34)} = 2.7$ $p < 0.01$; $t_{2(9)} = 2.1$, $p = 0.063$]. There was a significant negative correlation between participants' OPT scores and *non-local* antecedent choice rates for the SDP structures [$r_{(35)} = -0.35$, $p < 0.05$], however, no participant categorically chose *non-local* responses.

EXPERIMENT 1 SUMMARY

Participants' responses to the canonical condition B structures were highly consistent for both groups. While participants in the L1 group were overall more likely than those in the L2 group to choose the non-local antecedent, there was an overwhelming preference for the non-local antecedent in both groups, almost to the exclusion of any other response. This demonstrates that both L1 and L2 speakers are fully aware of the inaccessibility of the local antecedent, although the L1 group demonstrated more certainty than the L2 group. Participants' responses to the SDP structures were quite different, with the pronoun's ambiguity reflected in their antecedent choices. Both groups chose *either* at the highest

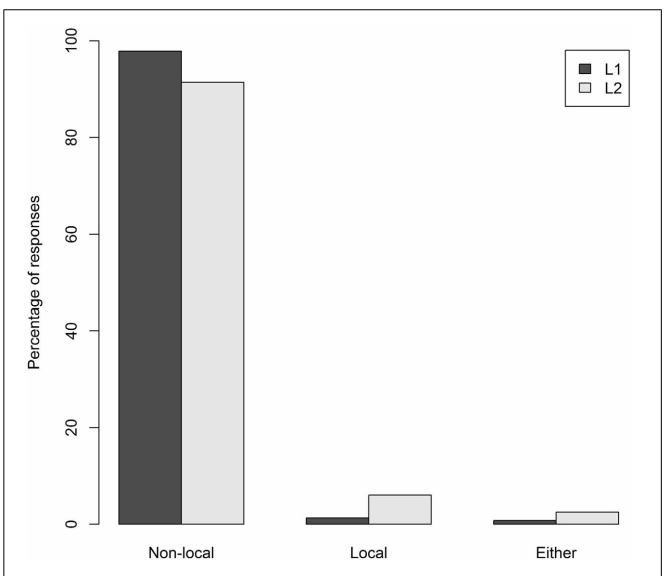


FIGURE 1 | Percentage of responses per group to the canonical condition B structures.

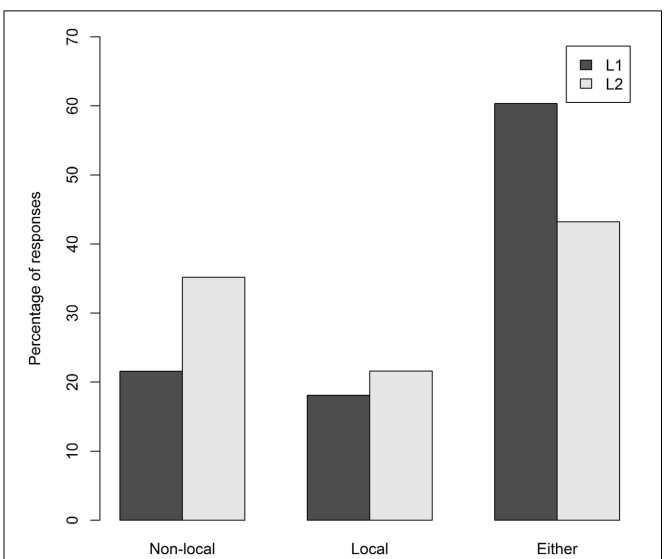


FIGURE 2 | Percentage of responses per group to the SDP structures.

rate, although the L2 group's rate of *either* responses was not significantly higher than their non-local responses. When choosing one particular antecedent (instead of the *either* option), the L1 group did not show a preference for either the local or non-local antecedent, whereas the L2 group displayed a slight preference for the non-local antecedent. This preference was related to OPT score; the lower a participant's OPT score, the more likely they were to choose the *non-local* referent. This may suggest that awareness of the ambiguity of SDPs increases with knowledge of English [6] . Taken together, the responses show that participants responded in line with condition B where appropriate, and displayed awareness of the ambiguity of SDPs.

MATERIALS AND METHODS, EXPERIMENT 2

Experiment 2 was designed to investigate the online application of condition B in sentences where only the local antecedent was accessible. We specifically sought to investigate whether L1 and/or L2 comprehenders would experience interference from the inaccessible antecedent at any point during processing.

PARTICIPANTS

The L1 participants were 34 native speakers of English (11 males) who were recruited from the University of Essex (UK) and the surrounding community. Their mean age was 25.9 (range: 18–54), and all confirmed that English was their first language. The L2 group consisted of 34 of the 35 native German speakers who took part in Experiment 1 (10 males, mean age 22.8, range 19–37), all of whom had learned English as their second language at school starting at the age from 5 to 13 (mean: 9.6, *SD*: 1.7). Their mean OPT score was 39/50 (*proficient*), range 30–48 (*lower intermediate* to *expert user*). All participants were paid for their participation, and all had normal or corrected-to-normal vision.

MATERIALS

Twenty-four experimental items were constructed. They were composed of three sentences: a lead-in sentence, a critical sentence that contained the pronoun and two potential antecedent NPs that were both proper names, and a wrap-up sentence. The gender match between the two names and the pronoun was manipulated to create three experimental conditions as shown in (8a–c) below[7] .

(8) Band practice was beginning to get rather dull.

(a) *Double match condition*
John remembered that Mark had taught him a new song on the guitar.

(b) *Local mismatch condition*
John remembered that Jane had taught him a new song on the guitar.

(c) *Non-local mismatch condition*
Jane remembered that John had taught him a new song on the guitar.
That really lifted everyone's spirits!

The names were matched in letter and syllable length, and were either typical male or typical female names (i.e., names that are not normally used for both genders). The names were counterbalanced across items to control for any potential frequency effects. The first name (the non-local antecedent) was always the main clause subject and was an accessible antecedent by virtue of being outside the local binding domain. The second name (the local antecedent) was always the subject of an embedded complement clause and a coargument of the pronoun. It was thus an inaccessible antecedent for the pronoun according to condition B. Half the pronouns were masculine and half feminine, and they were always object pronouns.

The experimental items were distributed across three presentation lists using a Latin-square design, and mixed and pseudo-randomized with 18 experimental items from Experiment 3 (described below) and 44 additional filler items, resulting in 86 items per list in total. The set of fillers included eight pseudo-fillers which were structurally similar to the experimental items but contained reflexive rather than non-reflexive pronouns, and another eight in which the structurally illicit antecedent for the pronoun was placed first. This was to ensure that participants were exposed to enough items that were similar to the experimental items but different in crucial factors (type of referring expression and position of the antecedent), to prevent them from developing expectations about the pronoun–antecedent relationships under investigation. Binary yes/no comprehension questions followed two thirds of the 86 items in each list, including the experimental items, to ensure that participants were paying attention and reading the items properly. A few of the comprehension questions following filler items directly probed the referent of a pronoun, to encourage participants to fully process the pronouns that they read. The experiment began with the presentation of six practice items to familiarize participants with the procedure, two of which were followed by a question.

PREDICTIONS

In the light of the different proposals regarding the primacy of condition B during processing, the following predictions can be made.

BAIF hypothesis

If structural information helps to rule out inaccessible antecedents at an early point, only the accessible (non-local) antecedent should be considered. This predicts that there will be a slow-down in reading times in condition (8c) (non-local mismatch) compared to the other two conditions. In addition, because the inaccessible antecedent is excluded from consideration on structural grounds, there should be no difference between condition (8a) (double match) and (8b) (local

[6] We remain cautious about this observation, firstly because of the limited range of the OPT scores, and secondly because the OPT gives placement scores (sufficient to demonstrate that all L2 participants were competent in English), rather than a direct and thorough measure of proficiency. Additionally, we did not set out to test the effect of proficiency here, and have made no specific predictions.

[7] A potential fourth condition in which neither name matched the pronoun in gender was not included in order to avoid presenting participants with too many unresolvable pronouns, which could have drawn their attention to the pronouns and encouraged strategic reading behavior. This is also the case for the materials of Experiment 3.

mismatch) because participants should not be sensitive to the gender of the inaccessible antecedent.

Defeasible filter hypothesis

Following Sturt's (2003) results for reflexives, it is possible that binding conditions act early to include or exclude certain antecedents, but the inaccessible antecedents are considered at a later point of processing. The defeasible filter account therefore predicts longer reading times for condition (8c), followed later by effects of the inaccessible antecedent which could manifest as either longer reading times in condition (8b) or as a competition effect with differences between condition (8a) and the other two conditions.

Feature-match hypothesis

If condition B does not immediately overrule other cues, then processing should also be sensitive to the gender features of the inaccessible antecedent initially. Readers may only home in on the accessible (i.e., the non-local) antecedent at later processing stages or sentence regions. Following Badecker and Straub (2002), if all antecedents with matching morphosyntactic or semantic features are activated on encountering the pronoun, regardless of the structural accessibility of the antecedents, participants might experience "retrieval interference" (Gordon et al., 2001; Lewis and Vasishth, 2005; Van Dyke, 2007) indexed as increased reading times when both antecedents match the pronoun in gender (condition 8a) compared to when only a single antecedent matches (conditions 8b and 8c).

PROCEDURE

The experimental and filler items were pseudo-randomized such that no two experimental items appeared adjacent to each other and were spread across three presentation lists in a Latin-square design. The experiment was divided into three blocks at which point participants could take a break if required. Forward and reverse orders of each list were constructed.

All items were presented in Courier New font (size 18), and displayed across up to three lines of text onscreen. Text was displayed in black on a white background. Eye movements were recorded using the EyeLink 1000 system (SR Research Ltd) at 500 Hz. Using the desktop system, the camera was located below the screen and participants placed their heads on a chin rest that was adjusted to allow a comfortable position. The distance between the eyes and the camera was 60 cm and the distance between eyes and screen 70 cm. Viewing was binocular but only the right eye was recorded. Each experimental session began with calibration of the eye-tracker on a nine-point grid. Calibration was repeated during the session if the experimenter noticed that measurement accuracy was poor. Before each trial, the screen displayed a marker positioned above the first word of the next trial. Participants were instructed to fixate upon this marker, and press a button to view the next trial, in order to control the placement of the initial fixations.

Participants read each text silently at their normal reading rate, pressing a button on a game pad once completed and after content questions requiring a yes/no push button response. The experiment session lasted approximately 30–45 min in total for L1 speakers. For the L2 participants the experiment took about 60 min because of the additional OPT, questionnaire (Experiment 1) and vocabulary test after the experiment. The vocabulary test consisted of a checklist containing all critical vocabulary items, and the learners were asked to read through the list carefully and circle any words that they were unfamiliar with.

The research was approved by the Ethics Committee of the University of Essex (L1, March 2011) and the ethics committee of the University of Potsdam (L2, application number 37/2011). Informed consent was obtained from all participants.

DATA ANALYSIS

Reading times for four regions of text are reported: the pronoun region, which contains the pronoun and the last three letters of the preceding word; the spillover region, which contains the two words following the pronoun [e.g., *a new* in (8a–c) above]; the next two words as the prefinal region [e.g., *song on* in (8a–c) above]; and the last two words of the sentence as the final region. For the statistical analysis, all reading time measures were log-transformed [$\log_e(x+1)$].

Five reading time measures will be reported for these regions. First fixation is the duration of readers' initial fixation within an interest area; first-pass reading time is the summed duration of fixations within an interest area until it is exited to either the left or the right for the first time; regression path time is the sum of all fixations on a region until this region is exited to the right; rereading time is the summed duration of all fixations in a region after it was first exited to either the left or right; and total viewing time is the summed duration of all fixations within a region. Reading times for trials in which track loss occurred, and reading times in regions which were initially skipped, were treated as missing data. For rereading time, trials in which a region was not refixated after the first-pass contributed a rereading time of zero to the calculation of averages.

Short fixations of 80 ms or below within one degree of visual arc of another fixation were automatically merged, and any other extremely short ($\leq$80 ms) or long ($>$1200 ms) fixations were removed. To explore whether the two participants groups patterned differently statistically, we carried out preliminary 3 × 2 ANOVAs with the factors Condition (*double match, local mismatch, non-local mismatch*) as within-subjects factor and Group (*L1, L2*) as a between-subjects factor, for each measure and interest region. Where interactions with the factor Group were found, the data from each group were analyzed separately[8].

RESULTS, EXPERIMENT 2

L1 participants answered 88% of the end-of-trial comprehension questions correctly and the L2 participants 86% overall, indicating that both groups paid attention to the task and read the stimulus items for meaning. Track loss accounted for 0.2% of the L1 and 0.13% of the L2 data. Skipping rates for the four reported

[8]Trials for which (L2) participants had indicated unknown vocabulary were not removed from the analysis reported here. A parallel analysis with unknown vocabulary trials excluded showed that excluding these did not affect the results.

regions were 25, 13, 11, and 6% in the L1 group and 9, 2, 4, and 0% in the L2 group.

Summaries of participants' reading times and of the ANOVA results are provided in **Tables 1**, **2** respectively. Results of subsequent pairwise comparisons are summarized in **Table 3**.

First-fixation durations, first-pass times and regression-path times in the region prior to the pronoun were also examined in order to check whether any effects of condition began before the pronoun was encountered. This precritical region consisted of the word before the pronoun (excluding the final three letters, which forms part of the pronoun region), and the previous word which was always an auxiliary verb. Skipping rates in this region were 11% for the L1 group and 2% for the L2 group. No effects of Condition, or Condition by Group interactions, were found in first-pass times or regression-path times. First-fixation durations did show a main effect of Condition (marginal in the F_2 analysis): [$F_{1(2,\ 132)} = 3.89$, $p < 0.05$; $F_{2(2,\ 46)} = 2.47$, $p = 0.09$.] Pairwise comparisons revealed that first-fixation durations were significantly longer in the local mismatch condition (8b) compared to the double match condition (8a) [$t_{1(67)} = 2.79$, $p < 0.05$; $t_{2(23)} = 2.26$, $p < 0.05$] and (marginally) longer than the non-local mismatch condition (8c) [$t_{1(67)} = 1.85$, $p = 0.07$; $t_{2(23)} = 2.12$, $p < 0.05$]. This effect is very fleeting, and is in a different direction from the effects seen at and beyond the pronoun region. It will therefore not be discussed any further.

PRONOUN REGION

Significant or partially significant main effects of Group were seen in all eye-movement measures, reflecting the fact that the L2 participants read the stimulus sentences generally more slowly than the L1 group. No main effects of, or interactions with, the factor Condition were found for first fixation durations or first-pass reading times. For both participant groups, regression path, rereading and total viewing times were longest in the non-local mismatch condition (8c), where the pronoun mismatched the accessible antecedent's gender, however. Significant main effects of Condition, unmodulated by the factor Group, were found for rereading and total viewing times. Subsequent *t*-tests on the collapsed L1 and L2 data confirmed that the pronoun region was reread significantly more slowly in the non-local mismatch condition (8c) compared to both the local mismatch (8b) and the double match condition (8a). The same statistical pattern was found for total viewing times.

SPILLOVER REGION

A similar pattern was seen at the spillover region. Main effects of Group were present in all measures other than rereading time. Both groups again showed the longest reading times in the non-local mismatch condition in regression path, rereading and total viewing times, giving rise to significant main effects of Condition unmodulated by the factor Group. Subsequent pairwise comparisons confirmed that in all three of these measures, the non-local mismatch condition elicited significantly longer reading times than the double match and local mismatch conditions.

The L2 group differed from the native readers in that the above reading-time pattern was also seen, numerically, in the L2 readers' first fixation durations and first-pass times at the spillover region. A Group by Condition interaction was found for first fixation durations that was significant by subjects only. To further explore this interaction, separate one-way ANOVAs for each group (L1 and L2) were carried out. These showed a significant effect of Condition for the L2 [$F_{1(2,\ 66)} = 3.81$, $p < 0.05$; $F_{2(2,\ 46)} = 5.02$,

Table 1 | Means (standard deviations in parentheses) for five eye-movement measures at four areas of interest in Experiment 2, for each participant group.

	L1					L2				
	First fixation duration	First-pass time	Regression-path time	Rereading time	Total viewing time	First fixation duration	First-pass time	Regression-path time	Rereading time	Total viewing time
PRONOUN REGION										
Double match	212 (71)	257 (132)	347 (323)	152 (233)	409 (279)	234 (72)	355 (250)	426 (364)	135 (260)	490 (388)
Local mismatch	215 (74)	267 (152)	329 (307)	115 (211)	382 (270)	234 (85)	346 (190)	415 (266)	135 (232)	481 (317)
Non-local mismatch	209 (70)	267 (154)	370 (431)	317 (463)	584 (477)	238 (85)	366 (233)	480 (453)	247 (377)	613 (420)
SPILLOVER REGION										
Double match	207 (76)	256 (162)	368 (402)	146 (230)	403 (261)	241 (96)	377 (220)	439 (368)	167 (336)	544 (418)
Local mismatch	211 (71)	260 (155)	328 (311)	142 (220)	402 (270)	233 (77)	366 (183)	420 (421)	136 (284)	502 (357)
Non-local mismatch	205 (60)	260 (148)	419 (460)	250 (350)	510 (386)	254 (103)	405 (228)	589 (595)	275 (421)	680 (508)
PREFINAL REGION										
Double match	210 (77)	297 (164)	381 (273)	171 (253)	468 (297)	239 (88)	422 (216)	530 (401)	238 (406)	661 (464)
Local mismatch	203 (67)	295 (171)	423 (396)	163 (225)	458 (282)	240 (88)	439 (227)	485 (298)	169 (276)	608 (355)
Non-local mismatch	205 (78)	288 (162)	503 (659)	216 (314)	504 (349)	249 (95)	466 (257)	646 (629)	267 (391)	733 (457)
FINAL REGION										
Double match	215 (85)	311 (202)	941 (1393)	130 (259)	441 (347)	245 (94)	520 (322)	1066 (1143)	181 (363)	701 (486)
Local mismatch	224 (106)	322 (198)	846 (1054)	105 (209)	427 (299)	246 (107)	508 (326)	1014 (1194)	151 (294)	659 (421)
Non-local mismatch	212 (91)	307 (198)	1199 (1685)	124 (229)	431 (314)	251 (104)	526 (342)	1236 (1386)	240 (426)	767 (552)

Table 2 | Summary of analyses of variance for the pronoun, spillover, prefinal and final regions in Experiment 2.

	Pronoun region		Spillover region		Prefinal region		Final region	
	*F*1	*F*2	*F*1	*F*2	*F*1	*F*2	*F*1	*F*2
FIRST FIXATION DURATION								
Group	(1, 66) 10.70**	(1, 23) 38.02***	(1, 66) 21.02***	(1, 23) 84.66***	(1, 66) 20.67***	(1, 23) 47.45***	(1, 66) 16.16***	(1, 23) 70.68***
Condition	(2, 132) 0.12	(2, 46) 0.18	(2, 132) 0.95	(2, 46) 1.62	(2, 132) 0.61	(2, 46) 0.49	(2, 132) 0.15	(2, 46) 0.23
Group × condition	(2, 132) 0.60	(2, 46) 0.55	(2, 132) 3.40*	(2, 46) 2.15	(2, 132) 1.77	(2, 46) 1.13	(2, 132) 1.72	(2, 46) 1.45
FIRST PASS TIME								
Group	(1, 66) 28.95***	(1, 23) 61.04***	(1, 66) 59.59***	(1, 23) 259.90***	(1, 66) 56.62***	(1, 23) 127.10***	(1, 66) 50.08***	(1, 23) 194.90***
Condition	(2, 132) 0.06	(2, 46) 0.34	(2, 132) 1.09	(2, 46) 1.54	(2, 132) 0.33	(2, 46) 0.57	(2, 132) 0.18	(2, 46) 0.15
Group × condition	(2, 132) 1.34	(2, 46) 0.96	(2, 132) 1.12	(2, 46) 0.57	(2, 132) 2.21	(2, 46) 1.58	(2, 132) 1.14	(2, 46) 0.83
REGRESSION-PATH TIME								
Group	(1, 66) 19.29***	(1, 23) 53.77***	(1, 66) 26.21***	(1, 23) 91.55***	(1, 66) 27.98***	(1, 23) 65.59***	(1, 66) 5.38*	(1, 23) 35.49***
Condition	(2, 132) 1.27	(2, 46) 1.99	(2, 132) 10.52***	(2, 46) 14.28***	(2, 132) 4.59*	(2, 46) 8.54***	(2, 132) 9.92***	(2, 46) 9.59***
Group × condition	(2, 132) 0.84	(2, 46) 0.72	(2, 132) 0.84	(2, 46) 0.41	(2, 132) 1.44	(2, 46) 1.79	(2, 132) 0.06	(2, 46) 0.13
REREADING TIME								
Group	(1, 66) 1.91	(1, 23) 8.14**	(1, 66) 0.93	(1, 23) 1.89	(1, 66) 0.21	(1, 23) 1.40	(1, 66) 0.21	(1, 23) 0.90
Condition	(2, 132) 21.73***	(2, 46) 23.68***	(2, 132) 17.38***	(2, 46) 19.16***	(2, 132) 5.45**	(2, 46) 2.35	(2, 132) 4.52*	(2, 46) 2.36
Group × condition	(2, 132) 1.62	(2, 46) 2.99(*)	(2, 132) 0.03	(2, 46) 0.30	(2, 132) 1.14	(2, 46) 0.52	(2, 132) 0.70	(2, 46) 0.87
TOTAL VIEWING TIME								
Group	(1, 66) 8.33**	(1, 23) 24.88***	(1, 66) 14.00***	(1, 23) 50.99***	(1, 66) 26.45***	(1, 23) 69.77***	(1, 66) 31.11***	(1, 23) 179.10***
Condition	(2, 132) 26.81***	(2, 46) 30.38***	(2, 132) 19.94***	(2, 46) 23.47***	(2, 132) 6.74**	(2, 46) 6.83**	(2, 132) 2.28	(2, 46) 1.72
Group × condition	(2, 132) 0.55	(2, 46) 1.60	(2, 132) 0.32	(2, 46) 0.31	(2, 132) 1.70	(2, 46) 1.72	(2, 132) 2.20	(2, 46) 1.99

*(*)$p < 0.1$; *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.*

Table 3 | Planned pairwise comparisons between conditions for Experiment 2.

	Pronoun region		Spillover region		Prefinal region		Final region	
	t1	*t2*	*t1*	*t2*	*t1*	*t2*	*t1*	*t2*
REGRESSION-PATH TIME								
Double match (a), local mismatch (b)			(67) 0.93	(23) 0.84	(67) 0.57	(23) 0.71	(67) 1.04	(23) 0.97
Double match (a), non-local mismatch (c)			(67) −3.23**	(23) −4.43***	(67) −2.27*	(23) −4.56***	(67) −3.09**	(23) −3.32**
Local mismatch (b), non-local mismatch (c)			(67) −4.48***	(23) −4.87***	(67) −2.66*	(23) −3.74**	(67) −4.34***	(23) −4.23***
REREADING TIME								
Double match (a), local mismatch (b)	(67) 1.20	(23) 0.87	(67) 0.92	(23) 0.64	(67) 0.99	(23) 0.77	(67) 0.96	(23) 0.75
Double match (a), non-local mismatch (c)	(67) −4.94***	(23) −4.68***	(67) −4.43***	(23) −4.73***	(67) −2.19*	(23) −1.76(*)	(67) −1.83(*)	(23) −1.53
Local mismatch, non-local mismatch	(67) −6.05***	(23) −6.18***	(67) −5.42***	(23) −5.73***	(67) −3.11**	(23) −2.40*	(67) −3.08**	(23) −2.03(*)
TOTAL VIEWING TIME								
Double match (a), local mismatch (b)	(67) 1.41	(23) 0.81	(67) 1.08	(23) 0.87	(67) 1.26	(23) 1.05		
Double match (a), non-local mismatch (c)	(67) −5.77***	(23) −5.42***	(67) −4.56***	(23) −5.59***	(67) −2.32*	(23) −2.67*		
Local mismatch (b), non-local mismatch (c)	(67) −6.26***	(23) −7.97***	(67) −6.02***	(23) −5.80***	(67) −3.46**	(23) −3.36**		

*(*)p < 0.1; *p < 0.05; **p < 0.01; ***p < 0.001.*

$p < 0.05$] but not for the L1 group [$F_{1(2,\ 66)} = 0.76$, $p = 0.47$; $F_{2(2,\ 46)} = 0.29$, $p = 0.75$]. In the L2 group first fixation durations were marginally longer, by items, in the non-local mismatch condition (8c) compared to the double match condition (8a) [$t_{1(33)} = 1.69, p = 0.10$; $t_{2(23)} = 2.61, p < 0.05$] and significantly longer compared to the local mismatch condition (8b) [$t_{1(33)} = 2.56, p < 0.05$; $t_{2(23)} = 2.68, p < 0.05$].

PREFINAL AND FINAL REGIONS

Main effects of Group were again seen at the prefinal and final regions, alongside main effects of Condition not modulated by Group. In the prefinal region significant condition effects were found in regression path and total viewing times, with the effect significant by subjects only in rereading times. Pairwise comparisons once again revealed significant differences between the non-local mismatch condition (8c) and both the double match (8a) and the local mismatch condition (8b) for regression path, rereading and total viewing times. In the final region there was a main effect of condition in the regression-path times (also a main effect significant by subjects in rereading times). Pairwise comparisons again revealed significant differences between the non-local mismatch condition (8c) and both the double match (8a) and the local mismatch condition (8b) for regression path times, with marginal differences in the same direction for rereading times.

SUMMARY, EXPERIMENT 2

In Experiment 2 the two participant groups patterned largely alike. Participants showed sensitivity to gender-mismatching non-local (i.e., accessible) antecedents but not to mismatching local (i.e., inaccessible) antecedents. These non-local mismatch effects were generally restricted to later reading-time measures, including total viewing times, with the exception of the L2 group's first fixation durations at the spillover region. This relatively minor between-groups difference might be due to the non-native readers' generally more "serial" reading strategy (as reflected by their lower skipping rates). Participants showed no evidence of considering the local antecedent at any point during processing, a finding that is consistent with the BAIF hypothesis.

The accessible-mismatch effects we observed are also in line with the results from the offline antecedent choice task, where both participant groups consistently chose the non-local antecedent.

The predictions of the defeasible filter hypothesis are not borne out here, because there is no evidence that either group considered the inaccessible antecedent at a later point during processing.

Note, however, that it is theoretically possible that the non-local mismatch effects seen in Experiment 2 reflect a general preference for matrix subject antecedents rather than the application of condition B. Examining the processing of SDPs should be able to confirm or rule out this hypothesis. It also allows us to see whether feature matching plays a more important role in L1 and/or L2 processing in the absence of a structural constraint which rules out one of the antecedents.

MATERIALS AND METHODS, EXPERIMENT 3

Our second eye-movement experiment examined the real-time processing of pronouns believed to be exempt from condition B.

Recall that in the offline task (Experiment 1), both L1 and L2 participants showed awareness of the ambiguity of SDPs. However, in cases where one specific antecedent was chosen, L2s preferred the non-local antecedent whereas for L1s there was no preference. Online, will L1 and L2 participants show sensitivity to the gender of the local or non-local antecedent, or both antecedents?

PARTICIPANTS

These were the same as in Experiment 2.

MATERIALS

The materials for this experiment included 18 experimental items which were again composed of three sentences each, a lead-in sentence, a critical sentence that contained the pronoun and two potential antecedents, and a wrap-up sentence. The gender match between the two names and the pronoun was manipulated to create three conditions as illustrated in (9a–c).

(9) Suddenly the lights went on and there were police everywhere.

(a) *Double match condition*
Barry saw Gavin place a gun near him on the ground with great care.

(b) *Local mismatch condition*
Barry saw Megan place a gun near him on the ground with great care.

(c) *Non-local mismatch condition*
Megan saw Barry place a gun near him on the ground with great care.

The robbery was definitely over now.

The names were again matched in letter and syllable length, were either typical male or typical female names, and were counterbalanced across the items. Half the pronouns were masculine and half feminine. As in the materials for Experiment 2, the first name (the non-local antecedent) was always the matrix subject. The second name (the local antecedent) was always the subject of an infinitival complement of a perception verb. Unlike in Experiment 2, the pronoun here appeared inside a prepositional phrase and thus was not a coargument of the local antecedent.

PREDICTIONS

Since SDPs are thought to be ambiguous and exempt from condition B, the predictions for Experiment 3 differ somewhat from those for Experiment 2 above.

Matrix-subject preference

If the parser initially searches for the matrix subject (i.e., the non-local antecedent), longer reading times are expected in the non-local mismatch condition (9c) compared to the other two conditions, similar to the results from Experiment 2.

Feature-match hypothesis

Where condition B does not rule out the local antecedent, the parser may be sensitive to gender mismatches between the pronoun and either or both potential antecedents. Participants might experience interference or competition when both antecedents match the pronoun in gender (condition 9a) compared to when only a single antecedent matches (conditions 9b and 9c), which would be reflected in longer reading times for the double-match condition (9a) compared to the two mismatch conditions.

Previous research on SDPs suggests that L1s are sensitive to their ambiguity in online processing tasks (Sekerina et al., 2004). For L2s there is evidence from eye-movement experiments on reflexives which indicates that they prefer linking these to the most discourse-prominent antecedent initially (Felser and Cunnings, 2012). In the light of these findings, we may expect the L2 group to show a different processing pattern from the L1 group here. While L1s might fail to show a clear antecedent preference for SDPs, or may be slowed down by antecedent competition in condition (9a), the non-native group might try to link SDPs to the matrix subject, giving rise to non-local gender mismatch effects.

PROCEDURES

The experimental, data cleaning and data analysis procedures for Experiment 3 were the same as in Experiment 2.

RESULTS, EXPERIMENT 3

Responses to the comprehension questions are reported in the Results section for Experiment 2. As for Experiment 2, we will report statistical analyses for four sentence regions. The pronoun region contained the pronoun and the last three letters of the preceding preposition, the spillover region contained the two words (e.g., *on the*) immediately following the pronoun, the prefinal region two words (e.g., *ground with*) following the spillover region and the final region the final two words of the sentence. Skipping rates for these regions were 11, 20, 9, and 20% in the L1 group and 5, 4, 2, and 5% in the L2 group.

Table 4 provides an overview of the reading time data and **Table 5** shows the between-groups ANOVA results of the log-transformed data in Experiment 3.

As for Experiment 2, a precritical region was examined in order to check whether any effects of condition began before the pronoun was encountered. This consisted of the preposition preceding the pronoun (excluding the final three letters) and the previous one or two words forming the object of the second verb. Skipping rates in this region were 5% for the L1 group and 1% for the L2 group, No effects of Condition, or Condition by Group interactions, were found in first-fixation durations, first-pass times or regression-path times.

PRONOUN REGION

At the pronoun region the native readers showed the longest regression path, rereading and total viewing times for the local mismatch condition (9b) numerically, whereas the L2 group consistently showed the longest reading times for the non-local mismatch condition (9c). No significant main effects or interactions (other than main effects of Group in all measures except rereading times) were found at this region, however.

SPILLOVER REGION

At the two words following the pronoun, main effects of Group were once again seen in all measures except rereading times. The L2 group—but not the L1 group—again showed the longest reading times in the non-local mismatch condition (9c) in all

Table 4 | Means (standard deviations in parentheses) for five eye-movement measures at four areas of interest in Experiment 3, for each participant group.

	L1					L2				
	First fixation duration	First-pass time	Regression-path time	Rereading time	Total viewing time	First fixation duration	First-pass time	Regression-path time	Rereading time	Total viewing time
PRONOUN REGION										
Double match	217 (93)	283 (164)	364 (332)	134 (223)	417 (293)	246 (85)	379 (199)	472 (490)	138 (285)	517 (353)
Local mismatch	213 (73)	287 (182)	413 (546)	154 (225)	441 (281)	244 (96)	381 (188)	442 (346)	147 (269)	528 (337)
Non-local mismatch	220 (93)	292 (169)	362 (323)	128 (203)	420 (259)	250 (99)	388 (190)	508 (697)	207 (375)	595 (415)
SPILLOVER REGION										
Double match	209 (74)	259 (218)	330 (324)	118 (200)	377 (288)	227 (74)	338 (166)	407 (343)	175 (341)	512 (426)
Local mismatch	201 (63)	239 (115)	350 (298)	121 (216)	361 (233)	229 (74)	353 (208)	473 (591)	148 (273)	501 (336)
Non-local mismatch	213 (73)	251 (138)	333 (358)	124 (236)	375 (278)	254 (114)	378 (222)	498 (420)	214 (368)	591 (444)
PREFINAL REGION										
Double match	194 (64)	293 (174)	423 (426)	123 (187)	416 (242)	225 (77)	439 (288)	542 (508)	191 (371)	630 (465)
Local mismatch	201 (80)	292 (171)	465 (616)	171 (265)	464 (304)	219 (61)	432 (253)	539 (488)	186 (341)	617 (454)
Non-local mismatch	204 (79)	296 (194)	478 (543)	146 (246)	442 (319)	215 (73)	392 (228)	539 (530)	211 (322)	603 (386)
FINAL REGION										
Double match	201 (96)	242 (136)	522 (629)	46 (112)	288 (168)	245 (113)	400 (249)	790 (1141)	88 (209)	487 (354)
Local mismatch	213 (89)	257 (158)	649 (1083)	77 (161)	334 (235)	233 (88)	380 (243)	751 (920)	85 (214)	465 (333)
Non-local mismatch	208 (102)	247 (140)	604 (840)	47 (112)	294 (196)	239 (127)	397 (285)	861 (1367)	132 (377)	529 (488)

five eye-movement measures numerically. The initial omnibus ANOVA revealed a main effect of Condition in first fixation durations, as well as significant Group by Condition interaction in regression path times in the analysis by subjects. Marginal interactions, by subjects only, were also found for rereading and total viewing times. As the observed (marginal) interactions, in the presence of significant main effects of Group, are indicative of between-group differences, we went on to analyze each group's reading-time data for the spillover region separately. Whilst the L1 group showed no significant effects at this region, the L2 group showed a significant main effect of Condition for first fixation durations [$F_{1(2,\ 66)} = 4.82$, $p < 0.05$; $F_{2(2,\ 34)} = 5.41$, $p < 0.01$] and significant effects, in the analyses by subjects, for regression path [$F_{1(2,\ 66)} = 5.46$, $p < 0.01$; $F_{2(2,\ 34)} = 2.97$, $p = 0.06$] and total viewing times [$F_{1(2,\ 66)} = 5.67$, $p < 0.01$; $F_{2(2,34)} = 3.22$, $p = 0.05$]. Planned pairwise comparisons showed that the non-local mismatch condition (9c) was read significantly more slowly than both the double match (9a) [$t_{1(33)} = 3.08$, $p < 0.01$; $t_{2(17)} = 2.73$, $p < 0.05$] and the local mismatch (9b) conditions [$t_{1(33)} = 2.36$, $p < 0.05$; $t_{2(17)} = 2.57$, $p < 0.05$] in first fixation durations, significantly more slowly (by subjects) than the double match condition in regression path [$t_{1(33)} = 3.34$, $p < 0.01$; $t_{2(17)} = 2.06$, $p = 0.05$] and total viewing times [$t_{1(33)} = 3.10$, $p < 0.01$; $t_{2(17)} = 1.95$, $p = 0.07$], and significantly more slowly than the local mismatch condition in total viewing times [$t_{1(33)} = 2.87$, $p < 0.01$; $t_{2(17)} = 2.13$, $p < 0.05$].

PREFINAL AND FINAL REGIONS

No significant effects or interactions, other than main effects of Group, were found at the prefinal region. At the final sentence region, interactions between Condition and Group were observed for both rereading times (marginal by items) and total viewing times. Here the L1 group showed the longest reading times for the local mismatch condition (9b) in these measures, whereas the L2 group again had longer reading times for the non-local mismatch (9c) than for the other two conditions. Subsequent per-group analyses only yielded a marginally significant main effect of Condition for the L1 group's total viewing times [$F_{1(2,\ 64)} = 2.98$, $p = 0.06$; $F_{2(2,\ 34)} = 2.5$, $p = 0.09$], and a marginal one in the by-items analysis for the L2 group's rereading times [$F_{1(2,\ 66)} = 2.46$, $p = 0.09$; $F_{2(2,\ 34)} = 0.85$, $p = 0.43$], however.

CORRELATION OF READING TIMES WITH OPT SCORE AND OFFLINE CHOICES

To investigate whether, for the L2 participants, the slower reading times in the non-local mismatch condition in the spillover region (9c) originate from a lack of knowledge about SDP structures among those participants with lower OPT scores, both OPT score and offline antecedent choice rates from Experiment 1 were correlated against reading times[9]. The difference between mean total viewing time in conditions (9b) and (9c) in the spillover region was calculated per participant as a measure of an individual's processing difficulty on encountering a mismatching non-local antecedent. However, there was no significant correlation between this reading measure and either OPT score [$r_{(34)} = -0.14$, $p = 0.4$] or antecedent choice rates [$r_{(34)} = 0.03$, $p = 0.8$].

SUMMARY, EXPERIMENT 3

In Experiment 3 we saw differences between the L1 and L2 groups' reading-time patterns, in particular in the spillover region. In the

[9]We thank the reviewers for this suggestion.

Table 5 | Summary of analyses of variance for the pronoun, spillover, prefinal, and final region in Experiment 3.

	Pronoun region		Spillover region		Prefinal region		Final region	
	*F*1	*F*2	*F*1	*F*2	*F*1	*F*2	*F*1	*F*2
FIRST FIXATION DURATION								
Group	(1, 66) 12.53***	(1, 17) 40.76***	(1, 66) 15.87***	(1, 17) 125.10***	(1, 66) 9.15**	(1, 17) 34.19***	(1, 65) 12.63***	(1, 17) 36.44***
Condition	(2, 132) 0.29	(2, 34) 0.49	(2, 132) 4.64*	(2, 34) 3.92*	(2, 132) 0.03	(2, 34) 0.02	(2, 130) 0.51	(2, 34) 0.82
Group × condition	(2, 132) 0.08	(2, 34) 0.02	(2, 132) 1.22	(2, 34) 0.98	(2, 132) 1.24	(2, 34) 1.28	(2, 130) 2.30	(2, 34) 2.29
FIRST PASS TIME								
Group	(1, 66) 31.01***	(1, 17) 136.50***	(1, 66) 40.99***	(1, 17) 98.59***	(1, 66) 39.68***	(1, 17) 111.00***	(1, 65) 41.26***	(1, 17) 88.55***
Condition	(2, 132) 0.32	(2, 34) 0.51	(2, 132) 1.05	(2, 34) 0.47	(2, 132) 0.43	(2, 34) 2.53(*)	(2, 130) 0.19	(2, 34) 0.54
Group × condition	(2, 132) 0.05	(2, 34) 0.17	(2, 132) 0.72	(2, 34) 0.63	(2, 132) 0.36	(2, 34) 1.13	(2, 130) 1.85	(2, 34) 1.49
REGRESSION-PATH TIME								
Group	(1, 66) 20.22***	(1, 17) 68.34***	(1, 66) 21.27***	(1, 17) 44.09***	(1, 66) 18.76***	(1, 17) 48.14***	(1, 65) 8.4**	(1, 17) 50.91***
Condition	(2, 132) 0.17	(2, 34) 0.41	(2, 132) 2.54(*)	(2, 34) 1.39	(2, 132) 0.17	(2, 34) 0.03	(2, 130) 0.42	(2, 34) 0.20
Group × condition	(2, 132) 0.63	(2, 34) 0.27	(2, 132) 4.44*	(2, 34) 1.94	(2, 132) 0.06	(2, 34) 0.09	(2, 130) 0.95	(2, 34) 0.84
REREADING TIME								
Group	(1, 66) 0.36	(1, 17) 1.05	(1, 66) 0.59	(1, 17) 1.62	(1, 66) 0.18	(1, 17) 0.07	(1, 65) 1.45	(1, 17) 4.51*
Condition	(2, 132) 2.41(*)	(2, 34) 2.59(*)	(2, 132) 0.62	(2, 34) 0.40	(2, 132) 1.18	(2, 34) 1.63	(2, 130) 0.67	(2, 34) 0.34
Group × condition	(2, 132) 1.37	(2, 34) 1.83	(2, 132) 2.93(*)	(2, 34) 1.92	(2, 132) 1.55	(2, 34) 2.52(*)	(2, 130) 3.20*	(2, 34) 2.67(*)
TOTAL VIEWING TIME								
Group	(1, 66) 14.93***	(1, 17) 34.58***	(1, 66) 21.65***	(1, 17) 74.72***	(1, 66) 22.55***	(1, 17) 56.16***	(1, 65) 29.61***	(1, 17) 102.8***
Condition	(2, 132) 1.52	(2, 34) 1.71	(2, 132) 2.05	(2, 34) 0.87	(2, 132) 0.33	(2, 34) 0.33	(2, 130) 0.58	(2, 34) 0.42
Group × condition	(2, 132) 1.22	(2, 34) 1.02	(2, 132) 2.79(*)	(2, 34) 1.79	(2, 132) 0.28	(2, 34) 1.16	(2, 130) 3.64*	(2, 34) 5.04*

*(*)$p < 0.1$; *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.*

pronoun region, the trend in the L1 data was for increased reading times in the local mismatch condition (9b) while the L2 trend was for increased times in the non-local mismatch condition (9c). Although these different patterns did not yield statistically reliable between-groups differences in the pronoun region, they gave rise to some interactions with the factor Group in later regions. In the spillover region the L1s showed no significant differences between the experimental conditions whilst the L2s showed increased reading times for the non-local mismatch condition (9c), indicative of trying to link the pronoun to the matrix subject. Analysis of the L1 data in the final region revealed a trend toward longer total viewing times in the local mismatch condition (9b). In the following section, the results from Experiment 3 will be discussed together with those from Experiments 1 and 2.

DISCUSSION

We set out to investigate the application and timing of condition B during L1 and L2 processing of English pronouns. Firstly, we discovered that both L1 and L2 groups were sensitive to the gender of the accessible antecedent online. There was an increase in reading times when the non-local (accessible) antecedent mismatched the pronoun's gender in canonical condition B environments. Secondly, we discovered that when both antecedents were structurally available (in SDP environments), L2s were again sensitive to the gender of the non-local antecedent (which was the matrix subject) while L1s experienced some difficulty with the local mismatch condition.

STRUCTURAL SENSITIVITY

Results from the offline questionnaire (Experiment 1) revealed that both the L1s and L2s ignored an inaccessible but gender-matching antecedent and instead chose the accessible antecedent almost exclusively, in line with condition B. This offline adherence to condition B was also reflected online in both groups, who showed longer reading times in the non-local mismatch condition in Experiment 2. This indicates a higher processing cost when the available antecedent mismatched in gender with the pronoun. No measurable processing cost was elicited by a mismatching inaccessible antecedent at any point, indicating that the inaccessible antecedent was not considered[10] . Furthermore, the results from Experiment 3 for the L1 group suggest that there may be no general preference for the first-mentioned antecedent, so it is unlikely that the Experiment 2 results were driven by such an underlying preference. These findings are line with the BAIF hypothesis, in which condition B gates access to the potential antecedents by filtering out structurally inaccessible ones. As such it adds to the evidence gained from the self-paced reading studies of Clifton et al. (1997, 1999), as well as self-paced reading and eye-tracking evidence from Chow et al. (in preparation). Because of the sensitivity of the eye-movement monitoring technique used in the current experiments, the evidence here suggests that previous support for the BAIF is not simply due to a less sensitive time measure which failed to pick up on short-lived, early effects.

The L1 data from Experiment 3 showed a trend for late processing difficulty in the local-mismatch condition, although this did not prove statistically reliable. This might nevertheless suggest that, while the native readers were largely unaffected by our manipulations of gender congruence between the pronoun and the potential antecedents, they had a weak preference for a local antecedent online. No such preference was visible in the L1 group's offline data, however. In the SDP environments both of the antecedents were accessible, and all experimental conditions contained at least one gender-matching accessible antecedent. This may explain the relative lack of any condition-specific processing difficulty in comparison to the condition B environments. The fact that the SDP items were processed differently despite being presented in same experimental session as the condition B items highlights that the L1 parser was sensitive to the subtle syntactic cues which distinguish SDP environments from those in which condition B applies.

TIMING

With respect to timing, it should first be noted that the L2 group showed sensitivity to our experimental manipulation in an earlier measure than did the L1 group in Experiment 2 (first fixation durations at the spillover region). In fact, the timing of the non-local mismatch effect in this experiment for the L1 group appears to be fairly late, appearing only in rereading times. The emergence of the L1 effect in rereading times could be due to a rapid reading strategy leading to fewer fixations and longer saccades, but increased regressive eye-movements in case of difficulty. In contrast, the L2s read more slowly, spending more time in each region. These differences in reading style might explain the seemingly earlier effects in the L2 group compared to the L1 group.

The timing of the effect in L1s, however, still stands in contrast to findings for inaccessible mismatch effects in previous (L1) studies with reflexives (e.g., Sturt, 2003). The comparison with reflexive studies is speculative because reflexives were not systematically tested in the current study. However some further consideration should be given to timing, since the study employs a method that is particularly sensitive to timecourse. It cannot be assumed that early and late reading measures are necessarily linked to distinct cognitive processes (see Pickering et al., 2004 for a discussion). As such, the effects in the rereading times could be behavioral echoes of much earlier processes. Even so, a later effect for pronouns fits in well with two considerations: first, pronouns are sensitive to a range of cues or information types which can help to determine their reference, so considering all these information sources may require more time; second, the nature of condition B, unlike condition A for reflexives, involves excluding rather than identifying an antecedent, and may require the generation of more than one semantic sentence representation (Reuland, 2001, 2011) or the consideration of pragmatic information (Huang, 1994).

[10]However, it should be noted that a previous analysis of the Experiment 2 data, in which the pronoun region contained only the pronoun itself, the L2 group did appear to be briefly distracted by a gender-matching, inaccessible antecedent. Following a reviewer's suggestion, this analysis was replaced due to high skipping rates and the resultant loss of data.

L1 vs. L2 PROCESSING

The L2 group showed a very similar pattern of results to the L1 group in Experiment 2, but a different pattern of results from the L1 group in Experiment 3. Although the results of Experiment 2 suggest that L2s do rule out the inaccessible antecedent in accordance with condition B (like the L1 group), results from Experiment 3 for the L2 group call this into question. In Experiment 3, the L2 participants were again sensitive to the gender of the non-local antecedent, despite their awareness of the ambiguity in the offline task (Experiment 1). This means that their sensitivity to the non-local antecedent in Experiment 2 may not be a result of applying condition B, but could instead be a general preference to link the pronoun to the matrix subject, even though offline the L2s show awareness of the ambiguity of the SDPs. This suggests firstly that L2s are less sensitive than L1s to the subtle syntactic cues that differentiate the SDP environments from the canonical condition B environments. Secondly, they appear to have a general preference for salient subjects, which may have driven the non-local mismatch effect for L2s in both Experiments 2 and 3. The discrepancy between L2s' offline knowledge and their use of this knowledge during online processing has been observed in previous studies, as well as a preference for (discourse-) salient antecedents (Felser and Cunnings, 2012 for reflexives). This finding is consistent with the hypothesis that L2 speakers tend to underuse structural information during processing and rely more on other cues such as discourse-level information instead (Clahsen and Felser, 2006).

A reviewer raises the question of whether the German participants' preference for non-local antecedents in Experiment 3 might reflect L1 transfer. Similar SDP configurations to those tested here also exist in German. To find out which, if any, antecedent native German readers might prefer online, we carried out a parallel eye-movement study on German (as yet unpublished). While L1 German readers showed an offline preference for the non-local antecedent, their reading-time patterns look similar to those of the native English group in the current study in that they did not show any measurable preference for either the local or non-local antecedent. The double-match condition tended to be the shortest one instead, a pattern that proved statistically significant only for total viewing times at the spillover region, however. This makes it unlikely that our Experiment 3 results reflect L1 transfer from German[11].

IMPLICATIONS FOR ANTECEDENT SEARCH MECHANISMS

The predictions of the BAIF hypothesis for pronouns appear to be very similar to those of a structured search mechanism for reflexives (Dillon, 2011; Dillon et al., 2013). If readers show sensitivity to the conditions governing both reflexives and pronouns, can they be assumed to exploit the same search mechanism? This makes the assumption that condition B is purely a structural constraint, a proposal which is contested by several theoretical accounts. A purely structured search to eliminate an inaccessible antecedent may therefore be inadequate. Nevertheless, a model of memory search for pronouns must incorporate (i) the ability to exclude an inaccessible antecedent from consideration even when it carries features that match the pronoun, and (ii) awareness of explicitly structural cues that distinguish, for example, canonical condition B environments from SDP environments. It is clear that native speakers make use of this information during processing, and that it plays a decisive role during the consideration of potential antecedents.

A slightly different question is whether there is a strict ordering of constraint application, as Nicol and Swinney imply in their original formulation of their hypothesis:

> "…the reactivation of prior referents is restricted by grammatical constraints. In the case where such information does not sufficiently constrain the list of potential antecedents to a single one, the pragmatic and other sentence/discourse processing procedures undoubtedly come into play, but, given the present evidence, only at a later point in processing."
>
> (Nicol and Swinney, 1989, p.18)

While the lack of interference from an inaccessible antecedent seems to imply that binding conditions are applied before other cues such as gender features are recruited, there is as yet no firm evidence that discourse cues, for example, are systematically withheld relative to binding constraints in the time-course of pronoun resolution. Given that discourse cues are increasingly found to act early and even predictively (e.g., Koornneef and Van Berkum, 2006; Cozjin et al., 2011), further research on the interaction between condition B and the discourse status of antecedents would be welcome, to confirm or disconfirm a strict ordering of constraint application.

In addition, any model of the retrieval process should be able to incorporate the profiles of both native and non-native comprehenders. As far as the L2 processing is concerned, the current study shows that the processing of pronouns may be driven by a search for a salient subject, rather than making use of a detailed structural analysis to distinguish condition B and SDP environments; this is not the case for L1 processing. This demonstrates a different sensitivity to structural cues in the two populations; generalizing a retrieval or processing model so that it applies equally well to L1 and L2 pronoun resolution could perhaps be achieved by assigning differing constraint weights in different populations.

CONCLUSION

Native English speakers appear to successfully apply condition B online so that they do not consider an inaccessible antecedent at any point during processing, which is in line with the BAIF hypothesis. They are also sensitive to syntactic cues that distinguish syntactic environments that either require, or do not require, the exclusion of a local referent. By contrast, non-native speakers do not appear to distinguish condition B environments from SDP environments online, appearing to opt for salient subject antecedents in both despite offline awareness of the difference. The different processing profiles of native and non-native speakers must be incorporated into models of retrieval, with particular reference to the relative importance of structural cues for different populations.

[11]Note that in order to draw any meaningful conclusions about possible L1 transfer, learner groups from different L1 backgrounds would need to be compared.

ACKNOWLEDGMENTS

Part of this research was completed as part of Clare Patterson's PhD dissertation at the University of Potsdam. She was supported by an ESRC postgraduate studentship awarded by the Department of Language and Linguistics at the University of Essex, and by an Alexander-von-Humboldt-Professorship to Prof. Harald Clahsen (Potsdam Research Institute for Multilingualism), which is gratefully acknowledged.

REFERENCES

Agresti, A. (2002). *Categorical Data Analysis, 2nd Edn.* New York, NY: John Wiley & Sons. doi: 10.1002/0471249688

Allan, D. (2004). *Oxford Placement Test*. Oxford: Oxford University Press.

Badecker, W., and Straub, K. (2002). The processing role of structural constraints on the interpretation of pronouns and anaphors. *J. Exp. Psychol. Learn. Mem. Cogn.* 28, 748–769. doi: 10.1037/0278-7393.28.4.748

Bertenshaw, N. (2009). *The Application of Binding Constraints by Japanese L2 Learners of English*. Ph.D. dissertation, University of Essex.

Chien, Y.-C., and Wexler, K. (1990). Children's knowledge of locality constraints in binding as evidence for the modularity of syntax and pragmatics. *Lang. Acquis.* 1, 225–295. doi: 10.1207/s15327817la0103_2

Chomsky, N. (1981). *Lectures in Government and Binding.* Dordrecht: Foris.

Clackson, K., Felser, C., and Clahsen, H. (2011). Children's processing of reflexives and pronouns in English: evidence from eye-movements during listening. *J. Mem. Lang.* 65, 128–144. doi: 10.1016/j.jml.2011.04.007

Clahsen, H., and Felser, C. (2006). How native-like is non-native language processing? *Trends Cogn. Sci.* 10, 564–570. doi: 10.1016/j.tics.2006.10.002

Clifton, C. Jr., Frazier, L., and Deevy, P. (1999). Feature manipulation in sentence comprehension. *Riv. Linguist.* 11, 11–39.

Clifton, C. Jr., Kennison, S. M., and Albrecht, J. E. (1997). Reading the words *her*, *his*, *him*: implications for parsing principles based on frequency and structure. *J. Mem. Lang.* 36, 276–292. doi: 10.1006/jmla.1996.2499

Cook, V. J. (1990). Timed comprehension of binding in advanced L2 learners of English. *Lang. Learn.* 40, 557–599. doi: 10.1111/j.1467-1770.1990.tb00606.x

Cozjin, R., Commandeur, E., Vonk, W. and Noordman, L. G. M. (2011). The time course of the use of implicit causality information in the processing of pronouns: a visual world paradigm study. *J. Mem. Lang.* 64, 381–403. doi: 10.1016/j.jml.2011.01.001

Cunnings, I., Patterson, C., and Felser, C. (2014). Variable binding and coreference in sentence comprehension: evidence from eye movements. *J. Mem. Lang.* 71, 39–56. doi: 10.1016/j.jml.2013.10.001

Dillon, B. W. (2011). *Structured Access in Sentence Comprehension*. Ph.D. dissertation, University of Maryland.

Dillon, B. W., Mishler, A., Sloggett, S., and Phillips, C. (2013). Contrasting intrusion profiles for agreement and anaphora: experimental and modelling evidence. *J. Mem. Lang.* 69, 85–103. doi: 10.1016/j.jml.2013.04.003

Ellert, M. (2010). *Ambiguous Pronoun Resolution in L1 and L2 German and Dutch*. MPI series in psycholinguistics, Vol. 58. Wageningen: Ponsen & Looijen.

Felser, C., and Cunnings, I. (2012). Processing reflexives in English as a second language: the role of structural and discourse-level constraints. *Appl. Psycholinguist.* 33, 571–603. doi: 10.1017/S0142716411000488

Felser, C., Sato, M., and Bertenshaw, N. (2009). The on-line application of binding Condition A in English as a second language. *Biling. Lang. Cogn.* 12, 485–502. doi: 10.1017/S1366728909990228

Gordon, P. C., Hendrick, R., and Johnson, M. (2001). Memory interference during language processing. *J. Exp. Psychol. Learn. Mem. Cogn.* 27, 1411–1423. doi: 10.1037//0278-7393.27.6.1411

Hestvik, A. (1991). Subjectless binding domains. *Nat. Lang. Linguist. Theory* 9, 455–496. doi: 10.1007/BF00135355

Huang, Y. (1994). *The Syntax and Pragmatics of Anaphora: A Study with Special Reference to Chinese.* Cambridge, UK: Cambridge University Press. doi: 10.1017/CBO9780511554292

Kaiser, E., Runner, J. T., Sussman, R. S., and Tanenhaus, M. K. (2009). Structural and semantic constraints on the resolution of pronouns and reflexives. *Cognition* 112, 55–80. doi: 10.1016/j.cognition.2009.03.010

Koornneef, A. W., and Van Berkum, J. J. A. (2006). On the use of verb-based implicit causality in sentence comprehension: evidence from self-paced reading and eye tracking. *J. Mem. Lang.* 54, 445–465. doi: 10.1016/j.jml.2005.12.003

Lees, R. and Klima, E. (1963). Rules for English pronominalization. *Language* 39, 17–28. doi: 10.2307/410759

Lewis, R., and Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cogn. Sci.* 29, 375–419. doi: 10.1207/s15516709cog0000_25

Nicol, J., and Swinney, D. (1989). The role of structure in coreference assignment during sentence comprehension. *J. Psycholinguist. Res.* 18, 5–20. doi: 10.1007/BF01069043

Patil, U. (2012). *An Evaluation of a Cognitive Architecture of Sentence Processing: Computational and Empirical Assessment*. Doctoral thesis, University of Potsdam.

Pickering, M. J., Frisson, S., McElree, B., and Traxler, M. (2004). "Eye movements and semantic composition," in *The On-Line Study of Sentence Comprehension: Eyetracking, ERPs and Beyond*, eds M. Carreiras and C. Clifton Jr. (New York, NY: Psychology Press), 33–50.

Reinhart, T. (1983). *Anaphora and Semantic Interpretation*. London: Croom Helm.

Reinhart, T., and Reuland, E. (1993). Reflexivity. *Linguist. Inq.* 24, 657–720.

Reuland, E. (2001). Primitives of binding. *Linguist. Inq.* 32, 439–492. doi: 10.1162/002438901750372522

Reuland, E. (2011). *Anaphora and Language Design*. Cambridge, MA: The MIT Press.

Roberts, L., Gullberg, M., and Indefrey, P. (2008). Online pronoun resolution in L2 discourse: L1 influence and general learner effects. *Stud. Second Lang. Acquis.* 30, 333–357. doi: 10.1017/S0272263108080480

Rooryck, J., and Vanden Wyngaerd, G. (2011). *Dissolving Binding Theory*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780199691326.001.0001

Runner, J., Sussman, R., and Tanenhaus, M. (2003). Assignment of reference to reflexives and pronouns in picture noun phrases: evidence from eye-movements. *Cognition* 89, B1–B13. doi: 10.1016/S0010-0277(03)00065-9

Sekerina, I., Stromswold, K., and Hestvik, A. (2004). How adults and children process referentially ambiguous pronouns. *J. Child Lang.* 31, 123–152. doi: 10.1017/S0305000903005890

Sturt, P. (2003). The time-course of the application of binding constraints in reference resolution. *J. Mem. Lang.* 48, 542–562. doi: 10.1016/S0749-596X(02)00536-3

Tenny, C. (2004). *Pronoun Binding and the Grammar Of Sentience. Workshop on Semantic Approaches to Binding Theory, ESSLLI 16*. Université Henri Poincaré, Nancy.

Van Dyke, J. A. (2007). Interference effects from grammatically unavailable constituents during sentence processing. *J. Exp. Psychol. Learn. Mem. Cogn.* 33, 407–430. doi: 10.1037/0278-7393.33.2.407

White, L. (1998). Second language acquisition and binding principle B: child/adult differences. *Second Lang. Res.* 14, 425–439. doi: 10.1191/026765898675690231

Xiang, M., Dillon, B., and Phillips, C. (2009). Illusory licensing effects across dependency types: ERP evidence. *Brain Lang.* 108, 40–55. doi: 10.1016/j.bandl.2008.10.002

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 15 October 2013; accepted: 05 February 2014; published online: 25 February 2014.

*Citation: Patterson C, Trompelt H and Felser C (2014) The online application of binding condition B in native and non-native pronoun resolution. Front. Psychol. **5**:147. doi: 10.3389/fpsyg.2014.00147*

This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology.

Copyright © 2014 Patterson, Trompelt and Felser. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 23 June 2015
doi: 10.3389/fpsyg.2015.00840

Structural constraints on pronoun binding and coreference: evidence from eye movements during reading

Ian Cunnings[1]*, *Clare Patterson*[2] *and Claudia Felser*[2]

[1] School of Psychology and Clinical Language Sciences, University of Reading, Reading, UK, [2] Potsdam Research Institute for Multilingualism, University of Potsdam, Potsdam, Germany

A number of recent studies have investigated how syntactic and non-syntactic constraints combine to cue memory retrieval during anaphora resolution. In this paper we investigate how syntactic constraints and gender congruence interact to guide memory retrieval during the resolution of subject pronouns. Subject pronouns are always technically ambiguous, and the application of syntactic constraints on their interpretation depends on properties of the antecedent that is to be retrieved. While pronouns can freely corefer with non-quantified referential antecedents, linking a pronoun to a quantified antecedent is only possible in certain syntactic configurations via variable binding. We report the results from a judgment task and three online reading comprehension experiments investigating pronoun resolution with quantified and non-quantified antecedents. Results from both the judgment task and participants' eye movements during reading indicate that comprehenders freely allow pronouns to corefer with non-quantified antecedents, but that retrieval of quantified antecedents is restricted to specific syntactic environments. We interpret our findings as indicating that syntactic constraints constitute highly weighted cues to memory retrieval during anaphora resolution.

Keywords: pronoun resolution, memory retrieval, quantification, eye movements, reading, English

OPEN ACCESS

Edited by:
Charles Jr. Clifton, University of Massachusetts Amherst, USA

Reviewed by:
Dave Kush, Haskins Laboratories, USA
Roger P. G. Van Gompel, University of Dundee, UK

***Correspondence:**
Ian Cunnings, School of Psychology and Clinical Language Sciences, University of Reading, Earley Gate, Reading RG6 7BE, UK
i.cunnings@reading.ac.uk

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: *16 February 2015*
Accepted: *03 June 2015*
Published: *23 June 2015*

Citation:
Cunnings I, Patterson C and Felser C (2015) Structural constraints on pronoun binding and coreference: evidence from eye movements during reading. Front. Psychol. 6:840. doi: 10.3389/fpsyg.2015.00840

Introduction

The successful interpretation of anaphoric elements during language comprehension involves forming dependencies between constituents that may span several words or sentences. Anaphora resolution thus provides a key test case for studying the memory system that subserves language comprehension, as the correct interpretation of anaphoric constituents crucially relies on the retrieval of a particular item, the antecedent, from memory. A growing number of studies have investigated how syntactic and non-syntactic factors combine to cue the retrieval of an antecedent during the resolution of different types of anaphora (Badecker and Straub, 2002; Sturt, 2003; Xiang et al., 2009; Clackson et al., 2011; Cunnings and Felser, 2013; Dillon et al., 2013; Chow et al., 2014; Clackson and Heyer, 2014; Cunnings and Sturt, 2014; Patterson et al., 2014). Most previous research has investigated constraints on the resolution of reflexives and object pronouns, where syntactic constraints (e.g., binding conditions A and B; Chomsky, 1981) restrict memory retrieval to an antecedent in a particular syntactic domain. Research on real-time pronoun resolution has investigated the extent to which such syntactic constraints interact with other sources of information, such as discourse prominence and gender/number congruence, to guide the retrieval

of a particular antecedent. While some have claimed that syntactic constraints act as "hard constraints" that restrict memory retrieval to syntactically licit antecedents (e.g., Dillon et al., 2013; Chow et al., 2014), others argue that syntactic constraints are violable and interact with other sources of information to cue antecedent retrieval (e.g., Badecker and Straub, 2002).

While binding conditions A and B have been well studied, to date little research has investigated the time-course of the application of syntactic constraints on the interpretation of pronouns linked to quantified and non-quantified antecedents as in (1) and (2) respectively.

(1a) The man heard that every boy at school said that he was happy.
(1b) The man who every boy at school heard said that he was happy.
(2a) The man heard that the boy at school said that he was happy.
(2b) The man who the boy at school heard said that he was happy.

In (1a), the subject pronoun *he* can refer to either the matrix subject *the man* or the quantified antecedent *every boy*. In (1b) however, when *every boy* appears inside a relative clause, it is not possible for the pronoun to be bound by it. Note that this is not an absolute restriction on antecedents inside relative clauses however, as the pronoun can freely refer to the non-quantified antecedent *the boy* in both (2a) and (2b). This contrast between quantified and non-quantified antecedents thus provides a particular challenge for memory retrieval mechanisms during language processing. For example, if retrieval operations disfavored antecedents inside relative clauses this would ensure that syntactically illicit quantified antecedents, as in (1b), are not retrieved, but would also rule out perfectly licit non-quantified antecedents as in (2b). Conversely, if subject pronouns routinely trigger retrieval of antecedents inside relative clauses, syntactically illicit quantified antecedents may be retrieved. Instead, successful pronoun interpretation requires selective retrieval of antecedents inside relative clauses, but this is dependent on the properties of the to-be-retrieved material.

The aim of the current study was to investigate the retrieval of quantified and non-quantified antecedents during pronoun resolution to further examine how syntactic constraints on memory retrieval are implemented during real-time anaphora resolution. To this end, we conducted an offline judgment task and three online reading experiments investigating pronoun resolution with quantified and non-quantified antecedents. We begin below by discussing theoretical accounts of the contrast between quantified and non-quantified antecedents as exemplified in (1) and (2), before discussing implications of this contrast for models of memory retrieval during language processing in more detail.

Background

Variable Binding and Coreference Assignment

In linguistic theory, it has been claimed that pronoun resolution can be achieved in different ways. Although theoretical accounts differ in their precise nature, a core idea is that pronouns can be resolved either in the discourse representation, via coreference assignment, or in logical syntax, via variable binding (e.g., Evans, 1980; Bosch, 1983; Reinhart, 1983; Reuland, 2001, 2011). Coreference assignment involves linking a pronoun to a referential antecedent in the discourse, as in the case of linking the pronoun *he* to either of the antecedents (*the man* or *the boy*) in (2). Quantified phrases (QPs), as in (1), however do not refer to a single individual in the discourse, and a pronoun linked to a QP co-varies in interpretation with the quantifier. Pronouns linked to QPs are thus said to involve variable binding rather than coreference assignment.

A long-standing observation in the linguistics literature is that variable binding is only possible in certain syntactic configurations. This restriction has traditionally been characterized in terms of c-command. C-command refers to a relationship between constituents in the phrase structure representation of a sentence based on the notion of hierarchical dominance. In the standard definition, a constituent c-commands its sister constituents and any constituents that these dominate (Reinhart, 1983). Variable binding is only possible between a pronoun and an antecedent that c-commands it (see e.g., Reuland, 2001, 2011). As such, in (3a), when the QP *every boy* c-commands the pronoun, the pronoun can be bound by it, while in (3b), when the QP does not c-command the pronoun, variable binding between the pronoun and QP is not possible. Coreference assignment to non-quantified determiner phrases (DPs) is not contingent on c-command however, and as such the pronoun can corefer with the referential antecedent *the boy* in both (2a) and (2b).

Memory Retrieval during Language Processing

Recent psycholinguistic research has motivated a cue-based model of memory retrieval during language processing (McElree, 2000; McElree et al., 2003; Lewis et al., 2006). In cue-based content-addressable models, retrieval is achieved by matching a set of retrieval cues with the contents of all items in memory in parallel. The item in memory that provides the best match to these cues becomes most highly activated and will thus be retrieved. The distinction between variable binding and coreference assignment has a number of implications for models of memory retrieval during language processing.

One theoretical implication relates to how the c-command constraint on variable binding is implemented in content-addressable memory. Content-addressable models are well-suited to utilize feature-based cues that target intrinsic properties of to-be-retrieved material. For example, it is straightforward to implement a [+masculine] feature for masculine pronouns to cue retrieval of a masculine antecedent. However, the c-command constraint on variable binding may be more difficult to implement, as it involves access to information about the *relation* between two items in memory (the pronoun and antecedent), rather than accessing an intrinsic *feature* of the antecedent (for discussion, see Kush, 2013; Kush et al., 2015). The primary aim of the current study was to investigate *if* the c-command constraint on variable binding restricts antecedent retrieval, rather than the question of *how* it is implemented. We do however return to this issue in the General Discussion.

A second implication that the distinction between variable binding and coreference assignment has for models of memory retrieval during language processing relates to how pronouns that are ambiguous with regards to variable binding and coreference assignment are resolved. Research in theoretical linguistics has claimed that syntactic variable binding is preferred over coreference assignment. Reuland (2001, 2011), for example, proposed an economy principle which predicts that variable binding should be computed before coreference assignment is attempted (see also Koornneef, 2008). This predicts that variable binding antecedents should preferentially be retrieved before coreference antecedents. Cunnings et al. (2014) tested this prediction in two reading experiments. They manipulated gender congruence between a pronoun and two potential antecedents in the discourse, and monitored participants' eye-movements as they read sentences as in (3).

(3a) Every soldier who knew that James/Helen was watching was convinced that he/she should wave as the parade passed.
(3b) It looked to James/Helen that every soldier was completely convinced that he/she should wave as the parade passed.

In their Experiment 1, exemplified in (3a), gender congruence was manipulated between the pronoun and a c-commanding QP (*every soldier*), and between the pronoun and a non c-commanding but linearly closer proper name coreference antecedent (*James/Helen*). They hypothesized that if variable binding is computed before coreference assignment, when participants encounter the pronoun, the c-commanding QP antecedent should be preferentially retrieved. In this case, the gender of the c-commanding QP should affect reading times at a point in time before the gender of the proper name. However, in contrast to this prediction, they observed that reading times at and shortly after the pronoun were longer when the pronoun mismatched in gender with the proper name antecedent, and were not significantly affected by the gender of the QP. This suggests that the proper name antecedent, rather than the QP, was preferentially retrieved upon encountering the pronoun. In their Experiment 2 however, exemplified in (3b), when the QP was linearly closer to the pronoun than the proper name antecedent, reading times at and shortly after the pronoun were reliably longer when the QP mismatched in gender with the pronoun. Together, these results indicate there is no overall preference for either variable binding or coreference assignment. For variable binding antecedents to be retrieved additional factors, such as antecedent recency, need to favor the QP antecedent.

A third issue relates to how the c-command constraint on variable binding, however it is implemented, interacts with other cues to antecedent retrieval. Cunnings et al. only investigated cases in which variable binding antecedents were syntactically licit, and did not test sentences containing QPs that did not c-command the critical pronoun. A key prediction of cue-based models is similarity-based interference (see e.g., Lewis et al., 2006; Van Dyke and Johns, 2012). As retrieval involves the matching of a set of retrieval cues with all items in memory in parallel, a distractor item that partially matches the retrieval cues may sometimes be retrieved instead of the intended retrieval target. This leads to the possibility that a QP antecedent that does not c-command a pronoun may occasionally be retrieved even though this dependency is ungrammatical.

Attraction effects in subject-verb agreement are a key example of such interference effects during language processing. For example, Wagers et al. (2009) reported longer reading times for sentences containing ungrammatical compared to grammatical subject-verb agreement (e.g., *the key to the cabinet was rusty* vs. *the key to the cabinet were rusty*). This ungrammaticality effect was however reliably attenuated when the structurally illicit distractor matched the agreement marking of the critical verb (e.g., *the key to the cabinets were rusty*). This attraction effect provides good evidence that structural cues (e.g., [+phrasal head]) and agreement (e.g., [+plural]) are equally weighted cues that combine to guide retrieval during subject-verb agreement. When no item in memory fully matches the cues at retrieval, a partially matching distractor can sometimes be retrieved. We will refer to this pattern of results as *facilitatory* interference, as the processing of ungrammatical sentences is facilitated by a partially matching distractor.

Although facilitatory attraction effects are well attested for subject-verb agreement, a number of studies have failed to observe this specific pattern of interference during anaphora resolution (e.g., Sturt, 2003; Xiang et al., 2009; Dillon et al., 2013; Chow et al., 2014; Cunnings and Sturt, 2014). Sturt, for example, manipulated gender congruence between a reflexive and two antecedents in a piece of discourse (e.g., *Jonathan/Jennifer remembered that the surgeon had pricked himself/herself with a used syringe needle*) and observed that while first-pass reading times at the reflexive were reliably longer when the structurally licit antecedent *the surgeon* mismatched in stereotypical gender with the reflexive, the gender of the structurally illicit antecedent (*Jonathan/Jennifer*) did not affect reading times in this measure. Results such as these have led some to claim that while equally weighted syntactic and agreement cues combine to guide retrieval for subject-verb agreement, anaphora resolution is guided by syntactic "hard constraints" that restrict retrieval to syntactically licit antecedents (Dillon et al., 2013; Chow et al., 2014). Although the question of whether or not structurally illicit antecedents are always ignored during anaphora resolution is debated (e.g., Badecker and Straub, 2002; Cunnings and Felser, 2013; Clackson and Heyer, 2014), the contrast in attraction effects observed for agreement and anaphora suggests these dependencies implement agreement cues in different ways. For anaphora, syntactic constraints appear to be more strongly weighted cues to retrieval than gender/number congruence.

Syntactic constraints on reflexives could potentially be implemented as highly weighted cues that trigger retrieval of an antecedent within a particular syntactic domain (e.g., the same clause as the reflexive; see Dillon et al., 2013). However, constraints on quantified and non-quantified antecedents are difficult to implement in this way, as it is not the case that antecedents within a particular syntactic domain (e.g., a relative clause) are categorically ruled out. Rather, sensitivity to constraints on variable binding and coreference assignment require retrieval operations to be able to selectively retrieve antecedents that do not c-command pronouns depending on their quantificational status. The contrast between variable

binding and coreference assignment thus provides a unique challenge to memory retrieval operations during language processing, which may leave variable binding more susceptible to facilitatory interference than has been observed for other types of anaphora, such as reflexives.

We are aware of only one study that has investigated the potential for facilitatory interference during the resolution of bound variable anaphora. Kush et al. (2015) recorded participants' eye-movements as they read sentences as in (4).

(4a) The troop leaders that the boy/girl scout had no respect for had scolded her after the incident at scout camp.
(4b) The troop leaders that no boy/girl scout had respect for had scolded her after the incident at scout camp.
(4c) The troop leaders were sure no boy/girl scout was afraid that she would be scolded after the incident at scout camp.

In (4a), the only syntactically licit antecedent for the pronoun *her* is the coreference antecedent *the boy/girl*. In (4b), the pronoun has no syntactically licit antecedent as the quantified phrase (*no boy/girl*) does not c-command it. Kush et al. hypothesized that if the pronoun triggers retrieval of the coreference antecedent in (4a), a gender mismatch effect should be observed, with longer reading times for gender mismatching (*the boy*) than gender matching (*the girl*) antecedents. If antecedent retrieval respects the c-command constraint, this contrast between gender matching (*no girl scout*) and gender mismatching (*no boy scout*) antecedents should not be observed in (4b). If the c-command constraint does not restrict antecedent retrieval however, Kush et al. hypothesized that the gender mismatch effect should be observed in both (4a) and (4b), as evidence of facilitatory interference. During first-pass processing at the pronoun Kush et al. observed a gender mismatch effect in (4a) but not (4b), suggesting the c-command constraint on variable binding restricts the early stages of antecedent retrieval. They did observe gender mismatch effects in (4c) however, when the quantified phrase c-commanded the pronoun. Kush et al interpreted these results as indicating that pronouns trigger retrieval of both c-commanding quantified phrases and non c-commanding coreference antecedents, but not non c-commanding quantified antecedents, suggesting that the c-command constraint restricts antecedent retrieval.

Against this background, the aim of the current study was to further investigate the implementation of the c-command constraint on variable binding during anaphora resolution. While Kush et al. compared antecedent retrieval for c-commanding and non c-commanding quantified antecedents in different sentence structures with different (subject and object) pronouns, we investigated variable binding and coreference resolution in maximally similar sentences with identical (subject) pronouns across four experiments. We also tested the universal quantifier *every* rather than the negative quantifier *no*. Together with the study reported by Kush et al., the current experiments provide a systematic examination of how constraints on retrieving quantified phrases and referential antecedents during anaphora resolution are implemented during language processing. Experiment 1 was an offline task that tested the extent to which naïve participants are sensitive to the c-command constraint on variable binding in an untimed task. Experiments 2–4 were online reading studies in which participants' eye-movements were monitored. Experiments 2–3 contrasted the retrieval of quantified and non-quantified referential antecedents in order to test the extent to which variable binding and coreference antecedents are retrieved in c-commanding and non c-commanding configurations. Experiment 4 tested the extent to which the c-command restriction on variable binding acts as a "hard constraint" on antecedent retrieval.

Experiment 1

Experiment 1 used a sentence judgment paradigm to assess sensitivity to the c-command constraint on variable binding in an untimed offline task. The materials consisted of sentences as in (5), which manipulate the factor "c-command" to test whether participants are willing to link pronouns to QPs in different syntactic configurations.

(5a) *C-commanding QP*
The surgeon suggested that every man on the waiting list definitely realized that he needed some help.
(5b) *Non c-commanding QP*
The surgeon who every man on the waiting list suggested definitely realized that he needed some help.

In (5a), the QP *every man* c-commands the pronoun *he* and as such the pronoun can be bound by the QP via variable binding. In (5b) however, the QP appears inside a relative clause and as such does not c-command the pronoun. In this case, the pronoun can only refer to the matrix subject *the surgeon*. We expect native English speakers to be sensitive to the c-command constraint on variable binding in this offline task. That is, participants should consider the QP as a possible antecedent for the pronoun in (5a) but not (5b).

Methods

Participants

32 native English speakers (17 males, mean age 21; range 18–30) from the University of Edinburgh community either received course credit or a small payment for taking part in Experiment 1[1]. All participants in Experiment 1, and Experiments 2–4, provided written, informed consent before the experiment began. Ethical approval for all experiments was granted by the Department of Psychology Research Ethics Committee at the University of Edinburgh.

Materials

Materials consisted of 16 experimental items constructed as in (5). In each item, the pronoun matched in definitional gender with the QP antecedent. The pronoun also always matched in stereotypical gender with the matrix subject to ensure that the texts were felicitous. The materials manipulated the factor "c-command" in two conditions, such that the QP either c-commanded or did not c-command the QP. A full list of experimental items is provided in Appendix A. In addition to the

[1]The participants in Experiment 1 also completed Experiment 3. All participants completed Experiment 3 before Experiment 1.

experimental items, 24 filler items were also constructed, some of which also contained pronouns but others which did not.

Procedure

The experimental and filler items were presented to participants as a questionnaire in Microsoft Word. A question appeared under each text with two possible answers. For the experimental items, the question always probed the interpretation of the pronoun. In (4), for example, the question was "Who does 'he' refer to?" with "(A) the surgeon" and "(B) every man" as possible answers. Participants provided a response by selecting one of five options from a drop-down menu that appeared beside each text. Possible responses were "(A) strongly preferred", "(A) mildly preferred", "(A) or (B) equally likely", "(B) mildly preferred" or "(B) strongly preferred". Across the 16 experimental items, the matrix subject and QP antecedents each appeared as options "(A)" and "(B)" an equal number of times. Fillers that did not include pronouns consisted of complex (ambiguous and unambiguous) sentences containing elliptical gaps. Two paraphrases, (A) and (B), were provided as answers which participants had to choose between using the same scale as in the experimental items.

The experimental and filler items were pseudo-randomized such that no two experimental items appeared next to each other. Items were spread across two presentation lists in a Latin-square design. Forward and reverse orders of each list were presented to the same number of participants. Participants were instructed to simply read each sentence and provide an answer to the questions using the drop-down menu.

Results

Responses were coded from −2 to 2, with −2 meaning "QP strongly preferred" and 2 meaning "DP strongly preferred." A score of 0 indicated either antecedent was equally likely, while −1 and 1 indicated a mild preference for the QP and DP respectively. The average rating in the c-commanding QP condition was −0.16 (SD 1.62) and in the non c-commanding condition 1.38 (SD 1.14). A pairwise comparison indicated that scores were significantly higher in the non c-commanding QP condition than the c-commanding QP condition [$t_{1(31)} = 8.19$, $p < 0.001$; $t_{2(15)} = 11.80$, $p < 0.001$]. This indicates that the DP antecedent was chosen more often when the QP did not c-command the pronoun compared to when it did. One sample *t*-tests indicated that the average scores in the c-command condition did not differ significantly from 0 [$t_{1(31)} = 0.90$, $p = 0.374$; $t_{2(15)} = 1.23$, $p = 0.237$], but that the scores in the non c-command condition were significantly higher than 0 [$t_{1(31)} = 11.12$, $p < 0.001$; $t_{2(15)} = 16.46$, $p < 0.001$]. This indicates that when the QP c-commanded the pronoun, participants considered either antecedent equally likely, but that the DP was preferred when the QP did not c-command the pronoun.

Discussion

The results of Experiment 1 align with intuitions from the theoretical linguistics literature. When the QP c-commanded the pronoun, participants were equally likely to interpret the pronoun as referring to either the QP or the DP antecedent. When the QP did not c-command the pronoun, participants preferred to interpret the pronoun as being coreferential with the DP. Experiment 1 thus suggests that naïve participants are sensitive to the c-command restriction on variable binding[2]. Experiment 2 tested how this constraint is implemented during online sentence processing.

Experiment 2

The aim of Experiment 2 was to investigate the application of the c-command constraint on variable binding during real-time language processing. Participants read a series of texts as in (6) while their eye-movements were monitored. The gender-mismatch paradigm (Sturt, 2003; Kazanina et al., 2007) was used as a diagnostic of dependency formation.

(6a) *C-commanding QP, gender match*
Being in hospital can be quite difficult at times.
The surgeon saw that every old man on the emergency ward silently wished that he could go a little bit faster.

(6b) *C-commanding QP, gender mismatch*
Being in hospital can be quite difficult at times.
The surgeon saw that every old woman on the emergency ward silently wished that he could go a little bit faster.

(6c) *Non c-commanding QP, gender match*
Being in hospital can be quite difficult at times.
The surgeon who every old man on the emergency ward saw silently wished that he could go a little bit faster.

(6d) *Non c-commanding QP, gender mismatch*
Being in hospital can be quite difficult at times.
The surgeon who every old woman on the emergency ward saw silently wished that he could go a little bit faster.

In (6a,b) the pronoun *he* is c-commanded by the QP *every old (wo)man*. In (6c,d) the pronoun is not c-commanded by the QP, as it appears inside a relative clause. In (6a,c) the QP *every old man* matches the gender of the pronoun, while in (6b,d) the QP *every old woman* does not. If participants attempt to retrieve the c-commanding QP upon encountering the pronoun, we expect to observe a gender mismatch effect such that reading times at or shortly after the pronoun should be longer in gender mismatch condition (6b) than gender match condition (6a). If the c-command constraint restricts antecedent retrieval during processing (Kush et al., 2015), no gender mismatch effect should be observed when the QP appears inside a relative clause, as in (6c,d). If however participants violate the c-command constraint during processing, we can expect to see gender mismatch effects in both (6a,b) and (6c,d). Sensitivity to the c-command constraint

[2]A reviewer notes that the results of Experiment 1 on their own could equally be explained in terms of a dispreference for linking pronouns to antecedents inside relative clauses, irrespective of quantification, rather than a specific constraint on variable binding to QPs. While this is a possible explanation of the results in Experiment 1, the results of Experiments 2 and 3 suggest that the restriction on binding to QPs is best characterised in terms of the c-command constraint, rather than a general dispreference against antecedents inside relative clauses.

is thus diagnosed statistically by an interaction between the main effects of c-command and gender, while main effects of gender would indicate constraint violation.

Methods

Participants

Thirty two native English speakers (8 males, mean age 19; range 17–23) from the University of Edinburgh community with normal or corrected-to-normal vision, and who did not take part in any of the other experiments reported here, took part in Experiment 2.

Materials

Twenty four experimental items as in (6) were constructed. A full list can be found in Appendix B. Each item began with a short context sentence that took up one line onscreen. The critical second sentence appeared across two lines, with the line-break always appearing before the adverb [*silently* in (6)] that appeared before the verb preceding the critical pronoun. The matrix subject of the critical sentence always matched the pronoun in stereotypical gender to ensure that a felicitous interpretation of the pronoun was always possible. The critical gender manipulation between the QP and pronoun always involved definitional gender (e.g., *every old man/ woman*).

In addition to the experiment items, 60 filler texts were also constructed that included a variety of different constructions, some of which included different types of anaphors. The fillers took up between two and three lines of text onscreen.

Procedures

Experimental and filler items were pseudo-randomized such that no two experimental items appeared adjacent to each other and were spread across four presentation lists in a Latin-square design. A different random order of items was presented to each participant. The experiment began with five practice items to familiarize participants with the procedure. All items were presented in Consolas fixed width font and displayed across up to three lines of text onscreen.

Eye movements were recorded using the EYELINK 2000 system, sampling at a rate of 1000 Hz. While viewing was binocular, eye movements were recorded from the right eye only. Each experimental session began with calibration of the eye-tracker on a nine-point grid, and any drift in calibration was compensated for via recalibration between trials if required. Before each trial, participants fixated on a fixation marker above the first word of the trial to be displayed. Upon fixation on this marker, the trial text appeared. Participants read each text silently at their normal reading rate, pressing a button on a control pad once completed. To ensure participants paid attention to the content of the sentences, comprehension questions requiring a yes/no push button response followed two thirds of all trials. The entire experiment lasted approximately 30–45 min in total.

Reading times are reported for four regions of text. The critical *pronoun region* consisted of the subject pronoun and the preceding complementiser (*that he*). We extended the pronoun region to the left of the critical pronoun rather than the right to avoid effects of first-pass processing at the pronoun being mixed with spillover effects at the post-pronoun region. As the perceptual span in English is approximately eight characters to the right of fixation (Rayner, 1998), fixations on the complementiser are likely to involve foveal processing of the pronoun. The *spillover region* comprised the two words after the pronoun (*could go*) while the *prefinal region* consisted of the next two words (*a little*). The *final region* consisted of the rest of the critical sentence (*bit faster*).

Four reading time measures are reported for each region of text. *First pass reading time* is the summed duration of fixations within a region during its first inspection, until it is exited to the left or right, while *regression path duration* is calculated by summing the duration of each fixation, starting with the first fixation when a region is entered from the left, up until but not including the first fixation in a region to the right. In addition to these two *first-pass* processing measures, we also calculated *second pass times*, which included all fixations within a region *after* it has been exited following the first-pass. *Total viewing times*, which sum all fixations in a region, are reported as a global measure of processing load. All trials in which track loss occurred were discarded, and regions which were initially skipped during reading were treated as missing data in the two first-pass measures. For second pass times, trials in which a region was not refixated after the first-pass contributed a second pass time of zero to the calculation of averages. Prior to the calculation of reading time measures an automatic procedure merged short fixations of 80 ms or below that were within one degree of visual arc of another fixation. All other fixations of 80 ms or below, as well as those above 800 ms, were removed. Outliers that were above or below 3.5 standard deviations from a participant's mean reading time for each measure were also removed before analysis.

Analysis was conducted using linear-mixed effects models with crossed random effects for subjects and items (Baayen, 2008; Baayen et al., 2008). For each reading time measure, the analysis included deviation-coded fixed main effects of "c-command" (c-command vs. non c-command), "gender" (match vs. mismatch) and their interaction. Subject and item random intercepts, as well as subject and item random slopes for each fixed effect, were included using a "maximal" random effects structure (Barr et al., 2013). If this maximal model failed to converge, the random effects structure was simplified by removing the random correlation parameters, which for the analyses reported here always led to convergence. For fixed effects, p-values were estimated from the t distribution (Baayen, 2008, p. 248). In the case of reliable interactions, planned comparisons compared gender mismatch effects separately for the two c-command and two non c-command conditions.

Results

Overall accuracy to the comprehension questions was 88% (all subjects above 73%), indicating that participants paid attention to the content of the sentences. Track loss accounted for 0.1% of the data and skipping rates for the pronoun, spillover, prefinal and

final regions were 26, 5, 19, and 10% respectively[3]. A summary of the reading time data is provided in **Table 1**. **Table 2** provides a summary of the statistical analysis.

At the pronoun region, there was a significant main effect of c-command in first-pass reading times, with reading times being longer in the two non c-command conditions (6c,d) than c-commanding conditions (6a,b). This likely reflects spillover processing as a result of the extra layer of syntactic embedding from the relative clause that appears in conditions (6c,d) but not (6a,b). There were significant c-command by gender interactions in both second-pass and total viewing times. Planned comparisons in both measures indicated that when the QP c-commanded the pronoun, reading times were longer in gender mismatch condition (6b) than gender match condition (6a) (for second-pass times, estimate = 71, SD = 30, $t = 2.407$, $p = 0.017$; for total viewing times, estimate = 80, SD = 31, $t = 2.586$, $p = 0.010$). The same comparisons in the two non c-command conditions were not significant (for both measures, $t < 1$, $p > 0.651$). This pattern of results, with gender mismatch effects in the c-command conditions only, is illustrated for second pass times in **Figure 1**. These results indicate that readers attempted to link the pronoun to the QP when it c-commanded the pronoun but not when it did not.

[3]Skipping rates at the pronoun region were quite high in Experiments 2–4. We thus conducted an additional analysis in which the two first-pass measures at the pronoun were calculated using a leftward-shifting procedure (see Sturt, 2003, p. 548). In this analysis, if the pronoun was initially skipped during reading, fixations up to four characters to the left of the region boundary were included in the calculation of first-pass and regression path times. This reduced skipping rates at the pronoun to below 8% across experiments, but did not alter the overall pattern of results compared to the non-shifted analysis reported in the main text.

At the spillover region, there were significant main effects of gender in both second-pass and total viewing times that were modulated by significant c-command by gender interactions in both measures. Again, planned comparisons in the c-command conditions indicated significantly longer reading times for gender mismatch condition (6b) than gender match condition (6a) (for second-pass times, estimate = 111, SD = 33, $t = 3.403$, $p < 0.001$; for total viewing times, estimate = 134, SD = 35, $t = 3.920$, $p < 0.001$), but there were no significant differences between the two non c-command conditions (for both measures, $t < 1$, $p > 0.563$). These results further indicate that readers retrieved the QP upon encountering the pronoun, but only when the QP c-commanded it.

At the prefinal and final regions there were marginally significant c-command by gender interactions in the regression path times. At the prefinal region, regression path durations were again numerically larger following a gender mismatch in the c-command conditions only, but here neither of the planned comparisons was significant (both $t < 1.2$, both $p > 0.236$). At the final region, regression path durations were marginally longer following gender mismatches in the two c-command conditions (estimate = 405, SD = 225, $t = 1.801$, $p = 0.073$). The same comparison for the two non c-command conditions was not significant ($t < 1$. $p > 0.470$).

Discussion

The results of Experiment 2 clearly show that readers readily retrieved the QP upon encountering the pronoun, but only when it was a syntactically licit antecedent. At both the critical pronoun and spillover regions, second-pass and total viewing times were longer when the QP mismatched in gender with the pronoun,

TABLE 1 | Reading times in milliseconds for four eye-movement measures at four regions of texts in Experiment 2 (SDs in parentheses).

	First pass reading time	Regression path time	Second pass time	Total viewing time
PRONOUN REGION				
C-commanding QP, gender match	226 (91)	260 (158)	105 (229)	328 (278)
C-commanding QP, gender mismatch	235 (126)	302 (278)	176 (271)	417 (297)
Non c-commanding QP, gender match	256 (131)	303 (230)	142 (240)	409 (287)
Non c-commanding QP, gender mismatch	259 (155)	330 (284)	137 (213)	390 (275)
SPILLOVER REGION				
C-commanding QP, gender match	262 (120)	303 (177)	170 (239)	419 (253)
C-commanding QP, gender mismatch	285 (166)	370 (263)	280 (319)	561 (331)
Non c-commanding QP, gender match	275 (170)	345 (354)	213 (284)	497 (356)
Non c-commanding QP, gender mismatch	274 (140)	339 (223)	205 (258)	478 (291)
PREFINAL REGION				
C-commanding QP, gender match	255 (155)	424 (638)	180 (230)	406 (263)
C-commanding QP, gender mismatch	266 (159)	531 (902)	188 (255)	440 (286)
Non c-commanding QP, gender match	275 (146)	488 (798)	177 (241)	439 (269)
Non c-commanding QP, gender mismatch	267 (123)	419 (394)	183 (247)	435 (270)
FINAL REGION				
C-commanding QP, gender match	287 (160)	1840 (2070)	139 (252)	437 (278)
C-commanding QP, gender mismatch	297 (174)	2288 (2234)	157 (2334)	493 (365)
Non c-commanding QP, gender match	295 (169)	2221 (2592)	138 (2592)	452 (316)
Non c-commanding QP, gender mismatch	290 (178)	2023 (2143)	146 (2143)	460 (349)

TABLE 2 | Summary of statistical analyses for four eye-movement measures at four regions of texts in Experiment 2.

	First pass reading time		Regression path time		Second pass time		Total viewing time	
	Estimate	*t*	Estimate	*t*	Estimate	*t*	Estimate	*t*
PRONOUN REGION								
C-command	25 (11)	2.392*	35 (23)	1.529	2 (17)	0.118	25 (19)	1.298
Gender	3 (12)	0.280	32 (25)	1.272	34 (22)	1.527	35 (27)	1.301
C-command* gender	1 (24)	0.035	14 (48)	0.288	74 (37)	2.012*	90 (38)	2.357*
SPILLOVER REGION								
C-command	2 (12)	0.135	8 (22)	0.359	16 (21)	0.795	1 (23)	0.035
Gender	11 (13)	0.807	28 (22)	1.230	52 (24)	2.162*	58 (27)	2.189*
C-command* gender	22 (24)	0.925	75 (47)	1.612	117 (41)	2.852*	161 (50)	3.216*
PREFINAL REGION								
C-command	12 (13)	0.911	14 (53)	0.269	4 (17)	0.263	9 (22)	0.419
Gender	2 (12)	0.152	17 (53)	0.313	7 (24)	0.275	18 (26)	0.685
C-command* gender	24 (22)	1.105	203 (118)	1.719(*)	1 (33)	0.027	41 (37)	1.095
FINAL REGION								
C-command	3 (12)	0.245	54 (172)	0.313	7 (24)	0.289	1 (23)	0.053
Gender	1 (13)	0.096	124 (163)	0.760	13 (25)	0.525	25 (29)	0.882
C-command* gender	16 (24)	0.671	546 (312)	1.751(*)	9 (38)	0.222	35 (50)	0.699

*Estimate = Model Estimate (SE in brackets). (*) = $p < 0.10$, * = $p < 0.05$, ** = $p < 0.001$.*

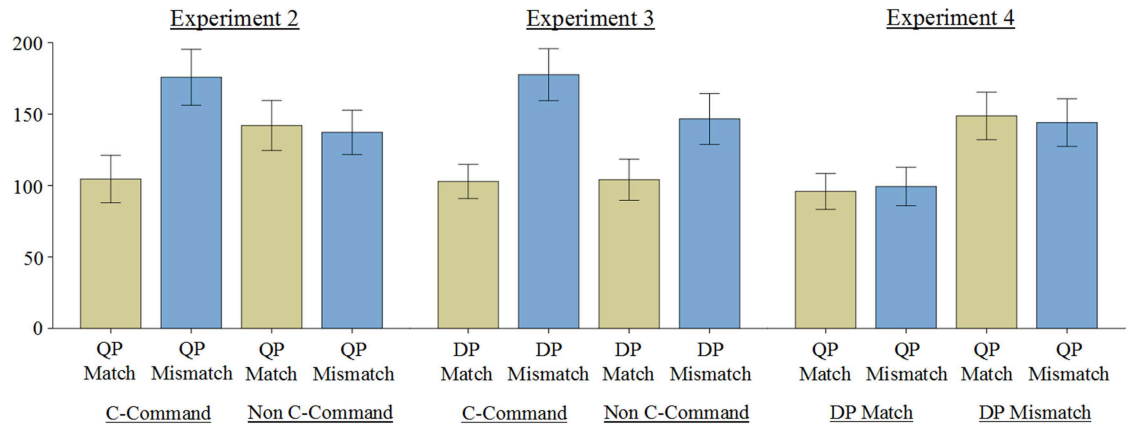


FIGURE 1 | Second pass times in milliseconds (with standard errors) at the pronoun region in Experiments 2–4.

but only when the QP c-commanded the pronoun. Similar trends were also observed in regression path times at later regions of text. At no point in time did we observe any reliable effect of the gender of the QP on participants' reading times when it did not c-command the pronoun. These results indicate that the c-command constraint on variable binding restricts antecedent retrieval during the resolution of subject pronouns.

One potential counterargument to this interpretation of our results is that the QPs inside relative clauses may have been ignored during retrieval not because of the c-command constraint on variable binding, but rather because antecedents inside relative clauses are comparatively non-discourse prominent. The results of Cunnings et al. (2014) however provide evidence against this interpretation. In their Experiment 1, they observed that readers would readily retrieve a non-quantified coreference antecedent inside a relative clause. This suggests that it is not the case that all antecedents inside relative clauses are ignored during retrieval, but rather they are readily retrieved only when syntactically licit.

However, it remains at least possible that there may have been subtle pragmatic differences between the texts used in Experiment 2 reported here and those used by Cunnings et al. (2014), which may have favored retrieval of the relative clause antecedent in Cunnings et al.'s study but not here. Note also that the coreference antecedent in Cunnings et al. was a proper name, which are known to be particularly discourse prominent (Sanford and Garrod, 1988). The aim of Experiment 3 was to investigate whether the selective retrieval profile observed in the

current experiment is truly a result of the c-command constraint on variable binding or results from differences in discourse prominence between c-commanding and non c-commanding antecedents in general.

Experiment 3

The aim of Experiment 3 was to investigate whether the c-command relationship between antecedent and pronoun affects the possibility of retrieval for non-quantified referential antecedents. The experimental materials used were identical to those from Experiment 2, except that the critical QP was replaced with a non-quantified referential DP as in (7).

(7a) *C-commanding DP, gender match*
Being in hospital can be quite difficult at times.
The surgeon saw that the old man on the emergency ward silently wished that he could go a little bit faster.

(7b) *C-commanding DP, gender mismatch*
Being in hospital can be quite difficult at times.
The surgeon saw that the old woman on the emergency ward silently wished that he could go a little bit faster.

(7c) *Non c-commanding DP, gender match*
Being in hospital can be quite difficult at times.
The surgeon who the old man on the emergency ward saw silently wished that he could go a little bit faster.

(7d) *Non c-commanding DP, gender mismatch*
Being in hospital can be quite difficult at times.
The surgeon who the old woman on the emergency ward saw silently wished that he could go a little bit faster.

As for QPs in Experiment 2, the DP c-commands the pronoun in (7a,b) but does not in (7c,d). In conditions (7a,c) the pronoun matches in gender with the DP, while in (7b,d) there is a gender mismatch. While variable binding between the pronoun and QP was syntactically illicit in conditions (6c,d) in Experiment 2, there is no constraint that restricts linking the pronoun to the DP in (7c,d) via coreference assignment. As such, if the results of Experiment 2 reflect application of the c-command constraint on variable binding, we expect to find different results with coreference antecedents in Experiment 3. That is, in contrast to the interactions observed in Experiment 2, in Experiment 3 main effects of gender should be observed such that reading times should be longer in gender mismatch conditions (7b,d) than gender match conditions (7a,c), irrespective of c-command.

However, if antecedents inside relative clauses are simply ignored during retrieval as they are not discourse prominent, we expect to observe similar results in Experiment 3 as were observed in Experiment 2. That is, we should observe reliable c-command by gender interactions, with gender mismatch effects being observed in c-command conditions (7a,b) but not non c-commanding conditions (7c,d).

Methods

Participants

32 native English speakers (17 males, mean age 21; range 18–30) from the University of Edinburgh community, none of whom took part in any of the other eye-tracking experiments reported here, took part in Experiment 3. All had normal or corrected-to-normal vision.

Materials

The 24 sets of experimental items from Experiment 2 were adapted as in (7). Experimental items were again interspersed with 60 fillers and pseudo-randomly distributed across four presentation lists in a Latin-square design.

Procedures

The procedure and data analysis were the same as outlined for Experiment 2.

Results

Average comprehension question accuracy was 90% (all subjects over 77%). There was no track loss and skipping rates for the pronoun, spillover, prefinal, and final regions were 32, 9, 16, and 12% respectively. Summaries of the reading time data and statistical analysis are provided in **Tables 3**, **4**.

At the pronoun region, there were significant main effects of gender in second pass and total viewing times, with reading times being longer in gender mismatch conditions (7b,d) compared to gender match conditions (7a,c). In contrast to Experiment 2, there was no hint of an interaction between c-command and gender in any measure at the pronoun region. This suggests that the DP was retrieved irrespective of whether or not it was inside a relative clause. This pattern of results for the second pass times at the pronoun region is shown in **Figure 1**.

The results of the spillover region replicated this pattern of results. In second pass times there was a marginal main effect of gender, with reading times again tending to be longer following a gender mismatch between the pronoun and DP compared to when there was a gender match. Total viewing times displayed the same pattern of results, with the main effect of gender being fully significant in this measure.

At the prefinal region, there was a significant c-command by gender interaction in first-pass reading times. Here, in the c-command conditions reading times were numerically longer in gender mismatch condition (7b) than gender match condition (7a). The planned comparison was however not significant ($t = 1.506$, $p = 0.133$). The opposite numerical pattern was observed in the two non c-command conditions, with gender match condition (7c) having numerically longer reading times than gender mismatch condition (7d). The planned comparison was however only marginally significant (estimate = 24, SD = 13, $t = 1.851$, $p = 0.065$). It is unclear what this numerical pattern might mean, and it is not replicated in any other measure. Indeed, in the regression path times at this region there was a significant main effect of gender, with reading times following the pattern observed at the pronoun and spillover regions, with reading times being longer following gender mismatches between the DP and pronoun compared to when there was a gender match.

TABLE 3 | Reading times in milliseconds for four eye-movement measures at four regions of texts in Experiment 3 (SDs in parentheses).

	First pass reading time	Regression path time	Second pass time	Total viewing time
PRONOUN REGION				
C-commanding DP, gender match	225 (94)	284 (233)	103 (165)	317 (194)
C-commanding DP, gender mismatch	261 (145)	311 (236)	178 (251)	413 (308)
Non c-commanding DP, gender match	246 (142)	323 (278)	104 (199)	339 (248)
Non c-commanding DP, gender mismatch	261 (144)	316 (262)	147 (247)	410 (305)
SPILLOVER REGION				
C-commanding DP, gender match	284 (150)	366 (322)	169 (217)	451 (273)
C-commanding DP, gender mismatch	295 (162)	378 (276)	211 (264)	493 (298)
Non c-commanding DP, gender match	273 (154)	358 (292)	171 (259)	431 (276)
Non c-commanding DP, gender mismatch	283 (175)	383 (338)	209 (331)	510 (377)
PREFINAL REGION				
C-commanding DP, gender match	243 (113)	455 (548)	119 (186)	358 (215)
C-commanding DP, gender mismatch	264 (133)	638 (963)	153 (203)	409 (248)
Non c-commanding DP, gender match	259 (143)	420 (536)	129 (219)	384 (263)
Non c-commanding DP, gender mismatch	234 (106)	518 (746)	143 (231)	373 (260)
FINAL REGION				
C-commanding DP, gender match	300 (220)	1364 (1391)	83 (212)	400 (299)
C-commanding DP, gender mismatch	293 (177)	1748 (1824)	114 (233)	431 (317)
Non c-commanding DP, gender match	311 (207)	1722 (2160)	92 (224)	426 (352)
Non c-commanding DP, gender mismatch	308 (198)	1783 (2082)	95 (241)	423 (314)

TABLE 4 | Summary of statistical analyses for four eye-movement measures at four regions of texts in Experiment 3.

	First pass reading time		Regression path time		Second pass time		Total viewing time	
	Estimate	***t***	**Estimate**	***t***	**Estimate**	***t***	**Estimate**	***t***
PRONOUN REGION								
C-command	9 (9)	0.922	15 (31)	0.474	14 (15)	0.952	0 (21)	0.021
Gender	17 (14)	1.261	5 (27)	0.171	58 (19)	3.116*	72 (24)	2.999*
C-command* gender	20 (21)	0.953	46 (50)	0.907	31 (27)	1.159	34 (40)	0.853
SPILLOVER REGION								
C-command	10 (12)	0.783	4 (21)	0.192	1 (20)	0.042	5 (19)	0.234
Gender	11 (12)	0.942	18 (22)	0.842	39 (21)	1.852(*)	59 (21)	2.796*
C-command* gender	0 (21)	0.011	8 (48)	0.165	5 (39)	0.141	40 (42)	0.958
PREFINAL REGION								
C-command	6 (9)	0.668	78 (51)	1.540	1 (18)	0.046	6 (20)	0.295
Gender	2 (9)	0.209	151 (63)	2.396*	22 (16)	1.431	20 (17)	1.174
C-command* gender	43 (18)	2.377*	64 (126)	0.512	22 (29)	0.754	59 (37)	1.604
FINAL REGION								
C-command	15 (18)	0.830	169 (131)	1.288	5 (15)	0.313	12 (29)	0.423
Gender	7 (18)	0.380	222 (127)	1.749(*)	18 (15)	1.204	13 (21)	0.599
C-command* gender	4 (37)	0.121	266 (257)	1.034	28 (29)	0.975	22 (43)	0.515

*Estimate = Model Estimate (SE in brackets). (*) = $p < 0.10$, * = $p < 0.05$, ** = $p < 0.001$.*

There was also a marginally significant main effect of gender in the regression path times at the final region, with reading times again tending to be longer when the pronoun mismatched in gender with the DP compared to when there was a gender match.

Discussion

The results of Experiment 3 are in clear contrast to those from Experiment 2. Whereas we observed significant c-command by gender interactions at the pronoun and spillover regions in Experiment 2, in Experiment 3 we observed only significant main effects of gender at these regions. This suggests that, in contrast to Experiment 2, in Experiment 3 participants were equally likely to retrieve the DP antecedent in both the c-command and non c-command conditions. Indeed, the relative time-course of mismatch effects across both experiments is very similar (compare graphs from Experiments 2 and 3 in

Figure 1). The crucial difference between the two is that while mismatch effects were restricted to the c-command conditions in Experiment 2, they appear irrespective of c-command in Experiment 3. This provides good evidence that the results of Experiment 2 cannot be explained in terms of antecedents inside relative clauses simply being non-discourse prominent. Rather, while both antecedents that c-command a pronoun and those that do not are readily retrieved, quantified antecedents are only retrieved when variable binding is syntactically licit. It is this contrast between syntactically licit and syntactically illicit pronoun-antecedent dependencies that appears to best explain the contrast in results between Experiments 2 and 3.

Although the results of Experiments 2 and 3 indicate that the c-command constraint restricts antecedent retrieval during language processing, one issue that remains is how the c-command constraint and gender congruence combine during anaphora resolution. In Experiments 2 and 3, there was always at least one gender-matching and syntactically licit antecedent in the discourse, namely the matrix subject DP [*the surgeon* in (5) and (6)]. To fully test how the c-command constraint and gender congruence interact to guide antecedent retrieval, it is also necessary to investigate anaphora resolution when the only syntactically licit antecedent available in the discourse provides only a partial match to the cues at retrieval. Experiment 4 was thus conducted to test this issue.

Experiment 4

The aim of Experiment 4 was to investigate how the c-command constraint and gender congruence combine to guide antecedent retrieval. Materials in Experiment 4 contained the two non c-command conditions from Experiment 2, additionally manipulating the stereotypical gender relationship between pronoun and matrix subject DP as in (8).

(8a) *DP gender match, QP gender match*
Being in hospital can be quite difficult at times.
The surgeon who every old man on the emergency ward saw silently wished that he could go a little bit faster.

(8b) *DP gender match, QP gender mismatch*
Being in hospital can be quite difficult at times.
The surgeon who every old woman on the emergency ward saw silently wished that he could go a little bit faster.

(8c) *DP gender mismatch, QP gender match*
Being in hospital can be quite difficult at times.
The surgeon who every old woman on the emergency ward saw silently wished that she could go a little bit faster.

(8d) *DP gender mismatch, QP gender mismatch*
Being in hospital can be quite difficult at times.
The surgeon who every old man on the emergency ward saw silently wished that she could go a little bit faster.

In (8), the QP always appears inside a relative clause and as such is not a syntactically licit antecedent of the pronoun. In each condition, the only syntactically licit antecedent is the matrix subject DP *the surgeon*. In (8a,b), this DP matches in stereotypical gender with the pronoun, whereas in (8c,d) there is a stereotypical gender mismatch. In (8a,c) the non c-commanding QP additionally matches the gender of the pronoun, while in (8b,d) it does not.

Different predictions with regards to the time-course of antecedent retrieval can be made depending on how the c-command constraint and gender congruence combine. If syntactic constraints on anaphora resolution constitute "hard constraints" that gate retrieval to syntactically licit antecedents (Dillon et al., 2013; Chow et al., 2014; Kush et al., 2015), we should observe main effects of the gender of the DP only. Reading times should be longer in DP gender mismatch conditions (8c,d) than in DP gender match conditions (8a,b). The gender of the syntactically illicit QP should not influence reading times at any point in the sentence.

Alternatively, if the c-command constraint and gender congruence combine to guide retrieval, we expect to observe facilitatory interference (e.g., Wagers et al., 2009). In this case, we would expect reading times to generally be longer in DP gender mismatch conditions (8c,d) than DP gender match conditions (8a,b). However, the size of the gender mismatch effect should be reliably attenuated when the structurally illicit QP matches in gender with the pronoun. In this case, reading times should be shorter in condition (8c), when the QP matches the gender of the pronoun, in comparison to (8d), when neither antecedent matches. This result would indicate that when no syntactically licit antecedent is available in the discourse that matches the pronoun's gender, a gender matching but syntactically illicit antecedent may sometimes be retrieved.

Another possibility is that we may observe a difference in the time-course of effects for syntactically licit and illicit antecedents. Sturt (2003) proposed the "defeasible filter" hypothesis which predicts that initially only structurally licit antecedents are considered, but that structurally illicit antecedents can subsequently be retrieved during later stages of processing. Applying this logic to the current experiment, we may observe an initial attempt to retrieve only the syntactically licit DP, followed by subsequent effects of the syntactically illicit QP. In this case, we should observe main effects of stereotypical gender mismatch between the pronoun and DP antecedent only at or shortly after the pronoun, with any effects of the gender of the structurally illicit QP antecedent being in comparison delayed.

Methods

Participants

32 native English speakers (12 males, mean age = 24; range = 18–49) from the University of Edinburgh community with normal or corrected-to-normal vision, none of which took part in Experiments 1–3, took part in Experiment 4.

Materials

The 24 experimental items from Experiment 2 were adapted as in (8), and again pseudo-randomly interspersed with 60 fillers across four presentation lists in a Latin-square design.

The stereotypical gender manipulations included items that had previously been pre-tested to ensure they displayed the intended stereotypes (Cunnings and Felser, 2013; Cunnings et al., 2014).

Procedures

The procedure and data analysis were the same as in Experiments 2 and 3.

Results

Overall accuracy to comprehension questions was 89% (all subjects above 77%). Track loss accounted for 0.1% of the data. Skipping rates for the pronoun, spillover, prefinal and final regions were 21, 7, 14, and 7% respectively. Summaries of the reading times and statistical analyses are shown in **Tables 5**, **6**.

At the pronoun region, we observed significant main effects of the gender of the DP in both second pass and total viewing times. In both measures, reading times were longer when the DP mismatched in stereotypical gender with the pronoun, as in (8c,d) compared to when there was a stereotypical gender match, as in (8a,b). The gender of the QP did not significantly affect reading times in any measure at this region. These results suggest that participants attempted to retrieve the syntactically licit DP antecedent. This pattern of results is illustrated in **Figure 1**, which shows the second pass times at the pronoun region.

At the spillover region, there was a significant main effect of DP gender in first pass times. Here, reading times were again longer when the DP mismatched in stereotypical gender with the pronoun compared to when there was a gender match. There was also a significant main effect of the gender of the DP in both second-pass and total viewing times, with reading times again being longer when the DP mismatched in stereotypical gender with the pronoun. There was additionally a marginally significant main effect of QP gender in total viewing times only. Here, reading times tended to also be longer when the QP mismatched in gender with the pronoun. There was no hint of an interaction however, as this numerical trend for longer reading times following gender mismatching QPs was observed in both the DP match and DP mismatch conditions.

At the prefinal region, there was again a significant main effect of DP gender in total viewing times, with reading times being longer when the DP mismatched in stereotypical gender with the pronoun. No significant effects of the gender of the QP were found at this region.

At the final region there was a marginally significant interaction in first-pass times. Here, in the DP stereotypical gender match conditions, reading times tended to be longer in QP match condition (8a) compared to QP mismatch condition (8b), but the planned comparison was not significant ($t = 1.476$, $p = 0.141$). The opposite numerical pattern was observed in the DP stereotypical gender mismatch conditions, but again the comparison was not significant ($t = 0.779$, $p = 0.437$). In regression path times the main effect of the stereotypical gender of the DP was significant, the main effect of the gender of the QP marginal, and the DP gender by QP gender interaction significant. In this measure, while reading times in DP match conditions (8a,b) did not differ ($t = 0.381$, $p = 0.703$), for the DP stereotypical gender mismatch conditions, reading times were longer in QP match condition (8c) than QP mismatch condition (8d) (estimate = 569, SD = 220, $t = 2.591$, $p = 0.010$). While this reading time measure thus provides evidence of the QP's gender

TABLE 5 | Reading times in milliseconds for four eye-movement measures at four regions of texts in Experiment 4 (SDs in parentheses).

	First pass reading time	Regression path time	Second pass time	Total viewing time
PRONOUN REGION				
DP gender match, QP gender match	251 (137)	356 (413)	96 (173)	332 (230)
DP gender match, QP gender mismatch	253 (124)	291 (185)	99 (184)	361 (278)
DP gender mismatch, QP gender match	256 (121)	334 (321)	149 (226)	405 (258)
DP gender mismatch, QP gender mismatch	253 (145)	312 (286)	144 (231)	375 (271)
SPILLOVER REGION				
DP gender match, QP gender match	278 (141)	368 (487)	168 (239)	445 (276)
DP gender match, QP gender mismatch	307 (168)	403 (442)	196 (265)	497 (308)
DP gender mismatch, QP gender match	323 (193)	411 (282)	230 (294)	542 (321)
DP gender mismatch, QP gender mismatch	339 (185)	482 (416)	226 (356)	564 (369)
PREFINAL REGION				
DP gender match, QP gender match	266 (144)	459 (507)	146 (182)	396 (228)
DP gender match, QP gender mismatch	279 (133)	556 (862)	144 (214)	410 (257)
DP gender mismatch, QP gender match	272 (128)	522 (682)	182 (255)	441 (271)
DP gender mismatch, QP gender mismatch	286 (163)	587 (845)	157 (228)	434 (269)
FINAL REGION				
DP gender match, QP gender match	342 (239)	1717 (1963)	74 (156)	423 (276)
DP gender match, QP gender mismatch	310 (228)	1799 (2285)	96 (220)	437 (327)
DP gender mismatch, QP gender match	338 (235)	2276 (2515)	116 (216)	486 (322)
DP gender mismatch, QP gender mismatch	366 (264)	1736 (1920)	80 (186)	464 (325)

TABLE 6 | Summary of statistical analyses for four eye-movement measures at four regions of texts in Experiment 4.

	First pass reading time		Regression path time		Second pass time		Total viewing time	
	Estimate	***t***	**Estimate**	***t***	**Estimate**	***t***	**Estimate**	***t***
PRONOUN REGION								
DP gender	4 (10)	0.377	4 (28)	0.144	47 (14)	3.448**	46 (19)	2.387*
QP gender	1 (12)	0.117	37 (35)	1.069	0 (16)	0.030	3 (20)	0.134
DP gender* QP gender	10 (21)	0.490	32 (57)	0.567	6 (27)	0.231	55 (36)	1.528
SPILLOVER REGION								
DP gender	37 (11)	3.209*	58 (46)	1.293	45 (21)	2.131*	81 (20)	3.968**
QP gender	22 (14)	1.617	51 (43)	1.194	12 (20)	0.603	38 (20)	1.884(*)
DP gender* QP gender	12 (27)	0.439	31 (61)	0.503	31 (39)	0.778	36 (39)	0.924
PREFINAL REGION								
DP gender	6 (12)	0.480	42 (78)	0.541	24 (15)	1.622	36 (18)	2.003*
QP gender	15 (12)	1.262	84 (65)	1.290	13 (20)	0.682	3 (19)	0.146
DP gender* QP gender	3 (25)	0.134	25 (122)	0.208	22 (33)	0.645	14 (35)	0.410
FINAL REGION								
DP gender	28 (17)	1.636	247 (125)	1.979*	10 (14)	0.709	48 (21)	2.261*
QP gender	7 (22)	0.303	250 (135)	1.846(*)	5 (13)	0.391	11 (21)	0.539
DP gender* QP gender	61 (33)	1.864(*)	627 (297)	2.109*	58 (37)	1.579	32 (49)	0.657

*Estimate = Model Estimate (SE in brackets). (*) = $p < 0.10$, * = $p < 0.05$, ** = $p < 0.001$.*

significantly influencing reading times, the direction of the effect in the DP stereotypical gender mismatch conditions is in the *opposite* direction to that predicted by facilitatory interference. Total viewing times at the final region exhibited reading times similar to earlier regions of text, with reading times being significantly longer when the DP mismatched in stereotypical gender with the pronoun compared to when there was a gender match. The QP did not significantly influence reading times in this measure.

Discussion

The results of Experiment 4 indicate that readers readily retrieved the syntactically licit DP antecedent upon encountering the pronoun. In a number of measures across all regions of text reported, we observed significantly longer reading times when the DP mismatched in stereotypical gender with the pronoun compared to when there was a gender match. Effects of the gender of the QP antecedent were more elusive and the one significant effect that we did observe was delayed in comparison to the effects that were observed of the DP's gender. While DP stereotypical gender mismatch effects were first observed in second pass and total viewing times at the pronoun, and first pass times at the spillover region, the only reliable effect of the gender of the QP was observed in the regression path times at the final region. We leave discussion of this delayed effect of the QP's gender until the General Discussion, but overall interpret the relative time-course of effects as indicating that the c-command constraint on variable binding restricts the initial stages of antecedent retrieval during comparatively earlier stages of anaphora resolution. We discuss the implications of these results, along with the other experiments reported above, in more detail below.

General Discussion

The aim of this study was to investigate if the c-command restriction on variable binding restricts antecedent retrieval during anaphora resolution. The results of Experiment 1 indicate that native English speakers are sensitive to the c-command restriction on binding by quantified antecedents in an offline judgment task. Experiment 2, which investigated the extent to which QP antecedents are retrieved upon encountering a pronoun during online processing, indicates that participants readily retrieved the QP upon encountering the pronoun, but only when the QP c-commanded the pronoun. The results of Experiment 3 showed that retrieval of DP antecedents, which is not contingent on c-command, was equally likely irrespective of whether or not the DP c-commanded the pronoun. The results of Experiments 2 and 3 together confirm that it is not the case that non c-commanding antecedents are generally ignored due to their lower discourse salience. Instead, both c-commanding and non c-commanding antecedents are readily retrieved, but only when they are syntactically licit antecedents for a pronoun. Finally, the results of Experiment 4 indicate that when only one syntactically licit antecedent is available in the discourse, that antecedent is preferentially retrieved over a syntactically illicit QP, even when the syntactically licit antecedent mismatches in gender with the pronoun. This different pattern of results across the three eye-movement experiments is illustrated in **Figure 1**. Together, these data indicate that the c-command constraint on variable binding restricts antecedent retrieval during anaphora resolution. Below we discuss the implications of these results with regards to how the c-command constraint on variable binding may be implemented in models of memory retrieval, and the relative weightings of different cues to antecedent retrieval during anaphora resolution.

Implementing the C-command Constraint

One potential way to help ensure that only syntactically licit QPs are retrieved during anaphora resolution might be to restrict at least initial memory access operations to antecedents that c-command a pronoun. This proposal would be similar to claims in the linguistics literature that variable binding is computed before coreference assignment (e.g., Reuland, 2001, 2011; Koornneef, 2008). In the current study, c-commanding antecedents always appeared in the main clause of the critical sentence, while non c-commanding antecedents appeared in relative clauses. The preference for retrieving a c-commanding antecedent in the current study could thus potentially be achieved by postulating that pronouns preferentially cue retrieval of an antecedent carrying a [+main clause] feature. However, this would also predict that non c-commanding DPs, even though they can be linked to the pronoun via coreference assignment, should also initially be ignored. Note however that while we observed selective retrieval of QPs in Experiment 2 and retrieval of DPs irrespective of c-command in Experiment 3, the time-course of gender mismatch effects across the pronoun and spillover regions in both experiments was very similar. If an initial retrieval favors antecedents carrying the [+main clause] feature only, we would have expected to see a delay in gender mismatch effects for non c-commanding DP antecedents compared to c-commanding DP antecedents in Experiment 3 that was not observed. The results from Cunnings et al. (2014) also clearly indicate that there is no initial preference for c-commanding over non c-commanding antecedents. Thus, we believe the hypothesis that initial retrieval operations should always simply ignore non c-commanding antecedents can be rejected. Nor can the retrieval operation initially target *only* referential antecedents or quantified ones, considering that Cunnings et al. (2014) observed no overall preference for either variable binding or coreference assignment.

Sensitivity to the c-command constraint on variable binding thus requires a restriction that selectively retrieves antecedents based on the c-command relationship between the pronoun and QP. As noted in the introduction, some have claimed that this type of *relational* constraint may be difficult to implement in content-addressable memory architectures (Kush, 2013; Kush et al., 2015). Kush et al. propose that one way to implement the c-command constraint on variable binding would be to encode all potential antecedents with an ACCESSIBLE feature that the parser is able to dynamically update based on the current state of the parse during incremental processing. That is, antecedents are always initially marked as [+accessible], but retrieval operations at specific points during an incremental parse may deactivate this feature if need be. We believe this proposal could account for our results as follows. In Experiments 2–4, the critical QP/DP (*every old man/woman; the old man/woman*) will initially be encoded as being [+accessible]. In the c-commanding QP/DP conditions, this feature will always remain activated. In the non c-commanding QP/DP conditions, upon reaching the right-most edge of the relative clause, a retrieval operation will access all antecedents within the relative clause, deactivating the ACCESSIBLE feature for QPs to ensure that they are no longer possible targets for retrieval, but leaving it unchanged for DPs. In this way, well-known clause "wrap-up" effects may in part involve updating items in a particular clause as being either accessible or inaccessible to further retrieval operations. Upon encountering the pronoun, the ACCESSIBLE feature will be a highly weighted cue to retrieval, activating DPs irrespective of c-command, but activating c-commanding QPs only.

Cue Weighting during Anaphora Resolution

The results of the current study indicate that the c-command constraint on variable binding, perhaps implemented using the ACCESSIBLE feature as above, is a highly weighted cue to antecedent retrieval. The gender mismatch effects observed in Experiment 2 indicate that participants will readily retrieve a c-commanding QP during processing, but we found no evidence of the QP being retrieved when it did not c-command the pronoun. In Experiment 4, when the QP was always syntactically illicit, we found that a number of reading time measures were significantly affected by the stereotypical gender of the syntactically licit DP only. The earliest measures where we observed this effect were those including first pass processing at the spillover region and second pass processing at the pronoun region.

Some models of memory retrieval assume that cues combine in an equally-weighted fashion to guide retrieval during language processing (e.g., Lewis et al., 2006). Evidence from facilitatory interference effects during subject-verb agreement processing for example, suggest that for at least some dependencies syntactic constraints and agreement markers are equally weighted cues to retrieval (e.g., Wagers et al., 2009). More recently however it has been claimed that retrieval cues during language processing are not always equally weighted (Van Dyke and McElree, 2011; Dillon et al., 2013). For example, Dillon et al. claimed that syntactic binding constraints constitute "hard constraints" that restrict retrieval to syntactically licit antecedents. We argued that the most obvious kind of evidence that the c-command constraint and gender congruence combine equally to guide retrieval would be from facilitatory interference effects similar to those observed for subject-verb agreement. However, we failed to observe this pattern of results in Experiment 4.

Although we failed to observe facilitatory interference, Badecker and Straub (2002) reported a different type of *inhibitory* interference in a series of self-paced reading experiments. They observed longer reading times when multiple antecedents matched in gender with a reflexive or pronoun compared to when there was only one gender matching antecedent. Such effects could indicate that when there are multiple gender matching antecedents in the discourse, both syntactically licit and illicit antecedents compete for retrieval. The clearest evidence of this type of interference in the current study would have been from longer reading times in Experiment 4 in multiple gender match condition (8a) compared to the single match condition (8b). However, we also failed to observe this type of effect.

The clearest evidence of the gender of the QP significantly affecting reading times that we did observe was in the *opposite* direction predicted by facilitatory interference, and was also dissimilar to the effects observed by Badecker and Straub (2002). In the regression path times for the final region in Experiment 4, reading times were significantly longer when the syntactically illicit QP matched the gender of the pronoun, but only when

the grammatically licit DP antecedent itself mismatched in gender with the pronoun. Note also that this effect of the QP's gender appears delayed in comparison to the significant main effects of the stereotypical gender of the syntactically licit DP.

In line with recent proposals that not all cues to memory retrieval are equally weighted during language processing (Van Dyke and McElree, 2011; Dillon et al., 2013), we argue that the c-command constraint on variable binding, implemented with the ACCESSIBLE feature, is a more highly weighted cue to antecedent retrieval than gender congruence during anaphora resolution. Whether or not the c-command restriction acts as a "hard constraint" that imposes a categorical ban on the retrieval of syntactically illicit antecedents is difficult to conclude. However, the relative time-course of effects observed for DP and QP antecedents in Experiment 4 may bear on this issue. Recall that in the DP stereotypical gender mismatch conditions in the regression path times of the final region in Experiment 4, we observed *longer* reading times when the QP matched the gender of the pronoun compared to when it mismatched. We remain cautious in interpreting precisely what this effect may index, but it could potentially indicate that readers sometimes attempted to coerce an interpretation in which the pronoun was linked to a syntactically illicit but gender matching QP antecedent, with this coercion of a syntactically illicit interpretation leading to longer reading times. The time-course of this effect, appearing at the sentence final region and delayed in comparison to stereotypical gender violations between the DP and pronoun, may indicate that it reflects a relatively late interpretive process that tries to coerce an otherwise dispreferred interpretation for the pronoun. Similar to Sturt's (2003) defeasible filter hypothesis, we propose that the time-course of effects observed in Experiment 4 may indicate that initially, retrieval operations attempt to retrieve syntactically licit antecedents only. Readers may sometimes try to coerce syntactically illicit interpretations during comparatively later stages of processing however, perhaps during reanalysis after initially retrieving a syntactically licit, but gender-mismatching antecedent. We note also however that other interpretations of this delayed effect are possible. Kush et al. (2015) for example, found that non c-commanding QPs did not influence reading times during early stages of anaphor resolution in sentences like (4b), but did find some suggestive evidence of delayed effects of non c-commanding QPs influencing processing in measures that included second-pass processing. They claimed that such delayed effects might index coercion of an additional referential antecedent for the pronoun from the set of antecedents implied by the quantifier. In this sense, when the pronoun matches the gender of a non c-commanding QP in sentences like (8c), the delayed effect we observed may index coercion of a referential antecedent (*an old woman*) from the set of antecedents implied by the QP (*every old woman*). We do not attempt to tease apart these two interpretations here. Irrespective of how these effects are to be interpreted, as they appear delayed in comparison to effects of syntactically licit antecedents, we maintain that retrieval operations initially attempt to retrieve grammatically licit antecedents only.

Finally, we note that counterexamples in which variable binding appears to be possible between a pronoun and antecedent irrespective of c-command have been discussed in the linguistics literature. For example, in *Every boy's mother says that he is special* the pronoun can be bound by the QP *every boy* even though the QP does not c-command the pronoun under the standard definition. Barker (2012) discusses a number of such counterexamples and claims that the restriction on variable binding should be recast in terms of semantic scope rather than c-command. The relative clause manipulation tested in the current study is a relatively clear-cut case where both traditional accounts and Barker would predict that variable binding is not permitted. Our results show that variable binding is not attempted during processing in such cases, at least during early stages of antecedent retrieval (see also Kush et al., 2015). The extent to which pronouns may trigger retrieval of non c-commanding QPs in other constructions is less well understood. Some researchers have investigated whether pronouns are linked to QPs in non c-commanding configurations other than the relative clause manipulation in the current study (e.g., Carminati et al., 2002; Kush et al., 2015, Experiment 1c), but these experiments used different diagnostics for dependency formation and did not use interference paradigms as in Experiment 4 here. One question that arises is whether retrieval of QPs upon encountering a pronoun is always, at least initially, restricted to c-commanding quantified antecedents, or whether in exceptional cases, as in sentences like *Every boy's mother says that he is special*, quantified antecedents are always accessible. How the c-command constraint and gender/number congruence interact to guide anaphora resolution in other constructions will thus be an important avenue of further research to investigate the extent to which the current findings generalize beyond the relative clauses tested here.

Conclusion

Across four experiments we investigated how constraints on pronoun interpretation influence the retrieval of quantified and non-quantified antecedents in different syntactic configurations. We found that variable binding between a pronoun and quantified antecedent was only attempted if the quantifier c-commanded the pronoun. Retrieval of coreference antecedents, which is not contingent on c-command, was attempted irrespective of c-command. We interpret these results as indicating that syntactic constraints restrict memory retrieval operations during anaphora resolution. We conclude that the c-command constraint on variable binding constitutes a highly weighted cue during anaphora resolution that, at least initially, guides retrieval operations to access syntactically licit antecedents only.

Acknowledgments

This research was supported by a British Academy Postdoctoral Fellowship to IC (pf100026) and by an Alexander-von-Humboldt-Professorship to Prof. Harald Clahsen (Potsdam Research Institute for Multilingualism), which is gratefully acknowledged. We also acknowledge the support of the Deutsche Forschungsgemeinschaft and the Open Access Publishing Fund of the University of Potsdam.

References

Baayen, H. (2008). *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R*. Cambridge: Cambridge University Press.

Baayen, H., Davidson, D., and Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *J. Mem. Lang.* 59, 390–412. doi: 10.1016/j.jml.2007.12

Badecker, W., and Straub, K. (2002). The processing role of structural constraints on the interpretation of pronouns and anaphors. *J. Exp. Psychol.* 28, 748–769. doi: 10.1037/0278-7393.28.4.748

Barker, C. (2012). Quantificational binding does not require c-command. *Ling. Inq.* 43, 614–633. doi: 10.1162/ling_a_00108

Barr, D., Levy, R., Scheepers, C., and Tily, H. (2013). Random-effects structure for confirmatory hypothesis testing: keep it maximal. *J. Mem. Lang.* 68, 255–278. doi: 10.1016/j.jml.2012.11.001

Bosch, P. (1983). *Agreement and Anaphora: A Study of the Role of Pronouns in Syntax and Discourse*. London: Academic Press.

Carminati, M., Frazier, L., and Rayner, K. (2002). Bound variables and c-command. *J. Semant.* 19, 1–34. doi: 10.1093/jos/19.1.1

Chomsky, N. (1981). *Lectures on Government and Binding*. Dordrecht: Foris.

Chow, W-Y., Lewis, S., and Phillips, C. (2014). Immediate sensitivity to structural constraints in pronoun resolution. *Front. Psychol.* 5:630. doi: 10.3389/fpsyg.2014.00630

Clackson, K., Felser, C., and Clahsen, H. (2011). Children's processing of reflexives and pronouns in English: evidence from eye-movements during listening. *J. Mem. Lang.* 65, 128–144. doi: 10.1016/j.jml.2011.04.007

Clackson, K., and Heyer, V. (2014). Reflexive anaphor resolution in spoken language comprehension: structural constraints and beyond. *Front. Psychol.* 5:904. doi: 10.3389/fpsyg.2014.00904

Cunnings, I., and Felser, C. (2013). The role of working memory in the processing of reflexives. *Lang. Cogn. Process.* 28, 188–219. doi: 10.1080/01690965.2010.548391

Cunnings, I., Patterson, C., and Felser, C. (2014). Variable binding and coreference in sentence comprehension: evidence from eye movements. *J. Mem. Lang.* 71, 39–56. doi: 10.1016/j.jml.2013.10.001

Cunnings, I., and Sturt, C. (2014). Coargumenthood and the processing of reflexives. *J. Mem. Lang.* 75, 117–139. doi: 10.1016/j.jml.2014.05.006

Dillon, B., Mishler, A., Sloggett, S., and Phillips, C. (2013). Contrasting intrusion profiles for agreement and anaphora: experimental and modelling evidence. *J. Mem. Lang.* 69, 85–103. doi: 10.1016/j.jml.2013.04.003

Evans, G. (1980). Pronouns. *Ling. Inq.* 11, 337–362.

Kazanina, N., Lau, E., Lieberman, M., Yoshida, M., and Phillips, C. (2007). The effect of syntactic constraints on the processing of backwards anaphora. *J. Mem. Lang.* 56, 384–409. doi: 10.1016/j.jml.2006.09.003

Koornneef, A. (2008). *Eye-Catching Anaphora*. Utrecht: LOT International Dissertation Series.

Kush, D. (2013). *Respecting Relations: Memory Access and Antecedent Retrieval in Incremental Sentence Processing*. Unpublished doctoral thesis, University of Maryland.

Kush, D., Lidz, J., and Phillips, C. (2015). Relation-sensitive retrieval: evidence from bound variable pronouns. *J. Mem. Lang.* 82, 18–40. doi: 10.1016/j.jml.2015.02.003

Lewis, R. L., Vasishth, S., and Van Dyke, J. A. (2006). Computational principles of working memory in sentence comprehension. *Trends Cogn. Sci. (Regul. Ed).* 10, 44–54. doi: 10.1016/j.tics.2006.08.007

McElree, B. (2000). Sentence comprehension is mediated by content-addressable memory structures. *J. Psycholinguist. Res.* 29, 111–123. doi: 10.1023/A:1005184709695

McElree, B., Foraker, S., and Dyer, L. (2003). Memory structures that subserve sentence comprehension. *J. Mem. Lang.* 48, 67–91. doi: 10.1016/S0749-596X(02)00515-6

Patterson, C., Trompelt, H., and Felser, C. (2014). The online application of binding condition B in native and non-native pronoun resolution. *Front. Psychol.* 5:147. doi: 10.3389/fpsyg.2014.00147

Rayner, K. (1998). Eye-movements in reading and information processing: 20 years of research. *Psychol. Bull.* 124, 372–422.

Reinhart, T. (1983). *Anaphora and Semantic Interpretation*. London: Croom Helm.

Reuland, E. (2001). Primitives of binding. *Ling. Inq.* 32, 439–492. doi: 10.1162/002438901750372522

Reuland, E. (2011). *Anaphora and Language Design*. Cambridge, MA: The MIT Press.

Sanford, A., and Garrod, S. (1988). Proper names as controllers of discourse focus. *Lang. Speech* 31, 43–56.

Sturt, P. (2003). The time-course of the application of binding constraints in reference resolution. *J. Mem. Lang.* 48, 542–562. doi: 10.1016/S0749-596X(02)00536-3

Van Dyke, J., and Johns, C. (2012). Memory interference as a determinant of language comprehension. *Lang. Linguist. Compass* 6, 193–211. doi: 10.1002/lnc3.330

Van Dyke, J., and McElree, B. (2011). Cue-dependent interference in comprehension. *J. Mem. Lang.* 65, 246–263. doi: 10.1016/j.jml.2011.05.002

Wagers, M., Lau, E., and Phillips, C. (2009). Attraction in comprehension: representation and processes. *J. Mem. Lang.* 61, 206–237. doi: 10.1016/j.jml.2009.04.002

Xiang, M., Dillon, B., and Phillips, C. (2009). Illusory licensing effects across dependency types: ERP evidence. *Brain Lang.* 108, 40–55. doi: 10.1016/j.bandl.2008.10.002

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2015 Cunnings, Patterson and Felser. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Appendix A

The materials for Experiment 1 are provided below. The c-command manipulation is denoted in square brackets, delimited with a forward slash (/).

1 The soldier [suspected that every old man at the bar/who every old man at the bar suspected] suddenly wished that he had drunk a bit less that afternoon.
2 The footballer [observed that every salesman in the shop/who every salesman in the shop observed] secretly thought that he could offer some advice.
3 The butcher [heard that every boy in the street/who every boy in the street heard] loudly said that he could not lift the heavy box.
4 The farmer [saw that every schoolboy in the field/who every schoolboy in the field saw] truly feared that he would disturb the wildlife.
5 The pilot [knew that every waiter in the café/who every waiter in the café knew] really hoped that he would avoid getting ill.
6 The sailor [trusted that every prince at the royal celebration/who every prince at the royal celebration trusted] clearly understood that he should not join in the dance.
7 The surgeon [suggested that every man on the waiting list/who every man on the waiting list suggested] definitely realized that he needed some help.
8 The builder [noticed that every father at the school gate/who every father at the school gate noticed] openly stated that he should get back to work soon.
9 The nurse [noticed that every old woman in the hospital/who every old woman in the hospital noticed] genuinely expected that she would be allowed to leave very soon.
10 The secretary [saw that every saleswoman in the meeting/who every saleswoman in the meeting saw] obviously worried that she was not making a contribution.
11 The babysitter [observed that every girl in the house/who every girl in the house observed] thoroughly wished that she could reach the biscuit tin.
12 The kindergarten teacher [trusted that every schoolgirl in the class/who every schoolgirl in the class trusted] earnestly believed that she could improve during the term.
13 The housekeeper [knew that every waitress at the event/who every waitress at the event knew] rightly realized that she could create a good atmosphere.
14 The cleaner [suspected that every princess in the castle/who every princess in the castle suspected] truly thought that she could find the lost set of keys.
15 The typist [heard that every woman in the office/who every woman in the office heard] suddenly decided that she should leave early that day.
16 The florist [suggested that every mother on the committee/who every mother on the committee suggested] sincerely hoped that she could get the flowers in time.

Appendix B

The materials for Experiment 2 are provided below. Gender manipulations are shown in parenthesis and square brackets denote the c-command manipulation, delimited with a forward slash (/).

1 Being in hospital can be quite difficult at times. The surgeon [saw that every old man (woman) on the emergency ward/who every old man (woman) on the emergency ward saw] silently wished that he could go a little bit faster.
2 A fight had broken out at the celebration dinner! The footballer [noticed that every waiter (waitress) in the hotel/who every waiter (waitress) in the hotel noticed] quickly realized that he should move away from the table.
3 There was only one way to sneak into the palace party. The soldier [suspected that every prince (princess) outside the castle/who every prince (princess) outside the castle suspected] honestly doubted that he could climb over the high wall.
4 The company was spending a lot of money on a new building. The builder [suggested that every businessman (businesswoman) in the firm/who every businessman (businesswoman) in the firm suggested] sincerely thought that he could do the job much better.
5 There had been a big fire at the local high-school. The firefighter [heard that every father (mother) in the crowd/who every father (mother) in the crowd saw] quickly declared that he should be allowed to go back.
6 The tour of the new local airport was very enjoyable. The pilot [saw that every boy (girl) near the large plane/who every boy (girl) near the large plane saw] clearly expected that he would go up the steps first.
7 There was an uneasy atmosphere after the big match. The boxer [trusted that every policeman (policewoman) in the room/who every policeman (policewoman) in the room trusted] soon realized that he could reach the back exit safely.
8 Trading always starts very early during the morning. The butcher [suspected that every salesman (saleswoman) at the market/who every salesman (saleswoman) at the market suspected] secretly hoped that he could decide on the new price.
9 It was an extremely hot afternoon, even out at sea. The sailor [believed that every schoolboy (schoolgirl) on the boat/who every schoolboy (schoolgirl) on the boat believed] genuinely thought that he could jump into the cold water.
10 It's not very good to be selfish all of the time. The farmer [knew that every little boy (girl) in the small village/who every little boy (girl) in the small village knew] slowly realized that he must share the apples from the orchard.
11 Water had been pouring into the kitchen for a whole day. The plumber [suggested that every young man (woman) in the house/who every young man (woman) in the house suggested] seriously thought that he could stop the leak very easily.
12 It is unusual to talk openly about personal finances. The priest [heard that every man (woman) at the village church/who

every man (woman) at the village church heard] suddenly claimed that he should give more money to charity.

13 Modern life can be a little bit too busy at times. The cleaner [knew that every old woman (man) in the care-home/who every old woman (man) in the care-home knew] truly wished that she had much more time to talk.

14 There had been a lot of changes to the menu that night. The fashion model [heard that every waitress (waiter) in the restaurant/who every waitress (waiter) in the restaurant heard] incorrectly feared that she would get the wrong evening meal.

15 It was the beginning of another very important day. The housekeeper [trusted that every princess (prince) in the old castle/who every princess (prince) in the old castle trusted] clearly understood that she must make a very good impression.

16 There were going to be some big changes at the large company. The secretary [suspected that every businesswoman (businessman) at the meeting/who every businesswoman (businessman) at the meeting suspected] correctly realized that she could change the terms of the contract.

17 Life can get very tiring if you don't find time to rest. The babysitter [trusted that every mother (father) in the big group/who every mother (father) in the big group trusted] eventually realized that she needed to relax now and again.

18 The flower festival was always a very busy time of year. The florist [believed that every girl (boy) in the beautiful garden/who every girl (boy) in the beautiful garden believed] wishfully thought that she could smell all the lovely roses.

19 The carnival always seems to get bigger every year. The nurse [knew that every policewoman (policeman) at the annual event/who every policewoman (policeman) at the annual event knew] fully understood that she must help with any big emergency.

20 Working hard can sometimes bring great rewards. The typist [noticed that every saleswoman (salesman) in the firm/who every saleswoman (salesman) in the firm noticed] never doubted that she would get a big bonus payment.

21 The baseball match was due to start in 2 h. The cheerleader [noticed that every schoolgirl (schoolboy) in the team/who every schoolgirl (schoolboy) in the team noticed] quietly expected that she would stand in the front row.

22 It was yet another really hectic day at the theater in town. The beautician [saw that every little girl (boy) in the glamorous show/who every little girl (boy) in the glamorous show saw] mistakenly worried that she would mess up the expensive costumes.

23 There had been a lot of rumors going around for a while. The fortune-teller [believed that every young woman (man) at the fair/who every young woman (man) at the fair believed) unexpectedly claimed that she was not telling the whole story.

24 It can sometimes be very difficult to ask for assistance. The kindergarten teacher [suggested that every woman (man) in the village/who every woman (man) in the village suggested] wrongfully thought that she could manage without any extra help.

frontiers in Psychology

ORIGINAL RESEARCH
published: 30 October 2015
doi: 10.3389/fpsyg.2015.01638

Active search for antecedents in cataphoric pronoun resolution

Leticia Pablos[1,2], Jenny Doetjes[1], Bobby Ruijgrok[1,2] and Lisa L.-S. Cheng[1,2]*

[1] Leiden University Center for Linguistics, Leiden University, Leiden, Netherlands, [2] Leiden Institute for Brain and Cognition, Leiden University, Leiden, Netherlands

Cataphoric dependencies where a pronoun precedes its antecedent appear to call on different mechanisms in language comprehension from forward dependencies where the antecedent precedes the pronoun. Previous research has shown that the resolution of cataphoric dependencies involves predictive processes such as the active search mechanism, which hypothesizes the automatic search for an antecedent immediately after encountering a cataphoric pronoun. The current study employs gender mismatch to investigate whether the active search for an antecedent of a cataphoric pronoun is restricted only to grammatically licit positions. We present results from an event-related potential experiment on the reading comprehension of cataphoric dependencies in Dutch. Results show that gender mismatch gives rise to an anterior negativity at grammatically licit antecedent positions only. We hypothesize that this negativity reflects the prediction failure for an antecedent after encountering a pronoun, rather than a gender mismatch. We discuss the timing, topography and functionality of this negativity with respect to previous studies and how this relates to the ERPs elicited in the processing of structural constraints on pronoun resolution.

Keywords: cataphora, active search, gender mismatch, anterior negativity, Principle C

OPEN ACCESS

Edited by:
Colin Phillips,
University of Maryland, USA

Reviewed by:
Stephen Politzer-Ahles,
University of Oxford, UK
Sol Lago,
University of Potsdam, Germany

****Correspondence:***
Leticia Pablos
l.pablos.robles@hum.leidenuniv.nl

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: *14 July 2015*
Accepted: *12 October 2015*
Published: *30 October 2015*

Citation:
Pablos L, Doetjes J, Ruijgrok B and Cheng LL-S (2015) Active search for antecedents in cataphoric pronoun resolution. Front. Psychol. 6:1638. doi: 10.3389/fpsyg.2015.01638

INTRODUCTION

The on-line interpretation of pronominal dependencies has raised several questions within theories of sentence comprehension. Forward pronominal dependencies – where the antecedent precedes the pronoun – and backward pronominal dependencies – where the pronoun precedes the antecedent – appear to call on different mechanisms in language comprehension. In the case of forward dependencies, their resolution requires retrieving the information about the antecedent at the position of the pronoun, which is closely connected with memory-retrieval processes (Chow et al., 2014). On the other hand, the resolution of backward dependencies (also called cataphoric dependencies) requires the search for an antecedent, which is related to predictive processes.

One of such predictive processes is the active search mechanism (ASM), found initially for the interpretation of *wh*-gap dependencies (Crain and Fodor, 1985; Stowe, 1986; Clifton and Frazier, 1989). In the case of backward dependencies, the ASM hypothesizes that the human parser automatically starts a search for an antecedent in the upcoming sentence immediately after encountering a cataphoric pronoun. This has been shown in behavioral studies through gender mismatch effect (GMME) observations in experimental paradigms where possible antecedents for cataphoric pronouns are restricted by grammatical principles (Sturt, 2003; Van Gompel and Liversedge, 2003; Kazanina et al., 2007; Yoshida et al., 2014). This paper presents an event related potential (ERP) study where we confirm that a similar effect can also be observed in

neurophysiological data. Results support the presence of an ASM for cataphoric dependency resolution that respects grammatical principles. The topography and timing of the ERP component generated at the mismatching antecedent position in our study was an anterior negativity, while previous forward antecedent/pronoun dependencies studies have found a P600 (Osterhout and Mobley, 1995; Van Berkum et al., 2007; Xu et al., 2013). We postulate that the ERP component observed in cataphoric dependencies is related to a failure of a prediction by the parser, in line with the active search approach, while in the case of forward antecedent dependencies the effect can only be connected to a gender mismatch as no prediction is made (the pronoun is not required to interpret the antecedent).

Cataphoric Dependencies

Cataphoric pronouns are pronouns that occur linearly before their antecedent. In other words, they are instances of referential dependencies in which the antecedent follows the referentially dependent element, as illustrated in (1). The index *i* indicates that he_i and $Peter_i$ refer to the same person.

(1) While **he$_i$** had a broken arm, **Peter$_i$** could not ride his bike.

Pronouns such as *he* in (1) pose an interesting case for parsing theories. In order to resolve the interpretation of the pronoun with an antecedent in the same sentence, the parser needs to wait until the appearance of the antecedent. When the antecedent is found, the pronoun can establish a link with it for its own interpretation. However, this is only possible when the grammar allows the link between the cataphoric pronoun and the antecedent to be established. Consider the pronoun *he* in (2) and the pronoun *his* in (3). In contrast to the pronoun in (1), the pronoun in (2) cannot take the proper name *Peter* as its antecedent (as indicated by the starred index of *j* – *he* cannot have the same index/reference as *Peter*). However, *Peter* can be the antecedent of the pronoun *his* in (3).

(2) **He$_{i/*j}$** said that **Peter$_j$** is sick.
(3) **His$_i$** brother said that **Peter$_i$** is sick.

The restriction of the pronominal reference in (1), (2), and (3) can be captured under the principles of the Binding Theory (Chomsky, 1981) that indicates the configurations in which nominal elements can or cannot establish a coreferential relation. There are three Binding Principles, each of which concerns a different type of nominal elements. Binding Principles A and B are concerned with two different types of pronouns (*himself* vs. *him*), while Principle C restricts the distribution of Referential Expressions, including proper names such as *Peter*.

We focus on Principle C, which prohibits a Referential Expression (e.g., proper name) from being *bound* (Chomsky, 1981). The pronoun *he* in (1) does not bind the referential expression *Peter*, because the pronoun is embedded in an adverbial clause that does not contain *Peter*. Given that *he* does not bind *Peter*, the two can have the same reference. On the other hand, the pronoun *he* in (2) binds the referential expression *Peter* structurally and in such a case, coreference is excluded. *His* in (3) on the other hand is more deeply embedded in the structure (i.e., in the noun phrase *his brother*), and therefore, it does not act as a binder of *Peter*. Thus, similar to (1), a cataphoric dependency can be established in (3). Referential expressions, such as *John* or *the man*, independently refer and select a referent from the domain of discourse. Given that Referential Expressions have independent reference, they do not need and in fact cannot tolerate a binder. The binder would act as an antecedent for the Referential Expression, which is in conflict with the referential status of the latter.

In this study, we investigate whether Principle C of the Binding Theory is respected in cataphoric pronoun processing. As illustrated in (1), (2), and (3), whether a referential expression can be a potential antecedent for a cataphoric pronoun depends on the structural configuration. If a coreferential relation is established between a referential expression and a cataphoric pronoun and as a result, the referential expression is bound by the pronoun, Principle C of the Binding Theory would be violated. This paper examines how this type of violation affects parsing. In particular, it uses gender mismatch to investigate whether a search for an antecedent is restricted by structural constraints. Given that the parser respects structural constraints such as Principles B and C of the Binding Theory when interpreting pronouns on-line as shown by behavioral studies that have examined reading times (e.g., Kazanina et al., 2007; Chow et al., 2014; Yoshida et al., 2014), we expect these effects to be visible through electroencephalography (EEG) as well.

Active Search Mechanism [or Active Filler Hypothesis (AFH)]

The ASM claims that an active search is automatically initiated for each uninterpreted element A encountered in a sentence, to find the element B which can help interpret A. The main evidence for the existence of the active search comes from the so-called filled-gap effects involving *wh*-dependencies, which demonstrate that (a) a search for a gap starts as soon as a *wh*-phrase is processed and (b) filling the gap position where the *wh*-word could be interpreted with an overt element (thus blocking the parser from interpreting the *wh*-phrase in that position) results in a longer processing time compared to a sentence where no *wh*-dependency was initiated (Crain and Fodor, 1985; Stowe, 1986; Lee, 2004). Thus, the ASM hypothesizes that the parser anticipates a gap as soon as a *wh*-phrase is processed (Clifton and Frazier, 1989; Frazier and Clifton, 1989).

In the case of pronoun interpretation, the ASM predicts that a search is initiated for an antecedent as soon as a pronoun is encountered (Clifton and Frazier, 1989; Kazanina et al., 2007), in order to resolve the interpretation of the pronoun. Even though pronouns may have antecedents outside of the sentence that contains them, the ASM assumes that the search for an antecedent within the sentence is the default strategy in cases where there is no preceding discourse.

Studies on the processing of cataphoric pronouns have examined whether the parser indeed searches for an antecedent in the sentence once a pronoun has been processed and when the grammar allows the establishment of the binding relation (Sturt, 2003; Van Gompel and Liversedge, 2003; Kazanina et al., 2007;

Yoshida et al., 2014). In these behavioral studies, which used eye-tracking or self-paced reading methodology, the parser searches for an antecedent in the upcoming input in positions where the coreference between the pronoun and the antecedent is allowed (i.e., such coreference does not lead to a violation of the Binding Theory). In such cases, when the potential antecedent does not match in gender with the preceding pronoun, reading times are longer than when the potential antecedent and the pronoun match in gender (Sturt, 2003; Van Gompel and Liversedge, 2003; Kazanina et al., 2007; Yoshida et al., 2014). This reading slowdown effect, known as the GMME, has been taken to be a sign of the parser's active search for an antecedent to interpret the pronoun. Importantly, the data in these studies show that the GMME does not occur if the coreference between the pronoun and the referential expression yields a violation of the Binding Theory (in particular, Condition C), suggesting that in such cases, the referential expression does not count as a potential antecedent for the pronoun.

The main hypothesis of Kazanina et al. (2007) word-by-word self-paced reading experiments is that the parser respects Principle C of the Binding Theory when searching for an appropriate antecedent for a pronoun. This can be illustrated on the basis of the four different conditions in (4), which are from their third experiment: no constraint match in (4a), no constraint mismatch in (4b), Principle C match in (4c) and Principle C mismatch in (4d).

(4) **a. No constraint/Match**
His$_i$ managers chatted amiably with some fans while **the talented, young quarterback$_i$** signed autographs for the kids, but **Carol** wished the children's charity event would end soon so she could go home.
b. No constraint/Mismatch
Her$_i$ managers chatted amiably with some fans while **the talented, young quarterback** signed autographs for the kids, but **Carol$_i$** wished the children's charity event would end soon so she could go home.
c. Principle C/Match
He$_i$ chatted amiably with some fans while **the talented, young quarterback** signed autographs for the kids, but **Steve$_i$** wished the children's charity event would end soon so he could go home.
d. Principle C/Mismatch
She$_i$ chatted amiably with some fans while **the talented, young quarterback** signed autographs for the kids, but **Carol$_i$** wished the children's charity event would end soon so she could go home.

In the no constraint match condition in (4a), the possessive pronoun *his*, being further embedded in the nominal structure, does not bind the referential expression *young quarterback,* allowing it to be a potential antecedent. In other words, in (4a), Principle C does not block the coreference relation between the pronoun *his* and the referential expression (the antecedent *young quarterback*), and these two elements match in gender. Therefore the cataphoric pronoun should be interpreted at the antecedent position. The no constraint mismatch condition in (4b) differs from the no constraint match condition in (4a), in that the gender of the pronoun *her* and that of the potential antecedent *young quarterback* do not match, creating a GMME. In the Principle C match condition in (4c), on the other hand, the pronoun *he* binds the referential expression *young quarterback* in the embedded clause. Thus, *young quarterback* is excluded as a potential antecedent of *he* due to a Principle C violation. Furthermore, both the pronoun *he* and the referential expression *young quarterback* match in gender, as both are masculine. Finally, in the Principle C mismatch condition in (4d) the pronoun *she*, binds the referential expression *young quarterback* in the embedded clause, just like in (4c); however, in this case, they mismatch in gender. Importantly, the GMME is expected to be absent in the Principle C mismatch condition (condition 4d) at the position of the referential expression *young quarterback*, relative to the Principle C match condition (4c), as the coreference relation is barred from being established due to Principle C, preventing the GMME to occur. Conversely, the GMME is expected to be present at the referential expression *young quarterback* position in the no constraint mismatch condition (4b), relative to the no constraint match condition (4a). The main findings of Kazanina et al. (2007) confirm these expectations. Their reading time results thus suggest that the parser abides by Principle C when it attempts to resolve the interpretation of cataphoric pronouns in real-time in that they only find a reading time difference, or GMME, in the no constraint conditions, in which the referential expression in the no constraint mismatch condition in (4b) elicited longer reading times than the no constraint match condition (4a) at the same position (in particular, at the noun *quarterback*), whereas this reading time difference was absent at the referential expression in the Principle C conditions in (4c) and (4d). Furthermore, Kazanina et al. (2007) claim that the active search for an antecedent in cataphoric configurations only occurs when the Binding Principles allow it.

Yoshida et al. (2014) examine the formation of cataphoric dependencies across a relative clause island in a word-by-word self-paced reading experiment and they expect to obtain a GMME, or longer reading times, only in cases where coreference between the pronoun and the antecedent is licit (i.e., not obeying Principle C). Further, the GMME would only be expected to occur if cataphoric dependencies were not to be sanctioned across relative clause islands. Similar to Kazanina et al. (2007), Yoshida et al. (2014) manipulated the sentence initial pronoun [nominative vs. (possessive) genitive], the gender of the pronoun and the first referential expression. Their stimuli are shown in (5). In (5a) and (5b) the pronouns *his/her* can corefer with the referential expression *Jeffrey Stewart* (thus, *Jeffrey Stewart* can be a potential antecedent), but in (5c) and (5d) coreference is not licit due to Principle C of the Binding Theory.

(5) **a. No Constraint/Match**
His$_i$ managers revealed that the studio that notified **Jeffrey Stewart$_i$** about the new film selected a novel for the script, but **Annie** did not seem to be interested in this information.

b. No constraint/Mismatch
Her$_i$ managers revealed that the studio that notified **Jeffrey Stewart** about the new film selected a novel for the script, but **Annie$_i$** did not seem to be interested in this information.
c. Principle C/Match
He$_i$ revealed that the studio that notified **Jeffrey Stewart** about the new film selected a novel for the script, but **Andy$_i$** did not know which one.
d. Principle C/Mismatch
She$_i$ revealed that the studio that notified **Jeffrey Stewart** about the new film selected a novel for the script, but **Annie$_i$** did not know which one.

A GMME or reading slowdown is found at the antecedent position *Jeffrey Stewart* (in particular, at the last name *Stewart*) in (5b) relative to (5a), where the pronoun and the antecedent could corefer (the coreference does not violate Principle C). Moreover, the GMME or reading time difference occurs despite the fact that the potential antecedent is contained within a relative clause island. The GMME generated in the no constraint conditions (5a) and (5b) in the self-paced reading experiment by Yoshida et al. (2014) confirms that online formation of a cataphoric dependency is not affected by island constraints in that coreference is established in (5a) and (5b) conditions when the grammatical constraint of Principle C does not ban this coreference. If island constraints affected the generation of a cataphoric dependency we will not expect a GMME to occur in no constraint conditions, which it does. Furthermore, these results support the claim in Kazanina et al. (2007), that the processing of cataphoric dependencies is modulated by a grammatically constrained ASM, which respects grammatical principles such as Principle C.

The current study aims to replicate the GMME results from previous studies (Sturt, 2003; Van Gompel and Liversedge, 2003; Kazanina et al., 2007; Yoshida et al., 2014; a.o.) using ERP, to identify a neural correlate of the ASM found in the on-line interpretation of cataphoric dependencies. If an active-search is initiated for these dependencies (as shown by previous behavioral studies through the generation of the GMME effect, which is a slowdown in the gender mismatching conditions relative to the gender matched ones), it should be possible to identify an effect (i.e., an ERP component) comparable to the reading time differences shown in behavioral studies with the ERP methodology. In other words, we predict there to be a GMME in the no constraint mismatch conditions such as (4b) and (5b) above, relative to the no constraint match conditions in (4a) and (5a).

Event-related Potential (ERP) Studies on Gender Agreement/Mismatch

Since the current study examines gender agreement mismatches at the antecedent position in cataphoric configurations, a brief overview of the ERP studies that have tackled gender agreement issues is in order. Gender agreement mismatches have been examined in the ERP literature using different paradigms. Wicha et al. (2004) found a P600 for gender disagreeing nouns in determiner-noun combinations in Spanish, where the expected noun mismatched in gender with the preceding determiner. Van Berkum et al. (2005) on the other hand tested the prediction for the likely appearance of a specific noun based on the previous discourse. Their aim was to examine how listeners use their discourse knowledge to predict specific nouns. If listeners anticipate a noun with a specific gender by the time they encounter the indefinite article (not gender marked) in the story, a gender-mismatched adjective (i.e., mismatched in accordance to the gender of the noun that is expected) would be a surprise, leading to an ERP effect at the adjective position. They tested Dutch sentences where the sentence continuations had either an adjective consistently gender-marked with the upcoming predicted noun and its gender, or an adjective inconsistently gender-marked with respect to the prediction made for the upcoming noun and its gender. Their results again showed a P600 for gender-mismatched adjectives.

In a different set of studies, gender agreement violations between a determiner and a noun, or between an adjective and a noun, showed a left anterior negativity (LAN) followed by a P600 at the noun position for Spanish, Italian, and German (Demestre et al., 1999; Gunter et al., 2000; Barber and Carreiras, 2005; Molinaro et al., 2008; a.o.), a P600 for English and Dutch (Hagoort and Brown, 1999) and a N400 followed by a P600 for Hebrew (Deutsch and Bentin, 2001).

Finally, in a third set of studies, gender violations were tested in forward pronoun resolution dependencies, i.e., dependencies in which antecedents occur before pronouns. Osterhout and Mobley (1995) tested sentences such as (6) where a masculine or feminine pronoun matched or mismatched in gender with a previously encountered antecedent. They found a P600 at the pronoun *he* that mismatched in gender with the previously encountered feminine antecedent *the aunt*. Note that coreference between *he* and *the aunt* is only blocked by the gender mismatch and not by the Binding Conditions, as pronouns, contrary to referential expressions, may be bound by their antecedent if the antecedent is located in a different clause (cf. Principle B of the binding theory).

(6) **The aunt** heard that **she/he** had won the lottery.

Similarly, studies that tested gender violations in comparable forward pronoun configurations in Dutch (Van Berkum et al., 2007) and Chinese (Xu et al., 2013) found a P600 at the position of the pronoun when it mismatched in gender with the preceding antecedent.

Taking into consideration the results in these studies that have manipulated gender agreement, it is clear that a P600 component emerges constantly, regardless of whether the relation is one between (1) a determiner and a noun; (2) an adjective and a noun; or (3) an antecedent and a pronoun. While the P600 is preceded by a LAN or by a N400 in some cases in pure pronoun resolution cases more akin to the manipulation in the current study, only a P600 is obtained at the position of the gender-mismatched pronoun.

The Current Study

As indicated above, the present study examines processing of pronouns and their antecedents in a *cataphoric configuration*,

where the pronoun linearly precedes the antecedent. To summarize, the aim of this study is threefold. (i) First is to examine whether there is a GMME when the parser encounters the first potential antecedent of the cataphoric pronoun that does not match in gender. This would be an indication that the parser starts actively searching for a matching antecedent after encountering the cataphoric pronoun, even though the antecedent of the pronoun could, in principle, be found outside of the sentence. We predict the GMME to be present in the case of a mismatch, and absent in the matching condition. (ii) Second, we examine if the search mechanism is modulated by grammatical constraints such as Principle C of the Binding Theory. For cases where co-reference may lead to Principle C violations, we predict no difference between the match and the mismatch conditions. We predict that an ERP component is elicited only for referential expressions that can legitimately establish a co-reference relation with the cataphoric pronoun. (iii) Third, we examine if cataphoric pronoun dependencies generate the same kind of ERP components as forward pronoun dependencies. As discussed above, previous studies (e.g., Osterhout and Mobley, 1995; Van Berkum et al., 2007; Xu et al., 2013; a.o.) examined forward dependencies. However, no ERP study has examined cataphoric dependencies where the pronoun precedes the antecedent.

We aimed to search for the neuronal correlates of the ASM by means of a technique that has an excellent temporal resolution and where the effects of the active search can be examined by looking directly at brain behavior.

MATERIALS AND METHODS

Materials

Thirty-six experimental items were constructed in Dutch. These 36 items were distributed across four lists in a Latin Square design, which implies that each participant saw nine trials per condition. We decided on the relatively small number of trials per condition for a number of reasons: (a) The GMME effect has been quite reliable in the behavioral literature. Thus, we expect the size effect of the gender mismatch to be robust; (b) we would like to avoid reading fatigue as well as participant developing different processing strategies derived from the high number of proper names included in the items. Note that previous studies, which investigated the processing of coreference involving repeated nouns with the ERP technique, used a higher number of trials per condition for their experiments (i.e., 40 trials per condition; see for example, Swaab et al., 2004; Ledoux et al., 2007). However, the research questions of these studies and our initial question do not overlap, since these studies were examining word repetition-priming effects and the impact this factor had on the modulation of the N400 ERP component, whereas our interest lays in the process of coreference itself. The vast majority of ERP experiments in the field present every participant with 20–40 items per condition, but this is because the ERP effects that the experimenters are after are often rather small. Likewise, the use of a large number of trials is often connected to the fact that usually some trials are discarded due to artifacts or to the type of ERP component that the researchers are after, which might be different in size (see for example, Luck, 2005; Kaan, 2007 for further discussion of this specific issue).

We followed closely the set-up of the English word-by-word self-paced reading experiment by Kazanina et al. (2007) while creating our ERP experiment, since we were interested in seeing the time-course of the GMME using ERPs. There are four experimental conditions, as shown in (7). First, No-Constraint conditions, which contain a possessive pronoun, in masculine (7a) or feminine form (7b) that matches or mismatches, respectively, in gender with the linearly first antecedent *Lodewijk* (masculine). Second, Principle C conditions, which contain a cataphoric nominative pronoun in masculine (7c) or feminine form (7d) that cannot co-refer with the referential expression *Lodewijk* in the embedded clause due to Principle C.

In all conditions, the test sentences always contain a licit antecedent for the pronoun. For example, in the No-Constraint mismatch condition in (7b) and in Principle C conditions in (7c) and (7d), the pronouns corefer with an antecedent that appears toward the end of each sentence [i.e., *Mirjam* in (7b) and (7d), and *Thomas* in (7c)]. Relevantly, even if pronouns could have co-reference with an antecedent outside of the sentence, the availability of an antecedent in the same sentence (i.e., *Mirjam/Thomas*) guarantees that the pronoun-antecedent relation is resolved within the sentence. Feminine and masculine pronouns and referential expressions were counter-balanced. Previous reading time studies found effects at positions immediately following the antecedent (see Yoshida et al., 2014). Based on this, we included proper names with a surname (such as *Lodewijk Boer*) in our data to ensure that there could be a region immediately following the proper name that was still connected to the antecedent position. However, considering the superior time accuracy of the ERP technique, our prediction was that the effect should be observable at the target position rather than at immediately following regions. Participants read 36 target stimuli such as those in (7; see Data Sheet in Supplementary Material for a whole list of stimuli) randomly interspersed with 35 unrelated fillers that were part of a different experiment that examined the processing of backward negative polarity item dependencies (Pablos et al., 2012).

(7) **a. No-Constraint/Match**
Zijn$_j$ assistenten kwamen erachter dat **Lodewijk$_j$ Boer** geen prijswinnaar *His assistants found out that Lodewijk$_{masc}$ Boer no prizewinner* geselecteerd had, maar **Mirjam$_i$** had geen interesse in de roddel.
selected had but Mirjam$_{fem}$ had no interest in the gossip 'His assistants found out that Lodewijk Boer had not selected a prizewinner, but Mirjam had no interest in the gossip.'
(b) No-Constraint/Mismatch
Haar$_i$ assistenten kwamen erachter dat **Lodewijk$_j$ Boer** geen prijswinnaar *Her assistants found out that Lodewijk$_{masc}$ Boer no prizewinner* geselecteerd had, maar **Mirjam$_i$** had geen interesse in de roddel. *selected had, but Mirjam$_{fem}$ had no interest in the gossip.*

'Her assistants found out that Lodewijk Boer had not selected a prizewinner, but Mirjam had no interest in the gossip.'

c. Principle C/Match

Hij$_i$ kwam erachter dat **Lodewijk$_j$ Boer** geen prijswinnaar
He found out that Lodewijk$_{masc}$ Boer no prize winner
geselecteerd had, maar **Thomas$_i$** had geen interesse in de roddel.
selected had, but Thomas$_{masc}$ had no interest in the gossip.
'He found out that Lodewijk Boer had not selected a prizewinner, but Thomas had no interest in the gossip.'

d. Principle C/Mismatch

Zij$_i$ kwam erachter dat **Lodewijk$_j$ Boer** geen prijswinnaar
She found out that Lodewijk$_{masc}$ Boer no prize winner
geselecteerd had, maar **Mirjam$_i$** had geen interesse in de roddel.
selected had, but Mirjam$_{fem}$ had no interest in the gossip.
'She found out that Lodewijk Boer had not selected a prizewinner, but Mirjam had no interest in the gossip.'

Participants

Twenty-four students of Leiden University participated in this study, which was conducted at the EEG Laboratory in the Faculty of Social Sciences of Leiden University. They were all native speakers of Dutch. All participants had normal or corrected-to-normal vision, were right-handed, gave informed consent and were paid €12.50 for their participation, which lasted around 30 min, excluding set-up time. The experiment followed the Ethics Committee regulations of the Faculty of Social Sciences of Leiden University, which approved its implementation.

Procedure

Participants were comfortably seated in a dimly lit testing room around 100 cm in front of a computer monitor. Sentences were presented one word at a time in black letters on a white screen using the presentation software E-prime (Psychology Software Tools Inc.). Each sentence was preceded by a fixation cross ("+") which appeared at the center of the screen and remained there for 1000 ms. The fixation point was followed by a blank screen interval of 300 ms, and then the sentence was displayed word by word.

Each word appeared on the screen for 300 ms, followed by a fixation cross ("+") at the center of the screen that remained visible for 300 ms. Participants were instructed to read the sentences carefully for comprehension. The last word of each sentence was marked with a period, and 1000 ms later a comprehension question appeared and prompted the participant to press a button to continue. Every experimental item was followed by a comprehension question. The comprehension questions targeted different positions of the sentence and some of them targeted the referential expressions *Lodewijk Boer* or *Thomas/Mirjam*. The comprehension questions were counter-balanced for yes and no answers and, for some items, they differed across conditions (see Data Sheet in Supplementary Material). Four counterbalanced lists derived from a Latin Square Design were used for the experiment. Before starting the experimental phase, eight warm-up practice trials were presented to the participants, which had no similarity to any of the targets or filler items in the experiment. Participants were able to ask clarification questions to the experimenter about the task at the practice time. The experimental session was broken up by two break periods, with a different number of items distributed across each block, with 35 and 36 sentences per block.

EEG Recording

The EEG signal was continuously acquired at a sampling frequency of 512 Hz using a BioSemi (Active Two) system from 32 Ag/AgC1 electrodes distributed in the scalp following the extended 10–20 convention (Fp1/2, FC5/, AF3/4, Fz, CP5/6, CP1/2, Cz, F7/8, F3/4, T7/8, C3/4, Pz, FC1/2, P3/4, O1/2, Oz, P7/8, PO3/4). EEG data was referenced on-line to two auxiliary electrodes: common mode sense (CMS) and driven right leg (DRL) and re-referenced off-line to the mean activity at the two mastoids. A high-pass filter with a cut-off frequency of 0.1 Hz was applied online to eliminate DC drifts. Vertical and horizontal eye movements were monitored with two electrodes at the infraorbital and supraorbital, and electrodes at the outer canthus of the right and left eyes. Electrode impedances were monitored during installation to ensure a low level of electronic noise.

EEG Analysis

For every subject, recorded EEG waveforms were post-processed before analysis to reduce noise and artifacts as much as possible. After applying a high-pass filter to remove slow drifts and DC offsets, ocular correction was performed using an implementation of the Gratton et al. (1983) algorithm. Other artifacts were removed both by visual inspection and by performing an automated detection based on gradient change rate. The process resulted in the rejection of 6% of the trials (51 out of 864) distributed among the experimental conditions as follows: (7a) 1%; (7b) 1%; (7c) 2%; (7d) 2%. To confirm that these small differences between conditions were not significant and did not introduce biases in the results, we ran a repeated measures mixed-logit analysis with Match (match/mismatch) and Constraint (No Constraint/Principle C) as independent variable and Subject as random factor. Both main effects and interactions were considered, and no significant difference in likelihood ratio between the fitted model and a null intercept only model was observed.

As a final step, a low-pass filter with a cut-off frequency of 30 Hz was applied to remove noise and non-neurological signals. After the data cleaning, a few electrodes identified as noisy or with intermittent connection were replaced by an interpolation based on neighboring channel responses.

Electroencephalography recordings were then segmented from 200 ms before to 800 ms after the onset of the significant region being analyzed (*Lodewijk*). A baseline correction was applied based on the average of the 200 ms prior to the stimulus onset.

Previous studies that have examined gender mismatches consistently reported a P600 component. In order to evaluate the presence of a P600 in our experimental data, the 500–700ms time window was tested by means of a 4-way repeated-measure

ANOVA, considering four within-subject factors. Two to evaluate the signal scalp distribution: *Hemisphere* [Left *(Fp1, F3, F7, C3, P3, O1)* Central *(Fz, Cz, Pz)*, Right *(Fp2, F4, F8, C4, P4, O2)*], and *Position* [Frontal *(Fp1, Fp2, F3, F4, F7, F8)*, Medial *(C3, Cz, C4),* and Parietal *(P3, Pz, P4, O1, O2)*]; and two to examine effects between conditions: *Constraint* (No Constraint/Principle C), and *Match* (Match/Mismatch). Mean voltage-amplitude was considered as the dependent variable in the analysis, and p-values where corrected for sphericity where required.

RESULTS

Comprehension Questions

Average accuracy rates were high and no participants were rejected on the basis of accuracy ($M = 84.59\%$, $SD = 5.44\%$). The accuracy scores were similar across conditions ($M_{\text{NoConstraintMatch}} = 81\%$, $M_{\text{NoConstraintMismatch}} = 84\%$, $M_{\text{PrincipleCMatch}} = 87\%$, $M_{\text{PrincipleCMismatch}} = 86\%$). The difference in mean values was not significant as shown by a 2×2 repeated-measures ANOVA randomized by subjects with *Constraint* and *Match* as independent factors and Response Accuracy as dependent variable ($p > 0.5$ for all main effects and interactions).

Event Related Potentials

We investigated ERPs at the subject position of the embedded clause, *Lodewijk*, which is the first potential antecedent position in the sentence if there is no Principle C violation. Four-way ANOVA performed in the pre-selected P600 time window (500–700 ms) did not result in any significant main effect or interaction ($p \geq 0.1$ in all cases), as shown in the right most column of **Table 1**. However, visual comparison of the grand average time traces in the anterior electrodes for the No-Constraint Mismatch condition (7b) versus No-Constraint Matched (7a) condition shows an apparent sustained negativity in the 200–600 ms region (**Figure 1**). The anterior topography of the negativity can be observed in **Figure 2**. No such negativity is observed for Principle C Match/Mismatch conditions (**Figures 3** and **4**). The asymmetry observed in the No-Constraint with respect to the Principle C conditions supports the expectation of the experimental manipulation, therefore, an exploratory analysis was performed to investigate the reliability and nature of this apparent difference.

An omnibus ANOVA performed in the complete 200–600 ms time window shows a significant 4-way interaction of *Constraint, Match, Hemisphere,* and *Position* [$F(4,92) = 2.572$; $p = 0.043$]. Follow-up simple interaction analysis for each level of the *Constraint* factor reveals no significant interaction or main effect in Principle C conditions, while a significant 3-way interaction between *Hemisphere* × *Match* × *Position* is present in No-Constraint [$F(4,92) = 3.202$, $p = 0.016$]. A further breakdown of this interaction for every level of the *Position* condition shows a significant effect of *Match* factor at the Anterior sites [$F(1,23) = 4.82$, $p = 0.038$], and no dependence on *Hemisphere*. The *No-Constraint Mismatch* condition (7b) waveform average amplitude is more negative than (7a) [t-test nearly significant difference $t'(23) = 1.989$, $p = 0.057$].

The same analysis was repeated using sliding 200 ms long windows to localize the effect with respect to the onset time of the stimuli. **Table 1** summarizes the omnibus ANOVAs and **Table 2** provides the follow up simple interaction evaluation for those regions with significant interaction in the omnibus ANOVA. (Only significant comparisons and effects are shown for readability. Values are corrected for sphericity where required – corrected p-values are reported).

Finally, **Table 3** shows a summary of the main effects and *post hoc* pairwise comparisons observed in the two time windows (200–400 ms, 300–500 ms) in the breakdown of the interactions observed in **Table 2**, which in all cases reflect a significant anterior negativity of the *Mismatch* condition for the *No-Constraint* case when compared with the *matched* counterpart.

However, the results of the exploratory analysis above present the multiple comparison problem (MCP). To limit the Family Wise Error Rate (FWER) to a 5% level, the individual comparisons reported in **Table 1** should have a p-value lower than $0.05/4 = 0.0125$. In addition, an individual 2×2 ANOVA – to verify the interaction of the *Constraint* and *Match* factors in the topographical regions of interest defined by the *Position* and *Hemisphere* factors considered in the above analysis – did not yield a significant interaction in neither of the time windows ($p > 0.10$). This result is very likely due to the low statistical power provided by the small number of electrodes in each region of interest, and the limited number of trials.

To address the problem of MCP and verify if the differences observed were reliable, the ERPs measured were analyzed with a repeated measures two-tailed cluster mass permutation test (Bullmore et al., 1999; Maris and Oostenveld, 2007) using the Matlab Mass Univariate ERP Toolbox (Groppe et al., 2011). This test provides a better spatial and temporal resolution and weak control of the FWER. We included

TABLE 1 | Multiple window ANOVA 4-way interaction results (p-values reported).

	Time window (ms)				P600 time window
	100–300	200–400	300–500	400–600	500–700 ms
Hemisphere × Position × Match × Constraint	0.085	0.037*	0.036*	0.069	0.101

*$p < 0.05$.

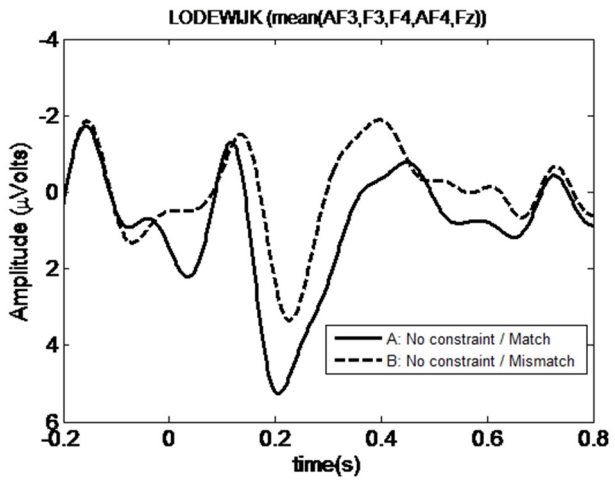


FIGURE 1 | Mean waveform at first potential antecedent *Lodewijk* position for *No-constraint Match* and *Mismatch* conditions.

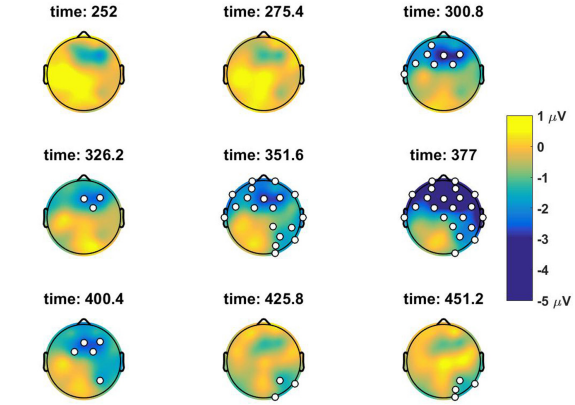


FIGURE 2 | Topographic scalp maps of the difference wave between the *No-Constraint Mismatch* and the *No-Constraint Match* condition at a series of discrete time positions. The electrodes that were significantly different between the two conditions in the cluster mass univariate analysis ($p < 0.07$) are marked in white.

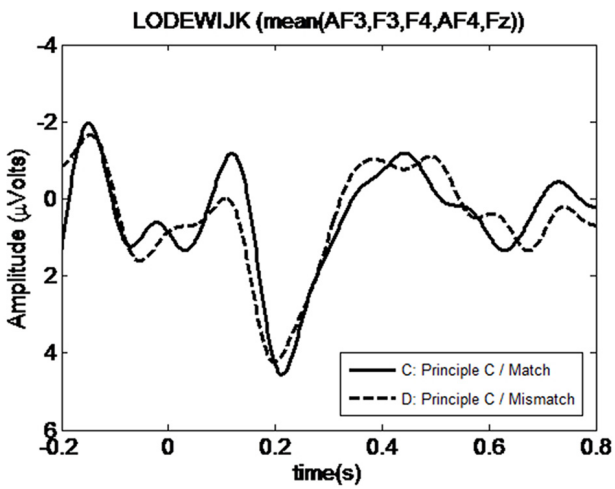


FIGURE 3 | Mean waveform at first potential antecedent *Lodewijk* position for *Principle C Match* and *Mismatch* conditions.

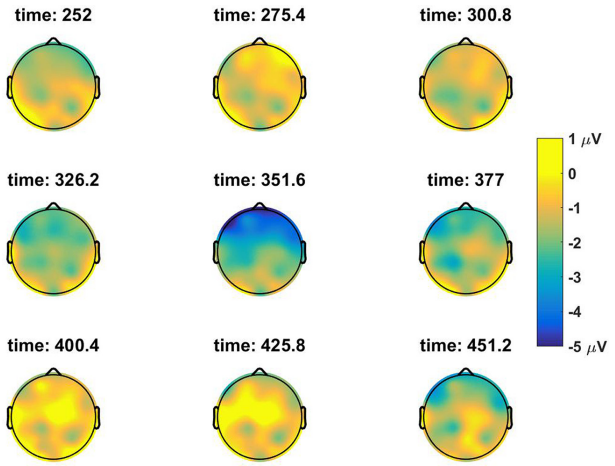


FIGURE 4 | Topographic scalp maps of the difference wave between the *Principle-C Mismatch* and the *Principle-C Match* condition at a series of discrete time positions.

all samples between 200 and 800 ms at all 32 electrodes. Electrodes within an approximate distance of 5.77 cm from each other were considered spatial neighbors for the cluster determination. Repeated measures *t*-tests were performed on the difference wave of the Match and Mismatch conditions for both *No-Constraint* and *Principle C* factor levels. *T*-test included the original data and 2500 random within-subjects permutations. With this technique, we tested separately the null hypothesis that the *Match* and *Mismatch* position do not differ in the No-Constraint and Principle C conditions. The maximum cluster-level mass procedure in the No-Constraint Match versus Mismatch comparison returned a cluster at the central-frontal electrodes extending temporally from 300 to ~420 ms with an alpha level $p = 0.07$ (see **Figure 2**). In contrast, the procedure in the Principle C conditions did not reject the null hypothesis to any level of significance ($p > 0.4$).

In conclusion, results show significant differences to an alpha level of ~0.07 between the *Match* and *Mismatch* conditions in the No-Constraint cases only, with anterior topographic distribution over a window around 300–420 ms. The observed difference is both in the direction expected based on the theoretical predictions, and with a coherent spatial and temporal localization. This reinforces that the effect is reliable even with the aforementioned reduced confidence level, compared to traditional 5% values. The presence of a positive result in an experiment with a relatively low power in terms of the number of trials observed per subject and condition (i.e., 9) suggests that the effect size is large and would be more prominent with an increased number of items [see Maxwell

TABLE 2 | Simple interactions follow-up.

		Time window (ms)	
		200–400	300–500
Principle C	Match	0.428	0.813
	Hemisphere × Match	0.834	0.653
	Position × Match	0.120	0.127
	Match × Hemisphere × Position	0.096	0.288
No-Constraint	Match	0.085	0.185
	Hemisphere × Match	0.129	0.169
	Position × Match	0.071	0.072
	Match × Hemisphere × Position	0.013*	0.018*

*$p < 0.05$.

TABLE 3 | Simple comparisons "No Constraint" condition.

		Time window (ms)	
No constraint		**200–400**	**300–500**
Anterior	Match	0.011* $t(23) = 2.52$, $p = 0.019$	0.025* $t(23) = 2.22$, $p = 0.036$
	Hemisphere × Match	0.061	0.053
Central	Match	0.202	0.333
	Hemisphere × Match	0.409	0.219
Posterior	Match	0.498	0.643
	Hemisphere × Match	0.074	0.109

*$p < 0.05$.

et al. (2008) for a discussion on sample size and statistical power].

DISCUSSION

Active Search for Antecedents

We have shown that, in cases such as (7b) (No-Constraint Mismatch), where there is a gender mismatch between the pronoun and the first potential antecedent for this pronoun, an anterior negativity is generated at the potential antecedent position *Lodewijk*. This is not the case for (7a), where the potential antecedent matches in gender with the preceding pronoun. The anterior negativity could be interpreted as a result of the gender mismatch between a cataphoric pronoun and its antecedent, as well as the effect of failing to find an antecedent at the first potential position. However, for (7c) and (7d), where the cataphoric pronoun cannot corefer with the referential expression *Lodewijk* due to Principle C, no component is generated at the referential expression position. This confirms our predictions that (i) an active search for an antecedent is initiated as soon as a cataphoric pronoun is processed and that, (ii) although the ASM can be automatically initiated for every pronoun, which referential expression will be considered by the ASM is constrained by grammatical principles (in this case, Principle C). This result is in line with the behavioral results (e.g., Kazanina et al., 2007) that found a GMME at the potential antecedent.

Forward vs. Backward Antecedent/Pronoun Dependencies and Prediction Failure

The differences observed in ERP components generated between our results in the case of cataphoric dependencies (anterior negativity) and the forward pronominal dependency studies (Osterhout and Mobley, 1995; Van Berkum et al., 2007; Xu et al., 2013; P600) raise questions on the nature of the effect observed.

In the current experiment, we focus on the relation between a *cataphoric* pronoun and its potential antecedent. In the case of forward antecedent-pronoun dependencies [as in (6)], there is no need to search for a pronoun after encountering the antecedent (e.g., *the aunt*) since this referential expression can be independently interpreted. In other words, we do not expect an active search for a pronoun in the case of forward dependencies. The P600 component in these cases, therefore, must correspond to a gender mismatch between the referential expression and the pronoun.

In backward, cataphoric pronoun-antecedent dependencies, on the other hand, the processes underlying the generation and interpretation of these dependencies are different since the interpretation of the pronoun needs to be resolved. It is therefore reasonable to hypothesize that the parser prefers to start a search as soon as a pronoun is encountered. The anterior negativity in our experiment could be interpreted as related to the searching process itself, namely, a failure of a prediction and not so much to the gender mismatch. The GMME provides the evidence that the antecedent search is active in the no-constraint cases, but it might not be the primary reason for the generation of the anterior negativity. Nevertheless, after having examined previous literature on gender mismatches, we might still wonder why no P600 as well is generated for the gender mismatch at *Lodewijk* in (7b) after encountering the feminine pronoun *haar*. We hypothesize that, in forward dependencies, the parser needs to retrieve the gender of the antecedent from memory and check for gender matching. The P600 could be a reflection of the gender mismatch alone. Conversely, in backward dependencies, the parser anticipates the appearance of an antecedent in the upcoming sentence as soon as it processes the pronoun. Thus, when the parser encounters the first potential antecedent position, it expects to find a matched antecedent. When it fails, there is a negativity generated instead of a P600 because the failure of finding a matching antecedent prevails over the GMME. With this claim we do not intend to imply that the gender mismatch does not occur at all or that it does not precede the expectation failure (since the failure of the prediction cannot occur before the mismatch is detected) rather that the failure of finding a matching antecedent veils the presence of a P600.

In the second experiment in Osterhout and Mobley (1995), a negativity (at anterior and temporal sites in the left-hemisphere between 300 and 500 ms) is found for a dependency where a specific verb form that agrees with the subject is predicted and fails. In our experiment, a negativity is found for a dependency where an antecedent for the cataphoric pronoun is predicted

and this prediction fails because of a gender mismatch. These two types of dependencies are different in nature (one involves subject-verb agreement and the other a pronoun-antecedent coreferential relation), but the mechanism of prediction failure seems to be the same in that there is a negative component generated in both cases. Despite of the fact that the negativities in these two studies are different in distribution, we suggest that they are connected to the same basic process, and that they reflect the failure of a previously established expectation. However, we have to consider that the presence of a negativity in agreement violations is currently under debate since not all the studies observed it (see Nevins et al., 2007; Mancini et al., 2011; Molinaro et al., 2011; a.o.).

Potential Task and Stimuli Presentation Effects

One of the potential sources for the lack of P600 for the gender mismatch in our study might connect to issues that previous studies have discussed (Bornkessel-Schlesewsky et al., 2011; Molinaro et al., 2011; Sassenhagen et al., 2014), such as the influence of task and the modality of stimulus presentation. The current study used word-by-word visual presentation of the sentences in which subjects had to read the sentence and answer a Yes/No comprehension question afterward. Studies that have shown P600 effects for gender mismatches in forward antecedent/pronoun dependencies (Osterhout and Mobley, 1995; Van Berkum et al., 2007; Xu et al., 2013) have all used visual presentation, so the mode of presentation does not seem to have an impact in the results. Differences between our study and previous studies rest in the task that participants were required to complete. Van Berkum et al. (2007) do not require any task from participants besides reading the sentences, whereas Osterhout and Mobley (1995) and Xu et al. (2013) ask their participants to conduct an acceptability judgment after reading each sentence. Sassenhagen et al. (2014) discuss the idea that the generation of a P600 can be task-dependent and that consciously detected violations might differ with respect to non-consciously detected violations in that the detected or attentive violations elicit both an early negative component and a P600, whereas the non-detected ones do not necessarily elicit a P600 (Hasting and Kotz, 2008; Batterink and Neville, 2013). Results from our experiment seem to align with this idea since we only get an early negativity and the study does not implement a task that highlights the mismatch.

Temporal Characteristics and Scalp Distribution of Negativities in Previous ERP Studies

Previous studies that have elicited negativities have looked at agreement mismatches with personal pronouns and subject-verb agreement failures (Osterhout and Mobley, 1995), at noun phrases that ambiguously referred to two equally suitable referents (Van Berkum et al., 2003, 2007), at incorrect cases of noun ellipsis (Martin et al., 2012), at pronoun and verb-agreement violations (Coulson et al., 1998), at verb subcategorization violations (Rösler et al., 1993), at phrase structure violations (Neville et al., 1991; Osterhout and Holcomb, 1992) and at conditions of increased memory load (Kluender and Kutas, 1993; King and Kutas, 1995; Friederici et al., 1996; Müller et al., 1997; Münte et al., 1998; Fiebach et al., 2001).

All the negativities found in these studies reflect syntactic processes and in many cases they represent a response to syntactic violations. However, they do not always have the exact same scalp distribution or topography as the negativity in our study. Osterhout and Mobley (1995) tested agreement mismatches involving personal pronouns in forward dependencies in their first experiment (discussed under the section on ERP Studies on Gender Agreement/Mismatch in the introduction) and found that a small sample of participants ($N = 4$) who judged the sentence as grammatical (and thus considered that there was an antecedent outside the clause for the pronoun) showed a sustained negativity in frontal electrodes in the 500–800 ms. The referentially induced frontal negativity (Nref) elicited by Van Berkum et al. (2003, 2007) was a widely distributed and frontally sustained negativity, emerging at about 300–400 ms after their acoustic onset, whereas Martin et al. (2012)'s negativity had a broad central distribution and emerged between 400 and 1000 ms after word onset. In Coulson et al. (1998), the negativity elicited by ungrammatical pronouns was largest at left anterior sites while that elicited by ungrammatical verbs was centro-parietal and slightly larger over the right hemispheres. This effect was largest between 300 and 500 ms after stimulus onset. ERPs for syntactic violations in Rösler et al. (1993) were negative between 400 and 700 ms after target onset and were more pronounced at anterior sites and over the left hemisphere. In Neville et al. (1991), the phrase structure violations generated a negative response between 300 and 500 ms over temporal and parietal regions of the left hemisphere while in Osterhout and Holcomb, 1992, the negativity occurred between 300 and 500 ms post stimulus at left hemisphere anterior sites.

If we look at the studies with increased memory load, the sustained negativity in Fiebach et al. (2001) started at about 400 ms after the onset of the first prepositional phrase and was maximal at left-anterior electrode positions. Friederici et al. (1996) found a left anterior negativity for the syntactic-category violation condition in auditory and visual tasks in the time windows between 400 and 600 ms (for auditory) and 350 and 500 ms (for visual) after word onset. The ERPs to the verbs in Object relative clause sentences (i.e., *The reporter who the senator harshly attacked admitted the error*) in King and Kutas (1995) showed more prolonged negativity over left anterior regions of the scalp than those in Subject relative clause sentences (i.e., *The reporter who harshly attacked the senator admitted the error*), and in Kluender and Kutas (1993), a difference was seen in the ERP between 300 and 500 ms. post stimulus when wh-questions were compared to yes/no questions at a position early in the matrix clause. Finally, in Müller et al. (1997), there was a large fronto-central negativity beginning at the gap in the Object relative clause sentences and a left frontal negativity in Münte et al. (1998).

Referential Dependencies that Generated Negativities in Previous ERP Studies

Among the ERP studies that have generated negativities, Martin et al. (2012) report a centrally distributed negativity at a position that renders a gender-mismatch effect [i.e., the determiner *otro* 'another (MASC)'], which mismatches in gender with the antecedent *camiseta* ['t-shirt (FEM)'] in cases of noun ellipsis in coordinated sentences. In their study, the gender mismatch results in an ungrammatical sentence (its interpretation cannot be recovered, unlike in (7b) in our study where a second potential antecedent *Mirjam* can be used to resolve the interpretation of the pronoun *haar*) and the position in which the mismatch is detected is a determiner that allows nominal ellipsis within the second coordinated sentence. Both Martin et al. (2012) and our study examine the resolution of dependencies where a referential entity and an antecedent are involved and both concern gender mismatches. However, similar to the first experiment on the study in Osterhout and Mobley (1995) on forward pronominal dependencies, in Martin et al.'s (2012) study, the interpretation of a determiner that allows nominal ellipsis and whose antecedent sits in the previous coordinated clause might involve a completely different process from the process required in the dependencies examined within the current study, since the antecedent does not necessarily start a search for the determiner in the second conjunct.

A sustained negativity (largest at anterior sites) has additionally been found in cases of referential ambiguity under the name of referentially induced frontal negativity (Nref; Van Berkum et al., 2003, 2007), where participants had to choose among a set of equally plausible referents for a specific noun phrase. The fact that Van Berkum et al. (2003, 2007) and our study both cover the processing of dependencies that involve referential expressions, might have contributed to the overlapping characteristics of the ERP components that were found.

In short, we have argued that the anterior negativity in this study can be connected to negativities found in previous studies in that it involves (1) a gender mismatch; (2) a dependency that contains referential expressions in which coreference needs to be established, and (3) a dependency in which an expectation of the parser fails. Thus, even if the studies discussed thus far have looked at different phenomena, it seems that there are some common processes underlying all these negativities, such as building a referential dependency on-line and predicting a specific upcoming element in the sentence.

CONCLUSION

In our ERP study on the processing of cataphoric pronoun dependencies in Dutch, we replicated earlier behavioral findings (Sturt, 2003; Van Gompel and Liversedge, 2003; Kazanina et al., 2007; Yoshida et al., 2014) supporting that the parser actively looks for an antecedent for a cataphoric pronoun in the upcoming sentence (even when this pronoun could have coreference with an antecedent outside of the sentence), but restricts its choice to grammatically licit positions. This is evidenced by the fact that no ERP effect is elicited at the potentially mismatched referential expression in the conditions where Principle C of the Binding Theory bars coreference. The overall results show that the GMME connected to longer reading times in previous behavioral experiments is reflected in the current ERP study as an anterior negativity elicited at the potential antecedent in cataphoric dependencies. We postulate that this anterior negativity reflects the prediction failure for an appropriate antecedent after encountering a sentence initial pronoun.

ACKNOWLEDGMENTS

We would like to thank Niels O. Schiller for his advice and for discussion of the materials presented here. We are grateful to Nina Kazanina for sharing experimental materials, for discussion and for helpful comments on earlier versions of this paper. We thank Masaya Yoshida for useful discussion of the research presented here. We would like to thank Bastien Boutonnet for his assistance regarding the statistical analyses applied to the EEG data and, at an earlier stage, Guido Band and Kalinka Timmer for assistance in conducting the experiment at the EEG laboratory at the Faculty of Social Sciences (FSW) in Leiden University. Finally, we would like to thank the two reviewers for very helpful comments and suggestions and for fruitful discussions on the content of the article. Earlier versions of this work were presented at the Architectures and Mechanism of Language Processing Conference (AMLaP) in 2011, at the CUNY Human Sentence Processing Conference in 2012, at GLOW 35 in 2012, and at CNS in 2014. We thank the audiences of these conferences for their input.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpsyg.2015.01638

REFERENCES

Barber, H., and Carreiras, M. (2005). Grammatical gender and number in Spanish: an ERP comparison. *J. Cogn. Neurosci.* 17, 137–153. doi: 10.1162/0898929052880101

Batterink, L., and Neville, H. J. (2013). The human brain processes syntax in the absence of conscious awareness. *J. Neurosci.* 33, 8528–8533. doi: 10.1523/JNEUROSCI.0618-13.2013

Bornkessel-Schlesewsky, I., Kretzschmar, F., Tune, S., Wang, L., Genç, S., Philipp, M., et al. (2011). Think globally: cross-linguistic variation in electrophysiological activity during sentence comprehension. *Brain Lang.* 117, 133–152. doi: 10.1016/j.bandl.2010.09.010

Bullmore, E. T., Suckling, J., Overmeyer, S., Rabe-Hesketh, S., Taylor, E., and Brammer, M. J. (1999). Global, voxel, and cluster tests, by theory and permutation, for a difference between two groups of structural MR images of the brain. *IEEE Trans. Med. Imaging* 18, 32–42. doi: 10.1109/42.750253

Chomsky, N. (1981). *Lectures on Government and Binding*. Dordrecht: Foris.

Chow, W. Y., Lewis, S., and Phillips, C. (2014). Immediate sensitivity to structural constraints in pronoun resolution. *Front. Psychol.* 5:630. doi: 10.3389/fpsyg.2014.00630

Clifton, C., and Frazier, L. (1989). "Comprehending sentences with long distance dependencies," in *Linguistic Structure in Language Processing*, eds G. N. Carlson and M. Tanenhaus (Boston: Kluwer Academic Publishers), 273–317. doi: 10.1007/978-94-009-2729-2_8

Coulson, S., King, J. W., and Kutas, M. (1998). Expect the unexpected: event-related brain response to morphosyntactic violations. *Lang. Cogn. Process.* 13, 21–58. doi: 10.1080/016909698386582

Crain, S., and Fodor, J. (1985). "How can grammars help parsers?," in *Natural Language Parsing. Psychological, Computational, and Theoretical Perspectives*, eds D. Dowty, L. Karttunen, and A. M. Zwicky (Cambridge: Cambridge University Press), 94–128.

Demestre, J., Meltzer, S., Garcia-Albea, J. E., and Vigil, A. (1999). Identifying the null subject: evidence from event-related brain potentials. *J. Psychol. Res.* 28, 293–312. doi: 10.1023/A:1023258215604

Deutsch, A., and Bentin, S. (2001). Syntactic and semantic factors in processing gender agreement in Hebrew: evidence from ERPs and eye movements. *J. Mem. Lang.* 45, 200–224. doi: 10.1006/jmla.2000.2768

Fiebach, C. J., Schlesewsky, M., and Friederici, A. D. (2001). Syntactic working memory and the establishment of filler-gap dependencies: insights from ERPs and fMRI. *J. Psychol. Res.* 30, 321–338. doi: 10.1023/A:1010447102554

Frazier, L., and Clifton, C. (1989). Successive cyclicity in the grammar and the parser. *Lang. Cogn. Process.* 4, 93–126. doi: 10.1080/01690968908406359

Friederici, A. D., Mecklinger, A., and Hahne, A. (1996). The temporal structure of syntactic parsing: early and late event-related brain potential effects elicited by syntactic anomalies. *J. Exp. Psychol. Learn. Mem. Cogn.* 22, 1219–1248. doi: 10.1037/0278-7393.22.5.1219

Gratton, G., Coles, M. G. H., and Donchin, E. (1983). A new method for off-line removal of ocular artifact. *Electroencephalogr. Clin. Neurophysiol.* 55, 468–484. doi: 10.1016/0013-4694(83)90135-9

Groppe, D. M., Urbach, T. P., and Kutas, M. (2011). Mass univariate analysis of event-related brain potentials/fields I: a critical tutorial review. *Psychophysiology* 48, 1711–1725. doi: 10.1111/j.1469-8986.2011.01273.x

Gunter, T. C., Friederici, A. D., and Schriefers, H. (2000). Syntactic gender and semantic expectancy: ERPs reveal early autonomy and late interaction. *J. Cogn. Neurosci.* 12, 556–568. doi: 10.1162/089892900562336

Hagoort, P., and Brown, C. M. (1999). Gender electrified: ERP evidence on the syntactic nature of gender processing. *J. Psychol. Res.* 28, 715–728. doi: 10.1023/A:1023277213129

Hasting, A. S. A., and Kotz, S. A. S. (2008). Speeding up syntax: on the relative timing and automaticity of local phrase structure and morphosyntactic processing as reflected in event-related brain potentials. *J. Cogn. Neurosci.* 20, 1207–1219. doi: 10.1162/jocn.2008.20083

Kaan, E. (2007). Event-related potentials and language processing. A brief introduction. *Lang. Linguist. Compass* 1, 571–591. doi: 10.1111/j.1749-818X.2007.00037.x

Kazanina, N., Lau, E. F., Lieberman, M., Yoshida, M., and Phillips, C. (2007). The effect of syntactic contraints on the processing of backwards anaphora. *J. Mem. Lang.* 56, 384–409. doi: 10.1080/17470210902974120

King, J. W., and Kutas, M. (1995). Who did what and when? using word- and clause-level erps to monitor working memory usage in reading. *J. Cogn. Neurosci.* 7, 376–395. doi: 10.1162/jocn.1995.7.3.376

Kluender, R., and Kutas, M. (1993). Bridging the gap: evidence from ERPs on the processing of unbounded dependencies. *J. Cogn. Neurosci.* 5, 196–214. doi: 10.1162/jocn.1993.5.2.196

Ledoux, K., Gordon, P. C., Camblin, C. C., and Swaab, T. Y. (2007). Coreference and lexical repetition: mechanisms of discourse integration. *Mem. Cogn.* 35, 801–815. doi: 10.3758/BF03193316

Lee, M.-W. (2004). Another look at the role of empty categories in sentence processing (and grammar). *J. Psychol. Res.* 33, 51–73. doi: 10.1023/B:JOPR.0000010514.50468.30

Luck, S. (2005). "Ten simple rules for designing and interpreting ERP experiments," in *Event-Related Potentials: A Methods Handbook*, ed. T. C. Handy (Cambridge, MA: MIT Press), 17–32.

Mancini, S., Molinaro, N., Rizzi, L., and Carreiras, M. (2011). When persons disagree: an ERP study of Unagreement in Spanish. *Psychophysiology* 48, 1361–1371. doi: 10.1111/j.1469-8986.2011.01212.x

Maris, E., and Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *J. Neurosci. Methods* 164, 177–190. doi: 10.1016/j.jneumeth.2007.03.024

Martin, A. E., Nieuwland, M. S., and Carreiras, M. (2012). Event-related brain potentials index cue-based retrieval interference during sentence comprehension. *Neuroimage* 59, 1859–1869. doi: 10.1016/j.neuroimage.2011.08.057

Maxwell, S. E., Kelley, K., and Rausch, J. R. (2008). Sample-size planning for statistical power and accuracy in parameter estimation. *Annu. Rev. Psychol.* 59, 537–563. doi: 10.1146/annurev.psych.59.103006.093735

Molinaro, N., Barber, H., and Carreiras, M. (2011). Grammatical agreement processing in reading: ERP findings and future directions. *Cortex* 47, 908–930. doi: 10.1016/j.cortex.2011.02.019

Molinaro, N., Vespignani, F., and Job, R. (2008). A deeper reanalysis of a superficial feature: an ERP study on agreement violations. *Brain Res.* 1228, 161–176. doi: 10.1016/j.brainres.2008.06.064

Müller, H. M., King, J. W., and Kutas, M. (1997). Event-related potentials elicited by spoken relative clauses. *Cogn. Brain Res.* 5,193–203. doi: 10.1016/S0926-6410(96)00070-5

Münte, T. F., Schiltz, K., and Kutas, M. (1998). When temporal terms belie conceptual order. *Nature* 395, 71–73. doi: 10.1038/25731

Neville, H. J., Nicol, J. L., Barss, A., Forster, K. I., and Garrett, M. F. (1991). Syntactically based sentence processing classes: evidence from event-related brain potentials. *J. Cogn. Neurosci.* 3, 151–165. doi: 10.1162/jocn.1991.3.2.151

Nevins, A., Dillon, B., Malhotra, S., and Phillips, C. (2007). The role of feature-number and feature-type in processing Hindi verb agreement violations. *Brain Res.* 1164, 81–94. doi: 10.1016/j.brainres.2007.05.058

Osterhout, L., and Holcomb, P. J. (1992). Event-related brain potentials elicited by syntactic anomaly. *J. Mem. Lang.* 31, 785–806. doi: 10.1016/0749-596X(92)90039-Z

Osterhout, L., and Mobley, L. A. (1995). Event-related brain potentials elicited by failure to agree. *J. Mem. Lang.* 34, 739–773. doi: 10.1006/jmla.1995.1033

Pablos, L., Doetjes, J., Ruijgrok, B., and Cheng, L. L. (2012). "Backward NPI dependencies in Dutch: an ERP investigation," *Poster Presented at the 25th Annual CUNY Sentence Processing Conference*, New York, NY.

Rösler, F., Piitz, P., Friederici, A., and Hahne, A. (1993). Event-Related brain potentials while encountering semantic and syntactic constraint violations. *J. Cogn. Neurosci.* 5, 345–362. doi: 10.1162/jocn.1993.5.3.345

Sassenhagen, J., Schlesewsky, M., and Bornkessel-Schlesewsky, I. (2014). The P600-as-P3 hypothesis revisited: single-trial analyses reveal that the late EEG positivity following linguistically deviant material is reaction time aligned. *Brain Lang.* 137, 29–39. doi: 10.1016/j.bandl.2014.07.010

Stowe, L. (1986). Parsing wh-constructions: evidence for on-line gap location. *Lang. Cogn. Process.* 1, 227–245. doi: 10.1080/01690968608407062

Sturt, P. (2003). The time-course of the application of binding constraints in reference resolution. *J. Mem. Lang.* 48, 542–562. doi: 10.1016/S0749-596X(02)00536-3

Swaab, T. Y., Camblin, C. C., and Gordon, P. C. (2004). Reversed lexical repetition effects in language processing. *J. Cogn. Neurosci.* 16, 715–726. doi: 10.1162/089892904970744

Van Berkum, J., Brown, C. M., Hagoort, P., and Zwitserlood, P. (2003). Event-related brain potentials reflect discourse referential ambiguity in spoken language comprehension. *Psychophysiology* 40, 235–248. doi: 10.1111/1469-8986.00025

Van Berkum, J. J. A., Brown, C. M., Zwitserlood, P., Kooijman, V., and Hagoort, P. (2005). Anticipating upcoming words in discourse: evidence from ERPs and reading times. *J. Exp. Psychol. Learn. Mem. Cogn.* 31, 443–467. doi: 10.1037/0278-7393.31.3.443

Van Berkum, J. J. A., Koornneef, A. W., Ottena, M., and Nieuwland, M. S. (2007). Establishing reference in language comprehension: an electrophysiological perspective. *Brain Res.* 1146, 158–171. doi: 10.1016/j.brainres.2006.06.091

Van Gompel, R. P. G., and Liversedge, S. P. (2003). The influence of morphological information on cataphoric pronoun assignment. *J. Exp. Psychol. Learn. Mem. Cogn.* 29, 128–139. doi: 10.1037/0278-7393.29.1.128

Wicha, N. Y. Y., Moreno, E. M., and Kutas, M. (2004). Anticipating words and their gender: an event-related brain potential study of semantic integration, gender expectancy, and gender agreement in Spanish sentence reading. *J. Cogn. Neurosci.* 16, 1272–1288. doi: 10.1162/0898929041920487

Xu, X., Jiang, X., and Zhou, X. (2013). Processing biological gender and number information during Chinese pronoun resolution: ERP evidence for functional differentiation. *Brain Cogn.* 81, 223–236. doi: 10.1016/j.bandc.2012.11.002

Yoshida, M., Kazanina, N., Pablos, L., and Sturt, P. (2014). On the origin of islands. *Lang. Cogn. Neurosci.* 29, 761–770. doi: 10.1080/01690965.2013.788196

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2015 Pablos, Doetjes, Ruijgrok and Cheng. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 08 September 2015
doi: 10.3389/fpsyg.2015.01346

Interference in the processing of adjunct control

Dan Parker[1], Sol Lago[2] and Colin Phillips[2,3]*

[1] Linguistics Program, Department of English, College of William and Mary, Williamsburg, VA, USA, [2] Department of Linguistics, University of Maryland, College Park, MD, USA, [3] Language Science Center, University of Maryland, College Park, MD, USA

Recent research on the memory operations used in language comprehension has revealed a selective profile of interference effects during memory retrieval. Dependencies such as subject–verb agreement show strong facilitatory interference effects from structurally inappropriate but feature-matching distractors, leading to illusions of grammaticality (Pearlmutter et al., 1999; Wagers et al., 2009; Dillon et al., 2013). In contrast, dependencies involving reflexive anaphors are generally immune to interference effects (Sturt, 2003; Xiang et al., 2009; Dillon et al., 2013). This contrast has led to the proposal that all anaphors that are subject to structural constraints are immune to facilitatory interference. Here we use an animacy manipulation to examine whether adjunct control dependencies, which involve an interpreted anaphoric relation between a null subject and its licensor, are also immune to facilitatory interference effects. Our results show reliable facilitatory interference in the processing of adjunct control dependencies, which challenges the generalization that anaphoric dependencies as a class are immune to such effects. To account for the contrast between adjunct control and reflexive dependencies, we suggest that variability within anaphora could reflect either an inherent primacy of animacy cues in retrieval processes, or differential degrees of match between potential licensors and the retrieval probe.

Keywords: adjunct control, anaphora, agreement, sentence processing, memory retrieval

OPEN ACCESS

Edited by:
Tamara Swaab,
University of California, Davis, USA

Reviewed by:
Edward Matthew Husband,
University of Oxford, UK
Patrick Sturt,
University of Edinburgh, UK

****Correspondence:***
Dan Parker,
Linguistics Program, Department of English, College of William and Mary, P.O. Box 8795, Williamsburg, VA 23187, USA
dparker@wm.edu

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: *20 May 2015*
Accepted: *21 August 2015*
Published: *08 September 2015*

Citation:
Parker D, Lago S and Phillips C (2015) Interference in the processing of adjunct control. Front. Psychol. 6:1346. doi: 10.3389/fpsyg.2015.01346

Introduction

Linguistic dependencies are subject to diverse structural and morphological constraints. Recent studies have examined how these constraints are applied in real-time comprehension in order to gain a better understanding of how we mentally encode and navigate linguistic representations. A comparison of the findings across studies shows a mixed profile of successes and failures of real-time constraint application: some constraints on dependency formation are accurately applied, whereas others are susceptible to errors. The reasons for these failures remain poorly understood, but the mixed profile of constraint application has been argued to reflect the way in which different linguistic processes engage memory retrieval mechanisms (Lewis and Vasishth, 2005; Vasishth et al., 2008; Wagers et al., 2009; Phillips et al., 2011; Lewis and Phillips, 2015).

In this paper, we focus on a specific type of memory retrieval error that leads to an effect called 'facilitatory interference' (also known as 'intrusion' or 'attraction'). Facilitatory interference arises when a structurally inappropriate but feature matching item facilitates the processing of an ill-formed linguistic dependency. This eased processing can trigger 'illusions of grammaticality,' which have been argued to reflect limitations of the memory retrieval mechanisms used to implement

linguistic constraints (Vasishth et al., 2008; Wagers, 2008). Such effects have been reported for subject–verb agreement and negative polarity item processing (Clifton et al., 1999; Pearlmutter et al., 1999; Drenhaus et al., 2005; Vasishth et al., 2008; Staub, 2009, 2010; Wagers et al., 2009; Xiang et al., 2009; Dillon et al., 2013; Tanner et al., 2014; Tucker et al., 2015). For instance, Wagers et al. (2009) used self-paced reading and speeded acceptability judgments to investigate interference effects in the comprehension of subject–verb agreement dependencies like those in (1). They varied the presence of a plural distractor noun in grammatical and ungrammatical sentences.

(1) a. The key to the *cell(s)* unsurprisingly was rusty from many years of disuse.
b. *The key to the *cell(s)* unsurprisingly were rusty from many years of disuse.

In grammatical sentences like (1a), Wagers et al. (2009) found that the plural number of the structurally inappropriate noun *cells* did not impact acceptability judgments or reading times after the verb, relative to the singular noun condition. However, in ungrammatical sentences like (1b) the presence of the plural distractor *cells* increased rates of acceptance and facilitated reading times after the verb, relative to the no distractor condition, giving rise to an illusion of grammaticality.

The profile of facilitatory interference effects in sentences like (1) provides important evidence about the source of these effects in comprehension. First, immunity to distractors in grammatical sentences suggests that facilitatory interference effects do not reflect misrepresentation of the subject number or the use of "good enough" representations (e.g., Ferreira and Patson, 2007). Misrepresentation of the subject phrase would lead comprehenders to misperceive grammatical sentences like (1a) as ungrammatical, triggering an 'illusion of ungrammaticality,' which rarely occurs. Second, if illusions of grammaticality are not due to problems in the representation of the subject, then facilitatory interference effects might instead be due to properties of the retrieval mechanisms used to resolve linguistic dependencies. For instance, under a view where both structural and morphological constraints guide memory access, facilitatory interference effects could reflect failure to apply structural constraints during retrieval, or they could reflect the outcome of a competition between structural and morphological constraints. Crucially, the finding that comprehenders are not misled by structurally inappropriate items in grammatical sentences provides good evidence that structural constraints are actively used to guide retrieval, and suggests that facilitatory interference reflects competing structural and morphological information[1].

Wagers et al. (2009) argued that facilitatory interference in subject–verb agreement is a consequence of competing constraints. Under their account, encountering the verb *were* in (1b) triggers a retrieval that probes previous items in memory to recover a noun phrase that is both the subject of the sentence and has plural number. In ungrammatical sentences, neither the target nor the distractor is a perfect match to the requirements of the verb, and the competition between the structural and morphological constraints is relatively even: the true subject is in the appropriate structural position, but it is not plural, and the distractor is plural, but it is in a structurally inappropriate position. On a significant portion of trials, the structurally inappropriate distractor is incorrectly retrieved, which facilitates processing of the ungrammatical verb and triggers an illusion of grammaticality. In grammatical sentences, by contrast, there is no competition between the structural and morphological constraints, and the full matching subject is almost always retrieved, as it easily out-competes a non-matching distractor[2].

Facilitatory interference is robust for subject–verb agreement, but not all linguistic dependencies are susceptible to it. For example, Dillon et al. (2013) directly compared the processing of subject–verb agreement and reflexive-antecedent dependencies using closely matched sentences like those in (2).

(2) a. The new executive who oversaw the middle *manager(s)* apparently doubted himself/*themselves on most major decisions.
b. The new executive who oversaw the middle *manager(s)* apparently was/* were dishonest about the company's profits.

Dillon et al. (2013) found that subject–verb agreement was susceptible to facilitatory interference from structurally inappropriate distractors (e.g., *managers*), but reflexive-antecedent dependencies were not. These findings are consistent with a growing number of studies that have concluded that direct object reflexives resist facilitatory interference (Nicol and Swinney, 1989; Clifton et al., 1999; Kennison and Trofe, 2003; Sturt, 2003; Xiang et al., 2009; Clackson et al., 2011; Jäger et al., 2015a). Specifically, these studies have found that structurally inappropriate distractors either do not impact the processing of the direct object reflexives or cause increased processing difficulty. The contrast between subject–verb agreement and reflexives is striking since retrieval for both dependencies targets the same structural position, i.e., the subject of the local clause. These findings are important because they cast doubt upon the claim that all linguistic dependencies are uniformly resolved using an error-prone retrieval mechanism, as suggested in previous research (McElree, 2000; McElree et al., 2003; Lewis and Vasishth, 2005).

The puzzle of why reflexives and subject–verb agreement show contrasting profiles with respect to facilitatory interference remains unresolved. One explanation that is often suggested is that the contrast may reflect differences in the interpretive status of reflexives vs. agreement (see Dillon, 2011, for discussion).

[1]Facilitatory interference cannot be simply a case of proximity concord (Quirk et al., 1985) or local coherence (Tabor et al., 2004), as the effect is also observed when the plural distractor does not intervene between the verb and true subject (cf. Wagers et al., 2009). The effect is also not merely due to dialectal variation, as speakers agree on the unacceptability of sentences like (1b) when they have ample time to make their judgment (see Dillon et al., 2013).

[2]More specifically, the account proposed by Wagers et al. (2009) predicts a bimodal response from the mixture of two distributions: either retrieval recovers the structurally inappropriate distractor, facilitating reading times and triggering an illusion, or the structurally appropriate target, slowing down processing. As a result, the average reading times across trials are reduced for ungrammatical sentences with a feature matching distractor relative to ungrammatical sentences where no matching distractor is present.

Reflexive licensing involves constructing an interpreted anaphoric dependency since the meaning of the reflexive depends on the semantic properties of its antecedent. By contrast, subject–verb agreement licensing might involve a morphological process without interpretive consequences (e.g., Lau et al., 2008; but cf. Patson and Husband, 2015). However, it is unclear why the interpretive status of a dependency should determine its susceptibility to facilitatory interference. One possibility is that all interpreted anaphoric dependencies that are subject to syntactic constraints might engage a more conservative retrieval strategy to avoid misinterpretation and interference from structurally inappropriate items. Under this hypothesis, reflexives and agreement could engage qualitatively different retrieval mechanisms, or use distinct sets of retrieval cues to access the local subject. For example, reflexive licensing might engage the same retrieval mechanism as agreement, but might only use structural retrieval cues, implementing morphological constraints only as a post-retrieval check (Dillon, 2011).

In this paper, we do not solve the problem of why reflexives and subject–verb agreement show differential susceptibility to facilitatory interference effects. Instead, we address a critical part of the puzzle by focusing on the status of anaphors and their reported immunity to such effects. Specifically, we test the hypothesis that all anaphoric dependencies that are subject to structural constraints avoid facilitatory interference during real-time comprehension. Our results challenge this hypothesis by showing that adjunct control dependencies, which involve an interpreted anaphoric relation between a null subject and its licensor, are susceptible to facilitatory interference. We then investigate the source of facilitatory interference in adjunct control dependencies, and conclude with a discussion of why anaphoric dependencies should vary with respect to facilitatory interference effects.

Adjunct Control Dependencies

In this paper, we focus on temporal adjunct control constructions like those in (3), which involve a phonetically null anaphoric subject (represented as Ø)[3]. Like reflexives, null subjects must establish a structural, item-to-item dependency with a licensor to receive an interpretation. Specifically, null subjects in temporal adjunct control structures are licensed by the subject of the immediately higher clause. For instance, in (3a,b), the phonetically null subject of the adjunct clause Ø receives its interpretation from the subject of the immediately higher clause *the little girl*, i.e., it is the little girl who played in the yard.

(3) a. The mother said [that *the little girl* fell asleep (after Ø playing in the yard)].
b. *The little girl* talked to her mother (after Ø playing in the yard).

[3]Our discussion does not rely on whether the missing subject is an empty category, e.g., PRO. See Hornstein (2003) for a discussion of the debate over how to formally represent control clause subjects.

However, there are several differences between reflexives and null subjects that might impact their susceptibility to facilitatory interference. For example, reflexives are licensed by the subject of the local clause, whereas null subjects in temporal adjunct control structures are licensed by the subject of the immediately higher clause. Another difference is that retrieval for reflexive licensing is triggered by an independent anaphoric element, whereas retrieval for null subject licensing is triggered by a gerundive verb preceded by a subordinator (e.g., "*after playing*"). Lastly, unlike reflexives, null subjects do not require overt gender or number agreement with a licensor. Instead, null subject licensing in adjunct control structures has been argued to be subject to an animacy constraint. For example, Kawasaki (1993) reported that adjunct control structures are judged to be more acceptable with animate licensors than inanimate licensors (4a vs. 4b; see Landau, 2001, for supporting judgments). The preference for an animate subject does not appear to be a general property of embedded clauses or a consequence of lexical verb biases, since the acceptability contrast between (4a) and (4b) is neutralized when the licensors are the overt subjects of the verb, as in (5).

(4) a. The doctor was certified after Ø debunking the hypothesis.
b. The discovery was certified after Ø debunking the hypothesis.

(5) a. The journalist was surprised that the doctor debunked the hypothesis.
b. The journalist was surprised that the discovery debunked the hypothesis.

The current study contributes to a growing body of research on the processing of control (e.g., Kwon and Sturt, 2014; Sturt and Kwon, 2015) by using the animacy preference for adjunct control structures to probe for facilitatory interference during real-time dependency formation. Animacy features are promising candidates to test for interference effects, as they have been shown to be used in memory retrieval during processing of various linguistic dependencies, including thematic binding and reflexive licensing (e.g., Van Dyke and Lewis, 2003, 2007; Van Dyke and McElree, 2006, 2011; Jäger et al., 2015b). Specifically, we contrast two hypotheses about the nature of retrieval for anaphoric dependencies. Under a view that posits that all anaphoric dependencies are immune to facilitatory interference, null subjects in temporal adjunct control structures should pattern like reflexives and show no susceptibility to facilitatory interference during retrieval for a licensor. In contrast, if anaphoric dependencies do not behave homogenously, then null subject licensing might show facilitatory interference effects similar to those observed for subject–verb agreement.

We report the results from three experiments. In Experiment 1 (untimed acceptability ratings) we confirmed the animacy constraint on null subject licensing. In Experiments 2 and 3 (self-paced reading), we directly compared the comprehension of null subjects and subject–verb agreement, and found that null subjects show a facilitatory interference profile that is qualitatively similar to the profile observed for agreement. These results imply that not all anaphoric dependencies resist facilitatory interference,

and suggest that differences in interpretative status cannot be uniquely responsible for the contrasting interference profiles reported for agreement and reflexives in previous studies.

Experiment 1

Experiment 1 used untimed acceptability ratings to confirm that temporal adjunct control sentences are more acceptable with animate than inanimate licensors, and that the preference for animate licensors is specific to adjunct control constructions, rather than a general property of embedded clauses or lexical verb biases (Kawasaki, 1993).

Participants

Twenty-four participants were recruited using Amazon's Mechanical Turk web-service[4]. All participants in this and the following experiments provided informed consent. Experiment 1 lasted approximately 10 min, and participants were compensated $2.

Materials

Twenty-four sets of items like those in (4–5) were constructed. Two experimental factors were manipulated: ANIMACY of the main clause subject (animate vs. inanimate) and CONSTRUCTION (adjunct control vs. overt subject). The 24 item sets were distributed across four lists in a Latin Square design. Within each list, the 24 target sentences were combined with 48 filler sentences of similar length and complexity, for a total of 72 sentences. The ratio of grammatical to ungrammatical sentences was 1:1, including the inanimate adjunct control sentences as ungrammatical. The ungrammatical filler sentences involved subject–verb agreement errors, unlicensed verbal morphology, and selectional restriction violations.

Procedure

Sentences were presented using Ibex (Alex Drummond[5]). Participants were instructed to rate the acceptability of the sentences along a 7-point Likert scale ('7' = most acceptable, '1' = least acceptable), according to their perceived acceptability in informal, colloquial speech. Participants could take as much time as needed to rate each sentence, as long as they finished the experiment within the 30 min restriction imposed by the Mechanical Turk session. Each sentence was displayed in its entirety on the screen along with the rating scale. Participants could click boxes to enter their rating or use a numerical keypad. The order of presentation was randomized for each participant.

Data Analysis

Data were analyzed using linear mixed-effects models, with fixed factors for experimental manipulations and their interaction. Models were estimated using the *lme4* package (Bates et al., 2011) in the R software environment (R Development Core Team, 2014). Experimental fixed effects and their interaction were set up using orthogonal contrast coding, and items and participants were crossed as random effects (following Baayen et al., 2008; Bates et al., 2011). To determine whether inclusion of random slopes was necessary, we compared a model that included random by-participant and by-item intercepts with a model that included a fully specified (i.e., maximal) random effects structure with random intercepts and slopes for all random effects and their interaction by-item and by-participant (Baayen et al., 2008; Barr et al., 2013). A log-likelihood ratio test revealed that the maximal model provided a better fit to the data [$\chi^2_{(2)} = 67.36$, $p < 0.001$]. Therefore, we adopted the maximal model. For all statistical analyses reported in this paper, an effect was considered significant if its absolute t-value was greater than 2 (Gelman and Hill, 2007).

Results

The results of Experiment 1 are presented in **Figure 1**. Adjunct control sentences with animate subjects were rated higher than those with inanimate subjects (means: 4.81 inanimate subject vs. 6.09 animate subject). By contrast, sentences with animate and inanimate overt subjects received similar ratings (means: 6.43 inanimate subject vs. 6.40 animate subject). The statistical analysis revealed a main effect of subject ANIMACY ($\hat{\beta} = -0.64$, SE = 0.19, $t = -3.40$), a main effect of CONSTRUCTION ($\hat{\beta} = -0.96$, SE = 0.18, $t = -5.08$), and an interaction between subject ANIMACY and CONSTRUCTION ($\hat{\beta} = -1.25$, SE = 0.30, $t = -4.08$). The interaction was driven by the fact that animacy significantly modulated ratings in the adjunct control conditions ($\hat{\beta} = -1.26$, SE = 0.29, $t = -4.32$), but not in the overt subject conditions ($t < 2$).

Discussion

Experiment 1 confirmed that adjunct control sentences are more acceptable with animate than with inanimate licensors. However, since sentences with inanimate licensors received relatively high ratings, we believe that the animacy constraint should be regarded as a weak constraint for adjunct control, or that it

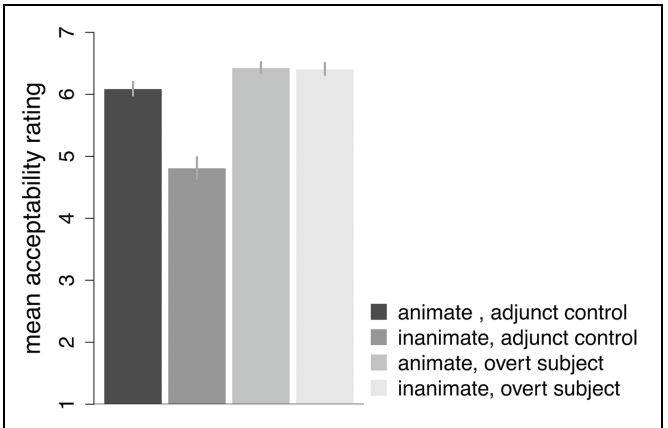


FIGURE 1 | Mean ratings and SE by participants for Experiment 1. Values are on a 7-point Likert scale, with '7' being most acceptable, and '1' the least acceptable. Error bars represent SEM.

[4]https://www.mturk.com

[5]http://spellout.net/ibexfarm

is a constraint that has a smaller impact on ratings because it does not block interpretability. Furthermore, the finding that the animacy manipulation did not impact ratings for sentences with an overt embedded subject implies that the animacy preference for adjunct control cannot simply reflect a general property of embedded clauses or lexical verb biases. Based on these findings, we conclude that comprehenders might use animacy as a cue to guide memory retrieval for null subject licensing, as has been reported for other linguistic dependencies, such as thematic binding (e.g., Van Dyke and Lewis, 2003; Van Dyke and McElree, 2006, 2011).

Experiment 2

The goal of Experiment 2 was to test the hypothesis that all anaphoric dependencies resist facilitatory interference during real-time comprehension. We used self-paced reading to investigate whether retrieval for null subject licensing is susceptible to interference from animate distractors in structurally inappropriate locations. Under the hypothesis that all anaphoric dependencies are immune to facilitatory interference, retrieval for null subject licensing should avoid facilitatory interference from structurally inappropriate animate distractors. Alternatively, if this hypothesis is incorrect, then we might observe facilitatory interference, yielding a profile similar to subject–verb agreement.

Participants

Thirty-two members of the University of Maryland community participated in Experiment 2. Participants were either compensated $10 or received credit in an introductory linguistics course. The self-paced reading task lasted approximately 40 min and was administered as part of a 1-hour session involving unrelated experiments.

Materials

The experimental materials consisted of 48 item sets, each containing eight conditions. The experimental conditions consisted of a 2 × 2 × 2 factorial design, which crossed the factors DEPENDENCY, GRAMMATICALITY, and DISTRACTOR. An example item set is provided in **Table 1**. The first factor, DEPENDENCY, varied the dependency of interest: adjunct control vs. subject–verb agreement. Subject–verb agreement conditions were included to provide an experiment-internal measure of facilitatory interference effects. Within each dependency type, the sentences were maximally similar and differed only in the manipulations of GRAMMATICALITY and DISTRACTOR.

All test items consisted of a passive main clause followed by an adjunct clause. Passive sentences were used because they naturally allow both animate and inanimate NPs in the main clause subject position, and provide a clear attachment site for the adjunct clause to the main clause VP, avoiding the possibility of an attachment ambiguity. In all conditions, the main clause subject was modified by an object relative clause that contained the distractor in subject position. The relative clause verb never overtly expressed agreement, and was always followed by an adverbial that signaled the end of the relative clause. The main clause verb phrase consisted of an auxiliary form of *be* (*was* or *were*) immediately followed by the main verb and an adjunct clause that consisted of a subordinator and gerundive verb.

TABLE 1 | Example set of experimental items for Experiment 2.

Adjunct control conditions
Grammatical, distractor
The doctor that the researcher described meticulously was certified after debunking the urban myth himself in the new scientific journal.
Grammatical, no distractor
The doctor that the report described meticulously was certified after debunking the urban myth himself in the new scientific journal.
Ungrammatical, distractor
The discovery that the researcher described meticulously was certified after debunking the urban myth himself in the new scientific journal.
Ungrammatical, no distractor
The discovery that the report described meticulously was certified after debunking the urban myth himself in the new scientific journal.
Subject–verb agreement conditions
Grammatical, distractor
The doctor that the researcher described meticulously was certified after debunking the urban myth in the new scientific journal.
Grammatical, no distractor
The doctor that the reports described meticulously was certified after debunking the urban myth in the new scientific journal.
Ungrammatical, distractor
The doctor that the researchers described meticulously were certified after debunking the urban myth in the new scientific journal.
Ungrammatical, no distractor
The doctor that the report described meticulously were certified after debunking the urban myth in the new scientific journal.

In the adjunct control conditions, the adjunct clause contained an emphatic reflexive that was licensed by the subject of the adjunct clause, i.e., the null subject. This configuration provided two points to measure susceptibility to facilitatory interference in the adjunct control conditions. The earliest point to measure the impact of the distractor was the gerundive verb. The second point was the emphatic reflexive. Since the reflexive must access the properties of the adjunct clause subject, it was meant to provide a probe of the properties of the licensor retrieved for the anaphoric null subject. In the subject–verb agreement conditions, the earliest point to measure susceptibility to facilitatory interference was the main clause verb.

The factor GRAMMATICALITY was manipulated by varying the animacy of the main clause subject in the adjunct control conditions and the number of the agreeing verb in the subject–verb agreement conditions. In the grammatical adjunct control conditions, the main clause subject was animate and matched the animacy of the reflexive, which satisfied the animacy requirement of the adjunct control structures. In the ungrammatical conditions, the main clause subject did not satisfy the animacy requirement and mismatched the reflexive in animacy. In the grammatical subject–verb agreement conditions, the main clause subject and the agreeing verb were always singular, and thus matched in number. In the ungrammatical conditions, the agreeing verb was plural and mismatched the number of the main clause subject. Lastly, the factor DISTRACTOR was manipulated by varying the animacy of the distractor in the adjunct control

conditions and the number of the distractor in the subject–verb agreement conditions. In order to avoid spurious effects due to lexical differences, the lexical content of the main clause was held constant across dependencies.

Procedure

Sentences were presented on a desktop PC in a moving-window self-paced reading display using Linger (Doug Rohde). Sentences were initially masked by dashes, with white spaces and punctuation intact. Participants pushed the space bar to reveal each word. Presentation was non-cumulative, such that the previous word was replaced with a dash when the next word appeared. Each sentence was followed by a 'yes/no' comprehension question, and onscreen feedback was provided for incorrect answers. The order of presentation was randomized for each participant.

Data Analysis

Only data from participants with at least 70% accuracy on the comprehension questions were used in the analysis. No participants were excluded due to poor accuracy. Reading times greater than 2500 ms were excluded from the analysis (following Hofmeister, 2011; Vasishth and Drenhaus, 2011). This trimming method affected less than 1% of the data. Reading times were then log-transformed to reduce non-normality. For the adjunct control conditions average reading times were compared between conditions in four regions of interest: the subordinator (v−1), the gerundive verb (v), the emphatic reflexive (refl), and the word immediately following the reflexive (refl + 1). For the subject–verb agreement conditions, average reading times were compared between conditions in two regions of interest: the agreeing verb (v) and the main verb (v + 1).

Reading time data were analyzed using linear mixed-effects models. Experimental fixed effects and their interaction were set up using orthogonal contrast coding, and items and participants were crossed as random effects (Baayen et al., 2008). To determine whether inclusion of random slopes was necessary, we compared an intercept-only model to a model with a fully specified random effects structure, which included random intercepts and slopes for all fixed effects and their interaction by items and by participants. A log-likelihood ratio test revealed that the maximal model did not provide a better fit to the data in the critical regions [subject–verb agreement: $\chi^2_{(18)} = 4.39$, $p = 0.92$; adjunct control: $\chi^2_{(18)} = 9.98$, $p = 0.93$]. Therefore, we adopted the intercept-only model, and for consistency, we applied the same model to all regions of interest.

Results

Subject–Verb Agreement Conditions

Figure 2 shows average reading times starting from the region preceding the agreeing verb to five regions beyond the main verb. No effects were observed at the critical verb (v). The word immediately following the critical verb (v + 1) showed a main effect of DISTRACTOR ($\hat{\beta} = 0.06$, SE = 0.02, $t = -2.22$) and crucially, an interaction between GRAMMATICALITY and DISTRACTOR ($\hat{\beta} = -0.17$, SE = 0.05, $t = -2.96$). This interaction was driven by a significant effect of DISTRACTOR in the ungrammatical conditions ($\hat{\beta} = -0.15$, SE = 0.04, $t = -3.56$), reflecting faster reading times for sentences with a plural distractor, relative to sentences with no distractor. No such difference was observed in the grammatical conditions ($t < 2$).

Adjunct Control Conditions

Figure 3 shows average reading times starting from the subordinator to three regions following the reflexive. No effects were observed at the subordinator region (v−1). At the gerundive verb (v), there was an interaction between GRAMMATICALITY and DISTRACTOR ($\hat{\beta} = -0.11$, SE = 0.04, $t = -2.48$). This interaction was driven by a significant effect of DISTRACTOR in the ungrammatical conditions ($\hat{\beta} = -0.07$, SE = 0.03, $t = -2.03$), reflecting faster reading times for sentences with an animate distractor relative to sentences with an inanimate distractor. No such difference was observed for the grammatical conditions ($t < 2$). No effects were observed at the reflexive (refl). The word immediately following the reflexive (refl + 1) showed a main effect of GRAMMATICALITY ($\hat{\beta} = -0.06$, SE = 0.02, $t = -3.02$) and an interaction between GRAMMATICALITY and DISTRACTOR ($\hat{\beta} = -0.10$, SE = 0.04, $t = -2.23$). The main effect of GRAMMATICALITY was due to slower reading times in the ungrammatical conditions relative to the grammatical conditions. The interaction was driven by a significant effect

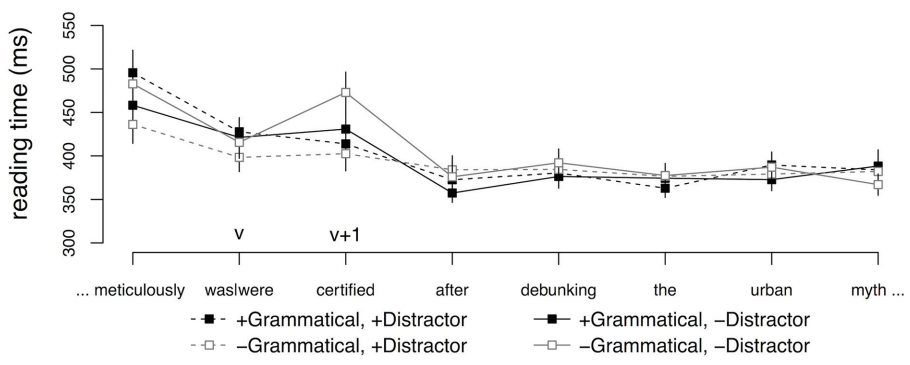


FIGURE 2 | Word-by-word reading times for subject–verb agreement conditions, Experiment 2. Error bars indicate SEM.

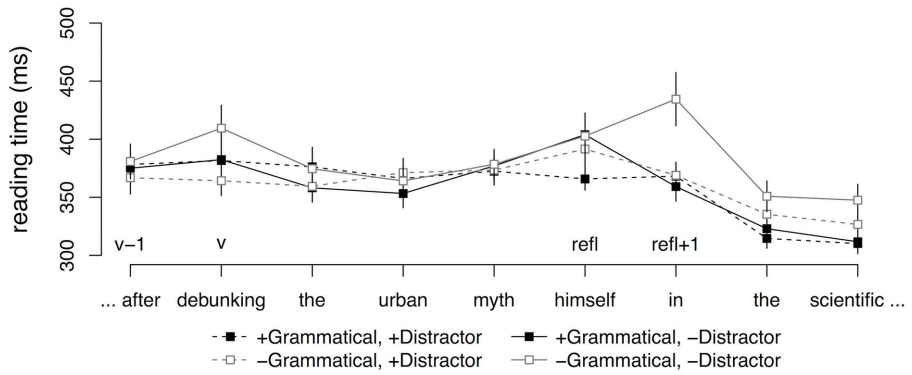


FIGURE 3 | Word-by-word reading times for adjunct control conditions, Experiment 2. Error bars indicate SEM.

of DISTRACTOR in the ungrammatical conditions ($\hat{\beta} = -0.07$, SE $= 0.03$, $t = -2.02$), reflecting faster reading times for sentences with an animate distractor relative to sentences with no distractor. No such difference was observed for the grammatical conditions ($t < 2$).

Discussion

Experiment 2 tested the hypothesis that immunity to facilitatory interference is a general property of anaphoric dependencies. Our results provide evidence against this hypothesis, since they show that adjunct control dependencies, which involve an anaphoric relation between a null subject and its licensor, are susceptible to facilitatory interference, similarly to subject–verb agreement. Facilitatory interference was observed at two different points in the adjunct control sentences. The first was at the gerundive verb, which was the earliest point where sensitivity to the structurally inappropriate distractor could be detected. At this region, reading times for ungrammatical sentences were facilitated by the presence of a structurally inappropriate animate distractor, leading to an illusion of grammaticality. The second region was the reflexive, which served as an additional probe of the properties of the licensor that was retrieved for null subject licensing. Reading times at this region showed a similar profile to the gerundive verb with respect to facilitatory interference. Taken together, these findings suggest that the structurally inappropriate animate distractor was sometimes retrieved as the subject of the adjunct clause, which licensed the reflexive without detection of the animacy violation.

The finding that null subject licensing exhibits facilitatory interference effects is striking, given that such robust effects have rarely been observed for anaphora before. Previous studies have consistently failed to find evidence of facilitatory interference in the comprehension of anaphoric dependencies, such as those involving direct object reflexives. In contrast, we found that null subject licensing shows an interference profile that is qualitatively similar to that observed for subject–verb agreement dependencies, which show strong interference effects.

The findings from Experiment 2 showed facilitatory interference for null subjects in adjunct control structures. However, our interpretation of the interference profile at the emphatic reflexive is based on the assumption that the reflexive was a faithful reflection of what was retrieved as the subject of the adjunct clause at the gerundive. This assumption is based on previous findings that reflexives generally only search for a licensor within the domain of their local clause (e.g., Sturt, 2003; Dillon et al., 2013). However, an alternative explanation of our results is that the interference effects observed at the gerundive and reflexive reflect independent effects, and that the profile observed at the reflexive is not predicated on the outcome of null subject licensing at the gerundive. For instance, the reflexive may not have tracked the interpretation of the subject of the adjunct clause but rather linked directly to one of the NPs in the higher clause (e.g., *the doctor*, *the report*). Since little is known about the processing of emphatic reflexives, it is possible that, unlike direct object reflexives, emphatic reflexives may trigger an error-prone retrieval that is not constrained to the domain of the adjunct clause, thus giving rise to an interference effect that is independent of the outcome of null subject licensing. We tested this possibility in Experiment 3.

Experiment 3

Experiment 3 tested the assumption that the reflexive in Experiment 2 tracked the interpretation of the subject of the adjunct clause, rather than linking directly to one of the NPs in the higher clause. We reasoned that if the reflexive accurately reflected the interference effect observed for null subject licensing, then eliminating interference for null subject licensing should also eliminate interference at the reflexive. To achieve this, we held constant the animacy of the target NP in the main clause and distractor NP in the relative clause, and manipulated their gender match with the reflexive instead, as shown in (6).

(6) The (harpist|drummer) that the (diva|guitarist) liked very much was congratulated after playing the beautiful song **herself** at the brand new recording studio.

As described earlier, reflexives require gender agreement with a licensor, but null subjects do not. Thus, the gender manipulation in (6) should not generate any interference effects at the gerundive in the adjunct clause, as only the correct licensor (*harpist*/*drummer*) should be retrieved for null subject licensing. Further, if the reflexive is a faithful reflection of what was retrieved for null subject licensing, then the reflexive should only be sensitive to the gender match of the structurally appropriate licensor (*harpist* vs. *drummer*), and thus pattern with the gerundive in the absence of interference effects. If, on the other hand, the reflexive links directly to either of the NPs in the higher clause, then different profiles might be obtained for null subject and reflexive licensing. In particular, although we do not expect interference at the gerundive, we might observe an interference effect at the reflexive when there is a structurally inappropriate but gender matching distractor in the relative clause (e.g., *diva*).

Participants

Thirty-two members of the University of Maryland community participated in Experiment 3. Participants were either compensated $10 or received credit in an introductory linguistics course. The task lasted approximately 40 min and was administered as part of a 1-hour session involving unrelated experiments.

Materials

The design of Experiment 3 was the same as Experiment 2, except that the animacy of the target and distractor NPs was held constant, and their gender match to the reflexive was manipulated. The experimental materials consisted of 48 item sets, each containing eight conditions. The experimental conditions consisted of a 2 × 2 × 2 factorial design, which crossed the factors DEPENDENCY, GRAMMATICALITY, and DISTRACTOR. An example item set is provided in **Table 2**. As in Experiment 2, the target NP appeared as the subject of the main clause, and was modified by an object relative clause that contained the distractor in subject position. The factor DEPENDENCY compared adjunct control conditions with subject–verb agreement conditions. The factor GRAMMATICALITY was manipulated by varying the stereotypical gender of the main clause subject in the adjunct control conditions and the number of the agreeing verb in the subject–verb agreement conditions. In the grammatical adjunct control conditions, the main clause subject was animate and matched the gender of the reflexive. In the ungrammatical conditions, the main clause subject was animate, but mismatched the gender of the reflexive. In the grammatical subject–verb agreement conditions, the agreeing verb was always singular and matched the number of the main clause subject. In the ungrammatical conditions, the agreeing verb was plural and mismatched the number of the main clause subject. The factor DISCTRACTOR was manipulated by varying the stereotyped gender of the distractor for adjunct control conditions and the number of the distractor for subject–verb agreement conditions. As in Experiment 2, the lexical content of the main clause was held constant across dependency types to avoid spurious effects due to lexical differences.

TABLE 2 | Example set of experimental items for Experiment 3.

Adjunct control conditions
Grammatical, distractor
The harpist that the diva liked very much was congratulated after playing the beautiful song herself at the brand new recording studio.
Grammatical, no distractor
The harpist that the guitarist liked very much was congratulated after playing the beautiful song herself at the brand new recording studio.
Ungrammatical, distractor
The drummer that the diva liked very much was congratulated after playing the beautiful song herself at the brand new recording studio.
Ungrammatical, no distractor
The drummer that the guitarist liked very much was congratulated after playing the beautiful song herself at the brand new recording studio.
Subject–verb agreement conditions
Grammatical, distractor
The harpist that the diva liked very much was congratulated after playing the beautiful song at the brand new recording studio.
Grammatical, no distractor
The harpist that the divas liked very much was congratulated after playing the beautiful song at the brand new recording studio.
Ungrammatical, distractor
The harpist that the divas liked very much were congratulated after playing the beautiful song at the brand new recording studio.
Ungrammatical, no distractor
The harpist that the diva liked very much were congratulated after playing the beautiful song at the brand new recording studio.

Procedure

The same self-paced reading procedure was used as in Experiment 2.

Data Analysis

The statistical analysis followed the same steps as in Experiment 2. Four participants were excluded from the analysis due to accuracy below 70% in the comprehension questions. Data trimming affected less than 1% of the data. Model comparisons revealed that a maximally specified random effects structure did not provide a better fit to the data in the critical regions than an intercept-only model [subject–verb agreement: $\chi^2_{(18)} = 11.16$, $p = 0.88$; adjunct control: $\chi^2_{(18)} = 14.53$, $p = 0.69$]. Therefore, we adopted the intercept-only model.

Results

Subject–Verb Agreement Conditions

Figure 4 shows average reading times starting from the region preceding the agreeing verb to five regions following the main verb. No effects were observed at the critical verb (v). The word immediately following the critical verb (v + 1) showed a main effect of GRAMMATICALITY ($\hat{\beta} = 0.18$, SE = 0.03, $t = -5.06$) and, crucially, an interaction between GRAMMATICALITY and DISTRACTOR ($\hat{\beta} = -0.21$, SE = 0.07, $t = -2.98$). This interaction was driven by a significant effect of DISTRACTOR in the ungrammatical conditions ($\hat{\beta} = -0.16$, SE = 0.05, $t = -2.97$), reflecting faster reading times for sentences with a

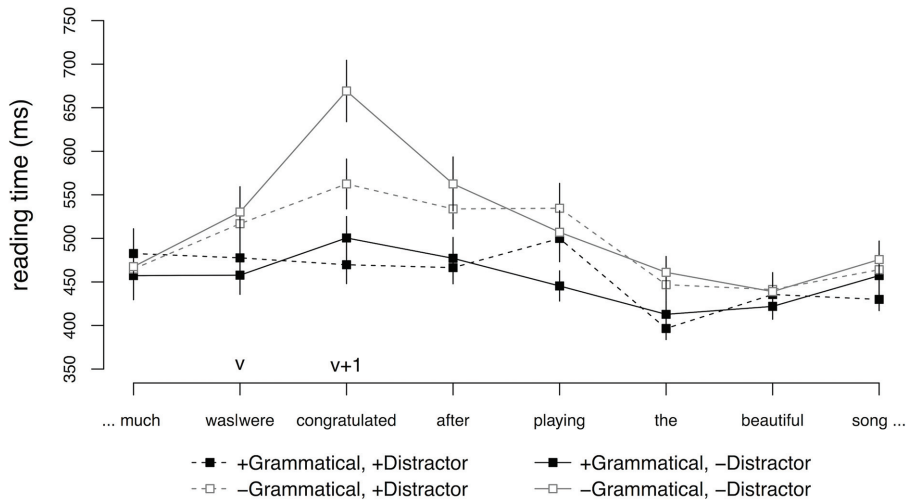


FIGURE 4 | Word-by-word reading times for subject–verb agreement conditions, Experiment 3. Error bars indicate SEM.

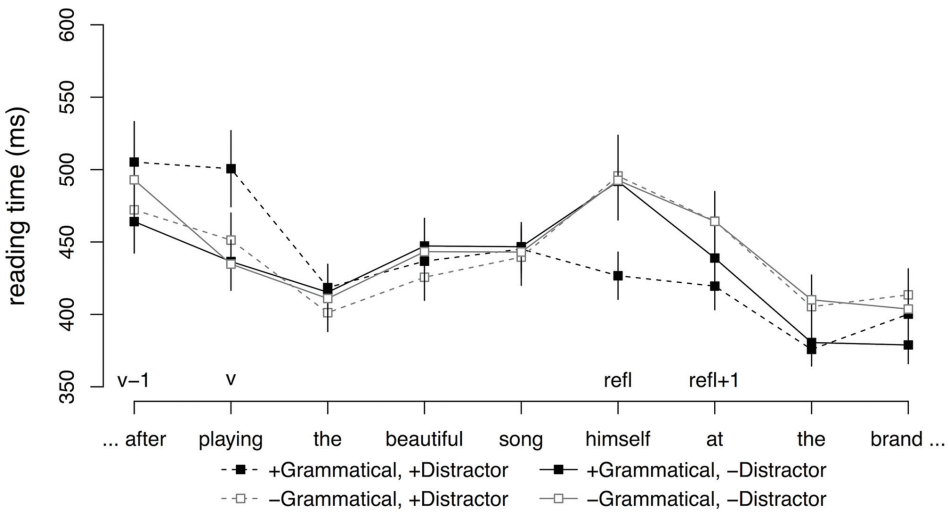


FIGURE 5 | Word-by-word reading times for adjunct control conditions, Experiment 3. Error bars indicate SEM.

plural distractor relative to sentences with no distractor. No such difference was observed for the grammatical sentences ($t < 2$).

Adjunct Control Conditions

Figure 5 shows average reading times starting from the subordinator to three regions following the reflexive. No effects were observed at the subordinator (v−1). At the gerundive verb (v), there was a main effect of distractor ($\hat{\beta} = -0.06$, SE $= 0.02$, $t = -2.31$). Pairwise comparisons revealed that this effect was due to a slowdown for grammatical conditions with an animate distractor ($\hat{\beta} = 0.10$, SE $= 0.04$, $t = 2.33$). No effect was observed in the ungrammatical conditions ($t < 2$). At the reflexive (refl), the grammatical condition with an animate distractor showed faster reaction times ($\hat{\beta} = -0.09$, SE $= 0.03$, $t = -2.41$). No other effects were observed at the reflexive (all *ts* < 2). The word immediately following the reflexive (refl + 1) showed a main effect of GRAMMATICALITY ($\hat{\beta} = -0.05$, SE $= 0.02$, $t = -2.09$), reflecting a slowdown for ungrammatical conditions relative to grammatical conditions. Crucially, and in contrast with Experiment 2, there was no effect of facilitatory interference at the word following the reflexive and no interaction was observed between GRAMMATICALITY and DISTRACTOR.

Discussion

Experiment 3 tested the assumption that the reflexive in the adjunct control constructions in Experiment 2 was a faithful reflection of what was previously retrieved as the subject of the adjunct clause. We reasoned that if the interference effect seen at the reflexive in Experiment 2 reflected the interference effect observed for subjects at the gerundive verb,

then eliminating interference at the gerundive verb should also eliminate interference at the reflexive. This outcome is not obvious, since different features are required to match to license the gerundive (animacy) and the reflexive (animacy, number, gender). As predicted, eliminating interference for null subject licensing also eliminated interference at the reflexive. These results provide preliminary evidence that the reflexive tracked the interpretation of the subject of the adjunct clause, rather than directly linking to one of the NPs in the higher clause. We discuss this further in the General Discussion.

Experiment 3 also revealed a main effect of distractor at the gerundive verb and the reflexive regions. Specifically, the presence of multiple gender matched licensors in the grammatical conditions (e.g., *The harpist that the diva... after playing... herself*) increased reading times at the gerundive verb, and later facilitated reading times in the same conditions at the reflexive. These effects were unexpected, and we believe that the effect of distractor at the gerundive might reflect a "fan" effect (Anderson, 1974; Anderson and Reder, 1999), which can arise in grammatical contexts when multiple items match the retrieval cues (Badecker and Straub, 2002; Autry and Levine, 2014; but cf. Chow et al., 2014).

In contrast with facilitatory interference effects at retrieval, fan effects have been argued to reflect interference at the encoding stage (Dillon, 2011). For example, encountering multiple items that overlap in morphological features can degrade the quality of memory representations for those items due to feature-overwriting (Nairne, 1988, 1990). Thus, the reading time slowdown at the gerundive for grammatical sentences with multiple match items may reflect impeded access to a degraded memory representation of the target at the point of retrieval for null subject licensing. Crucially, this effect does not entail that the structurally inappropriate licensor was retrieved during null subject licensing (see Dillon, 2011 for discussion). By contrast, the facilitation in the same conditions later at the reflexive could reflect a gender familiarity effect. After reading the gerundive verb, comprehenders might have been fairly confident that a gender matching item (*harpist* and *diva*) was present in the sentence, leading to facilitated processing (i.e., faster reading times) at the reflexive.

In sum, the critical finding from Experiment 3 is the absence of facilitatory interference effects in the ungrammatical conditions at the reflexive. These findings provide preliminary evidence that the reflexive in the adjunct control constructions from Experiment 2 tracked the interpretation of the subject of the same clause, rather than linking directly to one of the NPs in the higher clause. However, further research is necessary to better understand the source of the facilitation effect in the grammatical conditions.

General Discussion

Summary of Findings

The present study addressed one part of the puzzle of why reflexives and subject–verb agreement show contrasting profiles with respect to facilitatory interference effects. We tested the hypothesis that all anaphoric dependencies resist facilitatory interference from structurally inappropriate items during real-time comprehension. Specifically, we used an animacy manipulation to examine whether adjunct control dependencies, which involve an interpreted anaphoric relation between a null subject and its licensor, behave like reflexives in that they are immune to facilitatory interference effects.

In Experiment 1, we confirmed that null subject licensing in adjunct control structures obeys an animacy requirement, which we then used as a probe for interference effects in Experiment 2. In Experiment 2, we directly compared the reading time profiles of null subject licensing and subject–verb agreement dependencies. Our results revealed qualitatively similar profiles with respect to facilitatory interference, as illustrated in **Figure 6**. Specifically, we found reliable interference effects for null subject licensing at two points: at the gerundive verb and later, at a reflexive within the same clause, which served as an additional probe of what was retrieved as the subject of the gerundive verb.

The results from Experiment 2 challenge the hypothesis that all anaphoric dependencies resist facilitatory interference. Specifically, our results suggest that anaphors do not behave homogenously with respect to facilitatory interference, since null subject anaphors show interference, whereas reflexive anaphors typically do not. Thus, we believe that any account that claims that interference effects are linked to specific types of grammatical dependencies (e.g., anaphora vs. agreement) is unlikely to be successful. Furthermore, the results of Experiment 2 challenge the hypothesis that the contrast between reflexives and agreement seen in previous studies reflects differences based on their interpretive status. According to this hypothesis, anaphoric dependencies might engage a more conservative retrieval strategy to avoid misinterpretation and interference from structurally inappropriate items. Our results provide evidence against this hypothesis by showing that interpreted anaphoric dependencies involving null subjects are susceptible to facilitatory interference.

In Experiment 3, we tested the assumption that the reflexive in Experiment 2 tracked the interpretation of the subject of the same clause, rather than linking directly to one of the NPs in the higher clause. We tested a configuration that did not yield interference for null subject licensing, and found that the corresponding interference effect at the reflexive also disappeared, as shown in **Figure 6**. These results suggest that our assumption was justified.

However, there is an alternative explanation for the contrasting profiles at the reflexive between Experiments 2 and 3. Whereas in Experiment 2 the reflexive mismatched its licensor in both animacy and gender, in Experiment 3, the reflexive only mismatched its licensor in gender. This raises the possibility that the contrasting profiles at the reflexive between Experiments 2 and 3 may not reflect differences based on the outcome of null subject licensing. Rather, the contrast may have been caused by differences based on the degree of match between the reflexive and the candidate licensors (i.e., 2-feature mismatch in Experiment 2, but 1-feature mismatch in Experiment 3). Specifically, it is possible that the feature matching distractor was able to outcompete the target when the target mismatched the

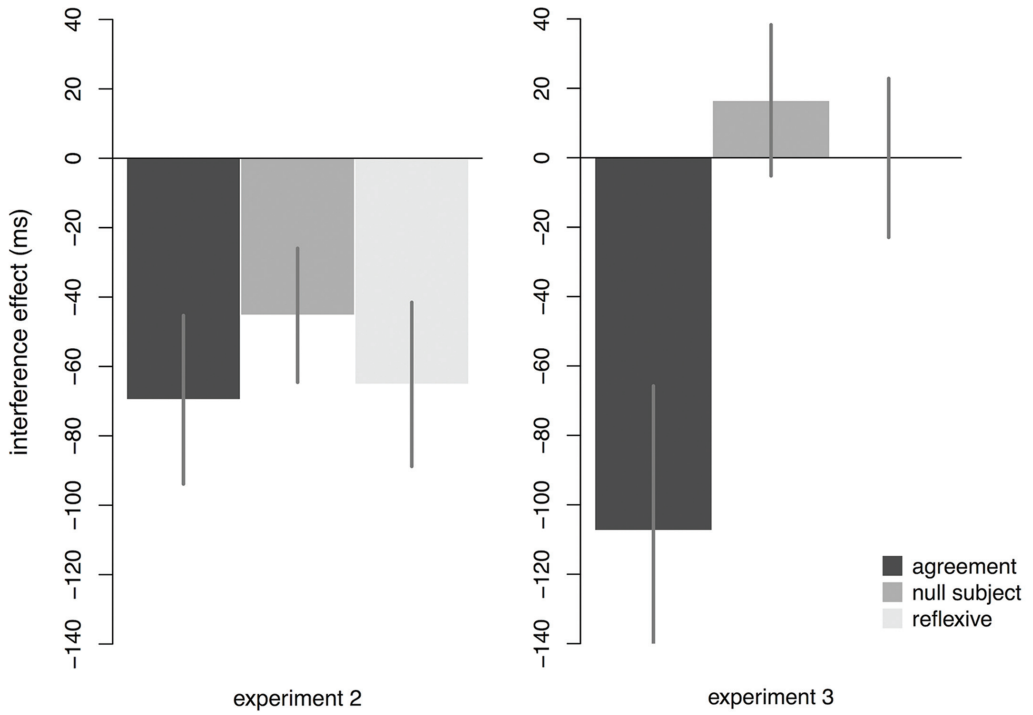


FIGURE 6 | Comparison of interference effects observed for subject-verb agreement (region: v + 1), null subject licensing (region: v), and reflexive licensing (region: refl + 1) in Experiments 2 and 3. Error bars indicate SEM. Interference effects (in ms) were estimated as the difference between the means of the two ungrammatical conditions.

reflexive in 2 features, but not when it mismatched the reflexive in only 1 feature. This difference could lead to interference in a 2-feature mismatch context (Experiment 2), but not in a 1-feature mismatch context (Experiment 3). We discuss this possibility further below.

Variability within Anaphora

The present study revealed that null subjects are susceptible to facilitatory interference in comprehension. These findings contrast with previous findings for reflexives, which typically resist interference. This raises the question of why anaphoric dependencies should behave differently at retrieval. We believe that there are two possibilities for why we should see variability within anaphora with respect to facilitatory interference.

First, previous studies on anaphora have failed to find evidence of facilitatory interference with designs that manipulated the gender or number match between the anaphor and its licensor. In contrast, we found evidence of facilitatory interference when we manipulated animacy. It is possible that the interference effects in our study reflect an inherent primacy of animacy information in anaphoric licensing. This could arise if animacy is a more reliable cue to the target subject in comprehension. For example, whereas a subject in a licensor position is typically animate, its gender and number may be more variable, leading comprehenders to prioritize animacy information at retrieval to access the target subject. This hypothesis aligns with recent findings on the psychology of memory, which suggest that animacy information is one of the most important dimensions in controlling memory retention (Nairne et al., 2013; Van Arsdall et al., 2013).

A second possibility is that the variability across studies could reflect the degree of feature match between the anaphor and its licensor (i.e., probe-to-target similarity). This possibility was raised earlier in our discussion of the contrasting profiles at the reflexive between Experiments 2 and 3. Specifically, we observed facilitatory interference at the reflexive when the target licensor mismatched the reflexive in two features (i.e., both gender and animacy), but failed to find evidence of interference when the target licensor mismatched the reflexive in only one feature (i.e., gender). These findings suggest that retrieval for reflexive processing might only be susceptible to facilitatory interference in configurations where the target mismatches the reflexive in more than one feature[6].

Our findings do not distinguish between the two possibilities discussed above, but they suggest some further directions. One avenue for future research would be to focus on the processing of direct object reflexives and to compare contexts in which

[6]A third possibility suggested by one of the reviewers is that retrieval involving animacy may be privileged over retrieval involving other features like gender or number, not because of the information-theoretic status of the antecedent, but because of the linguistic representation of feature hierarchies. Feature geometric approaches to agreement, e.g., Harley and Ritter (2002), propose that some features are represented as structurally higher than others, and we might expect that there are linguistic constraints on the order of feature access during retrieval.

the reflexive-antecedent dependency involves 1 vs. 2 feature mismatches. This test could help determine the source of the interference effects at the reflexive in Experiment 2 (see Parker and Phillips, 2014). Another avenue for future research would be to test the hypothesis that interference in anaphora is due to the privileged use of animacy information in retrieval. To achieve this, one could test languages where animacy and gender are not conflated, like Spanish or Polish. In these languages, gender is a syntactic property that is distinct from stereotypical or conceptual gender, such that a mismatch in animacy between an anaphor and its licensor does not entail a gender mismatch, like in English. Testing the impact of animacy independently of gender in these languages could help determine whether there is an inherent primacy of animacy in retrieval for anaphor processing.

Conclusion

This study explored the hypothesis that all anaphoric dependencies resist facilitatory interference during real-time comprehension. Our results challenged this hypothesis by showing that anaphoric dependencies do not behave homogenously with respect to facilitatory interference effects. Specifically, we found that adjunct control dependencies, which involve an anaphoric relation between a null subject and a licensor, are susceptible to facilitatory interference. In discussion, we explored several options for why anaphoric dependencies should vary with respect to facilitatory interference. We argued that variability within anaphora could reflect either an inherent primacy of specific content cues like animacy in retrieval processes, or the differential degree of match between the potential licensors and retrieval probe.

Acknowledgments

We would like to thank Brian Dillon, Norbert Hornstein, Dave Kush, Ellen Lau, Jeff Lidz, Luiza Newlin-Łukowicz, Shravan Vasishth, and Alexander Williams for helpful discussion. This work was supported in part by NSF-IGERT grant DGE-0801465 to the University of Maryland and by NSF grant BCS-0848554 to CP.

Supplementary Material

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpsyg.2015.01346

References

Anderson, J. R. (1974). Retrieval of propositional information from long-term memory. *Cogn. Psychol.* 6, 451–474. doi: 10.1016/0010-0285(74)90021-8

Anderson, J. R., and Reder, L. M. (1999). The fan effect: new results and new theories. *J. Exp. Psychol. Gen.* 128, 186–197. doi: 10.1037/0096-3445.128.2.186

Autry, K. S., and Levine, W. H. (2014). A fan effect in anaphor processing: effects of multiple distractors. *Front. Psychol.* 5:818. doi: 10.3389/fpsyg.2014.00818

Baayen, R. H., Davidson, D., and Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *J. Mem. Lang.* 59, 390–412. doi: 10.1016/j.jml.2007.12.005

Badecker, W., and Straub, K. (2002). The processing role of structural constraints on the interpretation of pronouns and anaphors. *J. Exp. Psychol.* 28, 748–769. doi: 10.1037/0278-7393.28.4.748

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for testing interactions in linear mixed-effects models. *Front. Psychol.* 4:328. doi: 10.3389/fpsyg.2013.00328

Bates, D., Maechler, M., and Bolker, B. (2011). *lme4: Linear Mixed-Effects Models Using S4 Classes*. Available at: http://CRAN.R-project.org/package=lme4

Chow, W. Y., Lewis, S., and Phillips, C. (2014). Immediate sensitivity to structural constraints in pronoun resolution. *Front. Psychol.* 27:630. doi: 10.3389/fpsyg.2014.00630

Clackson, K., Felser, C., and Clahsen, H. (2011). Children's processing of reflexives and pronouns in English: evidence from eye-movements during listening. *J. Mem. Lang.* 65, 128–144. doi: 10.1016/j.jml.2011.04.007

Clifton, C., Frazier, L., and Deevy, P. (1999). Feature manipulation in sentence comprehension. *Riv. Linguist.* 11, 11–39.

Dillon, B. (2011). *Structured Access in Sentence Comprehension*, Ph.D. Dissertation, University of Maryland, College Park, MD.

Dillon, B., Mishler, A., Slogget, S., and Phillips, C. (2013). Contrasting intrusion profiles for agreement and anaphora: experimental and modeling evidence. *J. Mem. Lang.* 69, 85–103. doi: 10.1016/j.jml.2013.04.003

Drenhaus, H., Saddy, D., and Frisch, S. (2005). "Processing negative polarity items: when negation comes through the backdoorm," in *Linguistic Evidence: Empirical, Theoretical, and Computational Perspectives*, eds S. Kepser and M. Reis (Berlin: de Gruyter), 145–165.

Ferreira, F., and Patson, N. D. (2007). The 'good enough' approach to language comprehension. *Lang. Linguist. Compass* 1, 71–83. doi: 10.1111/j.1749-818X.2007.00007.x

Gelman, A., and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.

Harley, H., and Ritter, E. (2002). Person and number in pronouns: a feature-geometic analysis. *Language* 78, 482–526. doi: 10.1353/lan.2002.0158

Hofmeister, P. (2011). Representational complexity and memory retrieval in sentence comprehension. *Lang. Cogn. Process.* 26, 376–405. doi: 10.1080/01690965.2010.492642

Hornstein, N. (2003). "On control," in *Minimalist Syntax,* ed. R. Hendrick (Oxford: Blackwell), 6–81. doi: 10.1002/9780470758342.ch1

Jäger, L. A., Benz, L., Roeser, J., Dillon, B., W., and Vasishth, S. (2015a). Teasing apart retrieval and encoding interference in the processing of anaphors. *Front. Psychol.* 6:206. doi: 10.3389/fpsyg.2015.00506

Jäger, L. A., Engelmann, F., and Vasishth, S. (2015b). Retrieval interference in reflexive processing: experimental evidence from Mandarin, and computational modeling. *Front. Psychol.* 6:617. doi: 10.3389/fpsyg.2015.00617

Kawasaki, N. (1993). *Control and Arbitrary Interpretation in English*, Ph.D. Dissertation, University of Massachusetts, Amherst, MA.

Kennison, S. M., and Trofe, J. L. (2003). Comprehending pronouns: a role for word-specific gender stereotype information. *J. Psycholinguist. Res.* 32, 355–378. doi: 10.1023/A:1023599719948

Kwon, N., and Sturt, P. (2014). The use of control information in dependency formation: an eye-tracking study. *J. Mem. Lang.* 73, 59–80. doi: 10.1016/j.jml.2014.02.005

Landau, I. (2001). *Elements of Control: Structure and Meaning in Infinitival Constructions*. Dordrecht: Kluwer.

Lau, E., Wagers, M., Stroud, C., and Phillips, C. (2008). "Agreement and the Subject of Confusion," in *Talk at the 21st CUNY Conference on Human Sentence Processing*, Chapel Hill, NC.

Lewis, R. L., and Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cogn. Sci.* 29, 375–419. doi: 10.1207/s15516709cog0000_25

Lewis, S., and Phillips, C. (2015). Aligning grammatical theories and language processing models. *J. Psycholinguist. Res.* 44, 27–46. doi: 10.1007/s10936-014-9329-z

McElree, B. (2000). Sentence comprehension is mediated by content-addressable memory structures. *J. Psycholinguist. Res.* 29, 155–200. doi: 10.1023/A:1005184709695

McElree, B., Foraker, S., and Dyer, L. (2003). Memory structures that subserve sentence comprehension. *J. Mem. Lang.* 48, 67–91. doi: 10.1016/S0749-596X(02)00515-6

Nairne, J. S. (1988). A framework for interpreting recency effects in immediate serial recall. *Mem. Cogn.* 16, 343–352. doi: 10.3758/BF03197045

Nairne, J. S. (1990). A feature model of immediate memory. *Mem. Cogn.* 18, 251–269. doi: 10.3758/BF03213879

Nairne, J. S., Van Arsdall, J. E., Pandeirada, J. N. S., Cogdill, M., and LeBreton, J. M. (2013). Adaptive memory: the mnemonic value of animacy. *Psychol. Sci.* 24, 2099–2105. doi: 10.1177/0956797613480803

Nicol, J., and Swinney, D. (1989). The role of structure in coreference assignment during sentence comprehension. *J. Psycholinguist. Res.* 18, 5–19. doi: 10.1007/BF01069043

Parker, D., and Phillips, C. (2014). "Selective priority for structure in memory retrieval," in *Poster at the 27th CUNY Conference on Human Sentence Processing,* The Ohio State University, Columbus, OH.

Patson, N. D, and Husband, E. M. (2015). Misinterpretation in agreement and agreement attraction. *Q. J. Exp. Psychol.* doi: 10.1080/17470218.2014.992445 [Epub ahead of print].

Pearlmutter, N., Garnsey, S., and Bock, K. (1999). Agreement processes in sentence comprehension. *J. Mem. Lang.* 41, 427–456. doi: 10.1006/jmla.1999.2653

Phillips, C., Wagers, M., and Lau, E. F. (2011). "Grammatical illusions and selective fallibility in real-time language comprehension," in *Experiments at the Interfaces,* Vol. 37, ed. J. Runner (Bingley: Emerald Publications), 147–180.

Quirk, R., Greenbaum, S., Leech, G., and Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. New York: Longman.

R Development Core Team. (2014). *R: A Language and Environment for Statistical Computing.* Vienna: R Foundation for Statistical Computing. Available at: http://www.R-project.org

Staub, A. (2009). On the interpretation of the number attraction effect: response time evidence. *J. Mem. Lang.* 60, 308–327. doi: 10.1016/j.jml.2008.11.002

Staub, A. (2010). Response time distributional evidence for distinct varieties of number attraction. *Cognition* 114, 447–454. doi: 10.1016/j.cognition.2009.11.003

Sturt, P. (2003). The time-course of the application of binding constraints in reference resolution. *J. Mem. Lang.* 48, 542–562. doi: 10.1016/S0749-596X(02)00536-3

Sturt, P., and Kwon, N. (2015). The processing of raising and nominal control: an eye-tracking study. *Front. Psychol.* 6:331. doi: 10.3389/fpsyg.2015.00331

Tabor, W., Galantucci, B., and Richardson, D. (2004). Effects of merely local syntactic coherence on sentence processing. *J. Mem. Lang.* 50, 355–370. doi: 10.1016/j.jml.2004.01.001

Tanner, D., Nicol, J., and Brehm, L. (2014). The time-course of feature interference in agreement comprehension: multiple mechanisms and asymmetrical attraction. *J. Mem. Lang.* 76, 195–215. doi: 10.1016/j.jml.2014.07.003

Tucker, M. A., Idrissi, A., and Almeida, D. (2015). Representing number in the real-time processing of agreement: self-paced reading evidence from Arabic. *Front. Psychol.* 6:347. doi: 10.3389/fpsyg.2015.00347

Van Arsdall, J. E., Nairne, J. S., Pandeirada, J. N. S., and Blunt, J. R. (2013). Adaptive memory: anaimacy processing produces mnemonic advantages. *Exp. Psychol.* 60, 172–178. doi: 10.1027/1618-3169/a000186

Van Dyke, J. A., and Lewis, R. L. (2003). Distinguishing effects of structure and decay on attachment and repair: a cue-based parsing account of recovery from misanalyzed ambiguities. *J. Mem. Lang.* 49, 285–316. doi: 10.1016/S0749-596X(03)00081-0

Van Dyke, J. A., and Lewis, R. L. (2007). Interference effects from grammatically unavailable constituents during sentence processing. *J. Exp. Psychol.* 33, 407–430. doi: 10.1037/0278-7393.33.2.407

Van Dyke, J. A., and McElree, B. (2006). Retrieval interference in sentence comprehension. *J. Mem. Lang.* 55, 157–166. doi: 10.1016/j.jml.2006.03.007

Van Dyke, J. A., and McElree, B. (2011). Cue-dependent interference in comprehension. *J. Mem. Lang.* 65, 247–263. doi: 10.1016/j.jml.2011.05.002

Vasishth, S., Brüssow, S., Lewis, R. L., and Drenhaus, H. (2008). Processing polarity: how the ungrammatical intrudes on the grammatical. *Cogn. Sci.* 32, 685–712. doi: 10.1080/03640210802066865

Vasishth, S., and Drenhaus, H. (2011). Locality effects in German. *Dialogue Discourse* 1, 59–82. doi: 10.5087/dad.2011.104

Wagers, M. (2008). *The Structure of Memory Meets Memory for Structure in Linguistic Cognition*, Ph.D. Dissertation, University of Maryland, College Park, MD.

Wagers, M., Lau, E. F., and Phillips, C. (2009). Agreement attraction in comprehension: representations and processes. *J. Mem. Lang.* 61, 206–237. doi: 10.1016/j.jml.2009.04.002

Xiang, M., Dillon, B., and Phillips, C. (2009). Illusory licensing effects across dependency types: ERP evidence. *Brain Lang.* 108, 40–55. doi: 10.1016/j.bandl.2008.10.002

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2015 Parker, Lago and Phillips. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

HYPOTHESIS AND THEORY
published: 10 February 2016
doi: 10.3389/fpsyg.2016.00082

On the Shallow Processing (Dis)Advantage: Grammar and Economy

Arnout Koornneef[1*†] *and Eric Reuland*[2*†]

[1] *Brain and Education Lab, Institute for Education and Child Studies, Leiden University, Leiden, Netherlands,* [2] *Utrecht Institute of Linguistics OTS, Utrecht, Netherlands*

In the psycholinguistic literature it has been proposed that readers and listeners often adopt a "good-enough" processing strategy in which a "shallow" representation of an utterance driven by (top-down) extra-grammatical processes has a processing advantage over a "deep" (bottom-up) grammatically-driven representation of that same utterance. In the current contribution we claim, both on theoretical and experimental grounds, that this proposal is overly simplistic. Most importantly, in the domain of anaphora there is now an accumulating body of evidence showing that the anaphoric dependencies between (reflexive) pronominals and their antecedents are subject to an economy hierarchy. In this economy hierarchy, deriving anaphoric dependencies by deep—grammatical—operations requires less processing costs than doing so by shallow—extra-grammatical—operations. In addition, in case of ambiguity when both a shallow and a deep derivation are available to the parser, the latter is actually preferred. This, we argue, contradicts the basic assumptions of the shallow–deep dichotomy and, hence, a rethinking of the good-enough processing framework is warranted.

Keywords: anaphoric dependencies, good-enough processing, variable binding, coreference, (reflexive) pronouns, economy hierarchy

OPEN ACCESS

Edited by:
Claudia Felser,
University of Potsdam, Germany

Reviewed by:
Clare Patterson,
Universität Potsdam, Germany
Peter Bosch,
University of Osnabrück, Germany

***Correspondence:**
Arnout Koornneef
a.w.koornneef@fsw.leidenuniv.nl;
Eric Reuland
e.reuland@uu.nl

†The names of the authors appear in alphabetical order.

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: *15 September 2015*
Accepted: *14 January 2016*
Published: *10 February 2016*

Citation:
Koornneef A and Reuland E (2016) On the Shallow Processing (Dis)Advantage: Grammar and Economy. Front. Psychol. 7:82. doi: 10.3389/fpsyg.2016.00082

INTRODUCTION

The marriage between linguistic theory and experimental psycholinguistics is a tumultuous one. On the one hand the one cannot live without the other, on the other hand, their relationship is characterized by frequent quarrels and misunderstandings. The tension is nicely shown by the following two quotes on "deep" vs. "shallow" processing. One is from Marantz (class lectures, 2000):

(1) *Deep processing* (our label, borrowed from the literature)

> "The split between linguistics and psycholinguistics in the 1970's has been interpreted as being a retreat by linguists from the notion that every operation of the grammar is a mental operation that a speaker must perform in speaking and understanding language. But, putting history aside for the moment, we as linguists cannot take the position that there is another way to construct mental representations of sentences other than the machinery of grammar....There is no retreat from the strictest possible interpretation of grammatical operations as the only way to construct linguistic representations."

The other position is aptly illustrated by a quote from Ferreira (2003):

(2) *Shallow processing*

> "The results .... suggest that a comprehensive theory of language comprehension must assume that simple processing heuristics are used during processing in addition to (and perhaps sometimes instead of) syntactic algorithms. Moreover, the experiments support the idea that language processing is often based on shallow processing, yielding a merely "good enough" rather than a detailed linguistic representation of an utterance's meaning."

Over the last decade, Ferreira's and related positions prompted a substantial line of research examining the driving forces in language processing. The broad scope of Ferreira's claim becomes particularly clear in a recent elaboration of the shallow, or "good-enough," processing position in Karimi and Ferreira (2015). That is, from their comprehensive overview of the literature and their implementation of a general "online cognitive equilibrium" model of language processing, one must conclude that they intend the good-enough processing position to hold generally across linguistic domains, arguing that "algorithmic procedures for sentence processing are not only too costly but sometimes outright unnecessary."

Since the shallow processing position has become an influential one, it deserves careful scrutiny. Yet, if we want to fully understand it, we are facing the fact that the mechanisms for shallow processing have not been formulated explicitly: it remains unclear how, precisely, they do their work. For instance, in order to fully understand the quote in (2) it is important to have a clear understanding of what counts as a heuristic. However, as Karimi and Ferreira state themselves, "the nature of the simple rules that guide heuristic processing is unclear." They do, however, provide the following helpful characterization: "We believe that this (=heuristic, K&R) processing relies more heavily on top-down information from semantic memory, whereas algorithmic processing seems to rely more heavily on linguistic knowledge to derive meaning in a bottom-up way, by organizing and combining the unfolding input using well-defined, successive linguistic rules. "It this 'top-down' vs. 'bottom-up' characterization of the relevant contrast, we will rely on in our discussions in the current contribution.

Furthermore, we would like to submit that if one wants to argue that in a particular situation people only assign a shallow interpretation, there is no escape from the requirement to make precise what this interpretation is, and which processes are involved in its derivation. Also in this respect the good-enough approach leaves some fundamental questions open.

The Nature of Good-Enough Representations in Language Processing

To illustrate the issues raised above further, let's think more carefully about the basic question of what counts as a good-enough representation. Note, that this question relates to rather fundamental questions about meaning representation. But, for the purpose of the present contribution we will try to stay as concrete as possible. Consider, then, the utterance in (3):

(3) The girl pushed the boy.

We would take it that in order to be good enough the utterance should be interpreted as representing a pushing relation rather than a kissing relation or an injuring relation. But, suppose we know that the pushing resulted in an injury, would a representation as an injuring relation still not be good enough? Perhaps it is, perhaps it isn't. In any case, it seems to us that a representation in which the boy does the pushing and the girl is pushed would certainly not count as good enough.

Now consider a case with quantifiers as in (4):

(4) Some girl pushed every boy.

Which scopal relation should count as good enough? Suppose *every boy* scoping over *some girl* is intended (with each boy being pushed by a different girl), is then the alternative with *some girl* scoping over *every boy* (that is, all the boys are being pushed by the same girl) still good enough? It seems, then, that the notion "good enough" in isolation is problematic. And in fact this is already illustrated in Ferreira's (2003) discussion. As she shows, in a remarkable number of cases participants assign a wrong interpretation to sentences. Thus, the question to ask is what representation, given limitations on attention and processing resources, will have to make do for a particular hearer in a particular situation, and how it is derived.

One area that provides a simple illustration of the different perspectives and the problems associated with the notion good enough is the interpretation of reversible passives as in (5) (Grodzinsky, 1995; Ferreira, 2003). As is well-known, children, agrammatic aphasics, but also certain typical speakers with no known deficit (Ferreira, 2003), show problems interpreting reversible passives[1]. For instance, agrammatic aphasics may show above chance performance on the active (5a), but only chance performance on the passive (5b), allowing (5c) as a possible interpretation.

(5) a. The girl pushes the boy. (*Above chance performance*).
b. The boy is pushed by the girl. (*Chance performance*).
c. The boy pushes the girl.

One may then hypothesize (as did Grodzinsky) that agrammatic aphasics cannot link the surface position of *the boy* to the object position in which it is assigned its semantic role. As such, *the boy* cannot receive a theme role. Or, in a less explicit manner, hypothesize, as did Ferreira (2003) in a study of typical participants, that there is a cost in following the grammatical algorithms. Consequently, in this view, the participants in these tests resort to an extra-grammatical interpretation strategy, based on the idea that there is a hierarchy of semantic roles and that the *agent* role is the most prominent role in this hierarchy[2].

[1]See Dąbrowska (2012) for the claim that such differences in interpretation reflect differences in acquisition, but Reuland (2012) for a demonstration that this claim is unwarranted.

[2]Or, hearers may adopt a simple Noun-Verb-Noun heuristic based on frequency patterns (Townsend and Bever, 2001). But this again raises non-trivial questions. That is, even if one assumes that this heuristic could be obtained on the basis of

(6) Assign the *agent* role to the leftmost NP of the clause as a default role.

This simple rule of *agent first* is indeed a good example of a top-down heuristic involving an extra-grammatical principle. Participants use this strategy to assign a role to *the boy* and interpret (5b) as (5c). Thus, crucially, in such a case the meaning representation arrived at is certainly not good enough. In fact it is not good at all. But it is *the* representation the participant arrives at and for him or her has to make do.

From these and similar results Ferreira (2003; see also e.g., Karimi and Ferreira, 2015) concludes that the field of language comprehension should adopt an approach similar to that taken in the Fast and Frugal Heuristics models (Gigerenzer et al., 1999; Gigerenzer, 2000), who take the position that "Models of rational choice which assume 'unbounded rationality' are unrealistic because the computations that are assumed to take place are often far too burdensome for real creatures operating in demanding environments." In the context of our present discussion, this would imply that a complete syntactic parse of a sentence—i.e., demanding the application of a wide range of grammatical computations—often yields a situation that is too burdensome for real creatures.

A crucial assumption is, then, that shallow processing—by avoiding a full syntactic parse and applying extra-grammatical heuristics instead—is cheaper than deep processing—which is based on the application of all available syntactic algorithms. This assumption clearly embodies an empirical claim. Let's call this assumption the *shallow advantage*. A second assumption, crucial for a fast and frugal heuristic to be viable at all is that using shallow strategies can in principle yield a similar interpretive result as grammatical computations—at least roughly so, even if speakers don't always do so. That is, even if such representations may be "incomplete," "lacking in detail," "sketchy," or "imprecise," they have to be good enough to be used (cf. Karimi and Ferreira, 2015). Let's refer to this as the *shallow equivalence* assumption; a strategy that can only lead to representations that for principled reasons fail to be at least moderately equivalent to what would have been derived by grammatical computations in a particular domain (i.e., a representation that could *never* be good-enough), may be frugal, but not very fruitful for the creatures using it. It is the aim of this contribution to critically assess these assumptions, which so far received too little attention in the literature from this perspective.

In the latter sense our goals are on a par with those of Karimi and Ferreira (2015), who in their recent proposal elaborated on the core assumptions of the good-enough processing framework (e.g., Ferreira, 2003; Ferreira and Patson, 2007). They put forward two fundamental processing principles that, in fact, closely mirror the two assumptions formulated above. More specifically, Karimi and Ferreira specified that "the reason why sometimes only fast and frugal heuristics rather than deep and time-consuming algorithms are applied during comprehension could be because heuristics offer a faster route to equilibrium (Principle 1). Similarly, the reason why the system is sometimes satisfied with a good-enough representation and does not exert the extra effort to engage in deeper processing could be because heuristics often provides enough equilibrium for the system, causing it to stay in that state for as long as possible… (Principle 2)" (pp. 6). Furthermore, following Kuperberg's (2007) syntactic-semantic model, Karimi and Ferreira claim that the algorithmic route of their implementation of the good-enough approach is syntactic in nature. The alternative route, on the other hand, relies more heavily on top-down information from semantic memory, and is capable of generating more global meaning representations of a sentence (intrasentential) or discourse (intersentential)[3].

Hence, Karimi and Ferreira's Principle 1 is identical to the shallow advantage assumption. Moreover, even though their Principle 2 is perhaps formulated less specifically than the equivalence assumption we ascribe to the good-enough position, Principle 2 reflects a similar core idea. That is, during the initial stages of processing there should be a perceived—or at least anticipated—equivalence between the output of the heuristics and algorithmic routes—after all, why should a creature be bothered with the construction of a mental representation that he or she knows will not be a reasonable reflection of the associated linguistic input?

To further substantiate their claims, Karimi and Ferreira (2015) present a comprehensive overview of studies that, in their opinion, are best explained by adopting a fast and frugal approach to language processing. These studies examined shallow linguistic processing for a wide range of different phenomena, such as the Moses Illusion, local syntactic ambiguities in garden-path sentences, quantifier scope ambiguities, erroneous interpretations of syntactically complex sentences, and the resolution processes of referring expressions (for references and more discussion, see Karimi and Ferreira, 2015).

As becomes clear from the discussion of Karimi and Ferreira (and as pointed out to us by one reviewer), a problem that arises if we set out to evaluate the shallow processing position is that the term "shallow processing" (originally due to Carter, 1985) is being used to refer to two different types of "shallowness" that must be kept apart—although they are not entirely unrelated. One involves the *top down* use of information from semantic memory, as briefly mentioned above. The other involves what one may call *reduced processing*.

That is, in some of the processing literature (for instance Stewart et al., 2007), and also some of the cases discussed by Karimi and Ferreira, shallow processing comes down to simply not fully processing part of the input—or at least delaying its integration (cf. Von der Malsburg and Vasishth, 2013). As Stewart

frequency—i.e., note that most canonical sentences only match this pattern if much material is ignored—it raises fundamental questions like, how is it stored, and how is it retrieved, and crucially how does it contribute to interpretation? In fact, it can only do so if the pattern is interpreted as "Subject Verb Object." But these are not surface encoded notions, but in fact already presuppose non-trivial analysis.

[3]Note that this use of the term semantics is different from ours. Whereas semantics in the model we will present refers to logical form (LF) representations, the label semantics in Kuperberg (2007) and Karimi and Ferreira (2015) in part reflects lexical semantics, and in part would be classified as discourse representations in our model.

et al. argue, in processing an input like *Paul$_i$ lent Rick$_j$ the CD before he$_{i/j}$ left for the holidays* the processor my simply disregard the temporal clause initially, and only yield a representation for Paul lending the CD to Rick. This type of shallow processing does not involve extra-grammatical heuristics. There is no top-down use of information coming from semantic memory and, in fact, it is compatible with deep processing using standard grammatical algorithms of whatever has been admitted to the processing buffer. For want of a better term, we will refer to it as "shallow-by-reduction" or *shallow-R*, in order to avoid confusion.

Shallow-R processing in the Stewart et al.'s sense (i.e., as partial non-processing) is not what we primarily address in this contribution—although it still raises non-trivial questions about the representations that are being derived. Instead, we will be focusing on claims about the type of shallow processing that explicitly involves the use of top-down information—including the use of extra-grammatical *heuristics*. We will refer to this notion of shallowness as "shallow-by-top-down," briefly *shallow-TD*.

This brings us to our main concern with these latter type of heuristics, which is actually three-fold, and can be summarized as follows: (1) it is unclear how they actually do the job they are taken to perform; (2) it is unclear whether they are necessary at all; (3) it is unclear—if they exist—why/whether they would be cheaper than the use of syntactic algorithms. We will start the discussion of these concerns on the basis of the *agent first* heuristic in the (shallow) interpretation of passive sentences—i.e., before moving on to anaphoric dependencies, which will be the main test case in the current contribution for the shallow-by-top–down position.

The Agent First Heuristic in Passive Sentences

Linguistic theory moves forward at a considerable pace. Consequently, considerations from the past need no longer apply to the current state of affairs. For instance, if it is claimed, after Slobin (1966), that "nonreversible sentences can be understood by going directly to the semantic roles without an intervening syntactic structure," we can easily see this is overly simplistic. As we now know, thematic role assignment is not just a matter of an argument "encountering" a predicate—containing an empty slot—in the mental working buffer and "filling the hole." Rather, the process involved minimally depends on verb and role type as shown for the contrasts between the processing of different types of intransitive verbs demonstrated in Koring et al. (2012).

Thus, even simple intransitive predicates have more internal structure than meets the eye, and this carries over to our initial example of passives. It is important to see that—in order for there to be a meaning representation at all—*the boy* in (5b) must be assigned a position to be formally identified as a subject (checking agreement, and/or case), that is, to function as an argument of the verb and its associated functional material. Given that under this construal it receives—mistakenly—the agent role, it must be able to semantically integrate with the verb in this capacity. Furthermore, *the girl* must be construed as the object and interpreted accordingly as bearing the theme role associated with this position. There is no escape from the assumption that in assigning this interpretation to the sentence, the processor has to treat the passive verb form as the active entry to which it is lexically related (Reinhart, 2002; Reinhart and Siloni, 2005). This it can only do if it *disregards* function words, such as *by* and *is*, and morphology like—*ed*. Thus, when (5b) is in fact interpreted as (5c), the "active" computation still needs to take place, which is not necessarily shallow at all. Or to put it bluntly, also *deriving a "wrong" interpretation requires explicit computations unless one advocates resorting to magic*.

But there is a further question. Namely, is an auxiliary, heuristics-based, interpretation strategy in fact necessary in this case? Recall that (5b) can only be interpreted as (5c) if the processor disregards the relevant functional elements. But note, that if it does so, the active interpretation is the only one that can be assigned. So in fact, no recourse to auxiliary strategies is needed. It is enough to assume that under certain conditions some functional elements—here, those necessary for a passive construction—will not enter the buffer of the processing system, and the processor simply works with what is has. From the perspective of Marantz's thesis in (1), then, one may assume that in order to interpret (5b) as (5c) a sufficiently articulate structure will be projected and interpreted by the rules the grammar contains. Projecting a structure that ignores the functional material that is present to license the passive interpretation (e.g., since it does not fit in the buffer due to cognitive overload, time pressure etc.), and subsequently using the active base form of the verb, will be quite enough to derive the interpretation observed. Hence, the most parsimonious assumption is that, at least in this domain, no extra-grammatical heuristics—other than disregarding functional elements—are involved at all. In short, here, shallow-TD reduces to shallow-R[4].

[4]Note that even the simplicity of a principle like *Agent first* is in fact not obvious. Properly considered it hides a considerable number of hidden assumptions. Consider, therefore, semantic role assignment in more detail:
Task: Assign a semantic role to DP_1

(i) DP_1 .... [$_{VP}$ V ....]

Options:

1. Apply compositional interpretation procedure to V, VP, and DP_1
2. Go to extra-grammatical set of heuristics, find *Agent first* and assign Agent to DP_1

The leading idea in heuristic based approaches is that option 2 is less costly than option 1. But the simplicity of Agent first is misleading. Consider the fuller structure in (ii):

(ii) DP_1 was [$_{VP}$ V-ed by DP_2]

In order to get the relevant (wrong) interpretation the processor has to:

1. Leave grammar
2. Access list of heuristic principles
3. Find Agent first
4. Apply Agent first
5. Disregard functional material (was, -ed, by)
6. Apply the predication rule between DP_1 and VP

But this raises issues such as:

- What do these heuristic principles arise from?
- Why does searching the list carry no cost?
- For a proper interpretation a non-trivial amount of syntactic/semantic computation will still have to be carried out (DP_1 will have to be interpreted

Of course, one may argue that at least in some cases utterances are interpreted by truly shallow processing. In case of high running emotions people may focus on one or two words in an utterance, completely ignoring any nuance and complexity the utterance may carry. And it is true that so far little is known about the syntax and semantics of exclamations. On the other hand, there is a growing literature on headlines and other similarly reduced linguistic expressions, which shows that these are far from arbitrary, and reflect an articulate linguistic structure underneath (De Lange, 2008). Thus, even if emotions highly limit the amount of items that are admitted into the processing system's buffer, this does not imply that whatever is admitted into the system is not subsequently structured and processed with grammatical means.

The Current Contribution

The discussion on passives as presented above nicely introduces the assumptions underlying the fast and frugal heuristics model, in particular the assumption that a full syntactic parse is complex, and hence, often more costly than the extra-grammatical strategies the lazy language user has at his disposition—what we refer to as the shallow advantage. It also becomes clear that the interpretation of passives is perhaps not the best domain to further evaluate the issue of *top-down* vs. *bottom-up* strategies—since shallow-TD can be reduced to shallow-R.

In the present contribution we will focus on the domain of anaphoric dependencies instead. The underlying reason is twofold. First, in their recent overview of the literature Karimi and Ferreira (2015) explicitly state that an important case of shallow processing in discourse is reference processing, a topic that in their opinion has not received enough attention in the good-enough literature. Second, and more importantly, we will argue that the by now firmly established linguistic theories on anaphoric dependencies allow us to more directly compare shallow and deep processing. Or to frame it more in terms of a good-enough approach (and Karimi and Ferreira's remark on the cost of algorithmic procedures for sentence processing): since grammatical computations and heuristic top-down principles are taken to compete, the well-defined grammatical (deep/bottom-up) and extra-grammatical (shallow/top-down) processing mechanisms of anaphoric dependencies present the perfect testing ground to critically assess whether grammatical computations are indeed "too cumbersome for real creatures"—i.e., as compared to their extra-grammatical alternatives.

In the following sections, we will argue that they are not. That is, we will argue—primarily on theoretical grounds—that the *shallow equivalence* assumption does not hold, at least not in this specific domain (Section *The equivalence assumption for shallow and deep anaphoric dependencies*). In addition, the *shallow-TD advantage* assumption has been evaluated in several experimental studies and, as we will demonstrate, shown to be false in the domain of anaphoric dependencies (Sections *The shallow-TD advantage assumption: Preparing the ground* and *The shallow-TD advantage assumption: The issue of economy*). To us, it seems that this provides enough reason to be skeptical about the aforementioned assumptions and, hence, this particular good-enough implementation of the heuristic/top-down approach to language processing.

THE EQUIVALENCE ASSUMPTION FOR SHALLOW AND DEEP ANAPHORIC DEPENDENCIES

It is a fundamental property of language in its relation to the world around us—and its mental representation—that different nominal expressions may receive the same value. Although nothing forces it—i.e., putting pragmatics aside for the moment—also nothing prevents that *the old baron* and *the driver* in the following sentence are used to refer to the same person.

(7) The old baron was crossing the bridge in a ramshackle carriage. The driver was visibly tired.

In this process of valuation, a linguistic expression is assigned a value from an extra-linguistic domain. Or more specifically, which value it receives is not grammatically determined. This provides a nice case of a potentially shallow-TD operation: Take an expression and assign it a value; prima facie no deep grammatical computations involved, and neither much searching if the referent is prominent in the context (in any case not more than general heuristics may be expected to require). Perhaps Karimi and Ferreira (2015) most clearly articulated this idea, since they state that "the processing of unambiguous referring expressions is facilitated because the comprehension system quickly reaches equilibrium by establishing the referential link between the referring expression and the antecedent through a simple, quick, and heuristics-based coindexation process, leading to little if any processing difficulty."

Whereas in sentence (7) we are dealing with two lexical noun phrases, the same option is available for pronominals, as in (8).

(8) *This soldier* has a gun. Will *he* attack?

He in (8) can be interpreted as the same individual as *this soldier*. However, this option is not available for

as a generalized quantifier, VP will have to be interpreted as a property that is a member of the GQ), and functional material will have to be left out of the computation).

- Why is there no cost involved in the clash between the result of *Agent first* and the functional material that is present?
- How does *Agent first* apply in the case of subject experiencing verbs (hate, admire), or unaccusatives, whose subjects are not agents?

It seems then that the purported simplicity of *Agent first* is not clear from the steps the processor has to take. Now compare this to the derivation by computation:

1. Retrieve V (θ_1, θ_2)
2. Disregard *–ed* and *by*
3. Assign θ_2 to DP_2
4. Disregard *was*
5. Apply the predication rule between DP_1 and VP and assign θ_1 to DP_1

None of the further issues arises and, the procedure applies both to agentive and to subject experiencer verbs, without unwarranted outcomes for other verb classes.

all nominal expressions. The mini-discourse in (9) is infelicitous:

(9) *No soldier* has a gun. *Will *he* attack?

This is due to the fact that *he* and *no soldier* cannot be co-valued. *No soldier* is quantificational, and does not introduce an entity *he* can refer to (an observation leading to the canonical distinction between binding and coreference in Heim, 1982; Reinhart, 1983, and subsequent work; see also Partee, 1978; Bosch, 1980, 1983). This makes anaphoric reference such as in (8) impossible. The same holds true of expressions like *every soldier*. Note that these well-known examples have important implications for how one should interpret the notion "discourse entity." That is, possible discourse antecedents are as diverse as soldiers, water, beauty, head-aches, dissatisfaction, etc. In addition to these nominal expressions, also sentences, verb phrases, prepositional phrases, adjective phrases, and tenses, admit anaphoric relations. Thus, the notion discourse entity must be broad enough to capture all these cases of anaphora, yet be restrictive enough to separate them from quantificational cases such as *no soldier*, or *every soldier*.

Crucially, although *he* cannot be co-valued with *no soldier*, it can depend for its interpretation on the latter. This is shown in (10):

(10) *No soldier* who has a gun hopes *he* will shoot.

That is, in (10) *he* can be semantically bound by *no* soldier. The semantic structure of (10) can be represented as in (11) where *he* is translated as a variable—x:

(11) No soldier who has a gun (λx.(x hopes [x will shoot]))

Under this construal the dependency of *he* on *no soldier* makes perfect sense. Here we have the relation of argument binding (A-binding), defined in terms of "logical binding," as in (12):

(12) A-binding (Reinhart, 2006)
α is A-bound by β iff α is the sister of λ-predicate whose operator binds β

A crucial difference between coreference and A-binding is that the latter, but not the former is subject to a structural condition, namely c-command. Briefly, as indicated by the definition in (12), the A-binder must be the sister of a constituent containing the bindee, as in (13):

(13) A-binder [. . . . bindee. . .]

The role of c-command is clearly illustrated by the contrasts in (14):

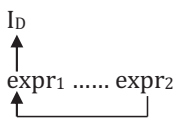
(14) a. The cop who found *the criminal* arrested *him*
b. *The criminal* found by the cop realized the latter would arrest *him*
c. *The cop who found *every criminal* arrested *him*
d. *Every criminal* found by the cop realized the latter would arrest him
e. *The cop who found *no criminal* arrested *him*
f. *No criminal* found by the cop realized the latter would arrest *him*

In (14a), *the criminal* does not c-command *him*, hence doesn't bind it, but since *the criminal* is referential and can have a discourse individual as its value, it can be co-valued with *him*, and there is no difference with (14b) where *the criminal* c-commands the pronominal; that is in both cases can *him* end up as covalued with *the criminal*. In (14c) *every criminal* does not c-command *him*, hence cannot bind it, hence the contrast with (14d). The same contrast is found between (14e) and (14f). Note, that one might argue that (14d) has a shallow counterpart that results by replacing *every criminal* by *all criminals*, and *him* by *them*, and possibly subjects might accept (14c) under such a construal (giving up on distributivity effects). But (14e,f) pose an insurmountable challenge to any such strategy; there is simply no alternative to a procedure in which *him* relates to the expression *no soldier*, and derives its interpretation from the instructions for interpretation this expression contains, since there is no discourse individual it could shallowly access instead[5].

In summary, these contrasts show that two different modes of interpretation must be distinguished: (1) (*shallow*), directly assigning two (or more) expressions the same discourse entity from the interpretation domain (I_D) as a value: *co-reference* as in (15a), and (2) (*deep*), interpreting one of the expressions first via another expression by grammatical—more specifically, semantic—means, as in (15b): *binding*[6].

(15) a. coreference:

b. binding:



It should be clear from these considerations that the contrast between co-reference and binding proves that a certain type of *bottom-up* (deep) dependency, namely binding, is plainly impossible to represent *top-down* (shallowly), without recourse to grammatical computation. Thus, the *equivalence* assumption underlying the shallow approach does not hold—i.e., in the domain of anaphora there is for principled reasons no extra-grammatical alternative to binding. This brings us to the discussion of the *shallow advantage* assumption—or more specifically, the *shallow-by-top-down (shallow-TD) advantage*—which cannot be refuted as easily on theoretical grounds alone, but requires recourse to an accumulating body of experimental evidence.

[5]Note, that it is crucial to distinguish between an expression and its value—for instance between the expression *John* and the individual John denoted by it. This distinction is often overlooked in the psycholinguistic literature, although clearly crucial if one wants to relate processing effects to the working of the memory system.

[6]One might think that such a two-route model is "uneconomical." However, clearly, each of these routes has its own independent motivation, and given their nature partial overlap between their effects is unavoidable.

THE SHALLOW-TD ADVANTAGE ASSUMPTION: PREPARING THE GROUND

While binding is subject to grammatical conditions, as we saw in (14), co-reference, by its very nature is not. Furthermore, binding is not only subject to the c-command requirement but elements to be bound can also be subject to a locality condition as illustrated in (16), with binding represented by co-indexing.

(16) a. Alice$_i$ expected [the King to admire her$_i$]
b. *Alice expected [the Queen$_i$ to admire her$_i$]

The upshot is that—as can be easily observed in the vast majority of languages—a pronominal may not be too close to its binder. In (16a) *Alice* is sufficiently far away from *her* to serve as its antecedent, but in (16b) *the Queen* is too close, matches in features, but yet is not allowed to bind *her*. This is one of the main patterns captured by Condition B of the Canonical Binding Theory (CBT, Chomsky, 1981, 1986)[7]:

(17) Condition B
A pronominal must be free (=not bound) in its governing category (=roughly, the domain of its nearest subject).

Condition B is a grammatical principle. But, one may wonder, why cannot the prohibition expressed by condition B be bypassed by using coreference? That is, even if *the Queen* in (16b) cannot bind *her*, why cannot the language system simply resort to the strategy (or top-down heuristic) in (15a), and assign the same individual to *the Queen* and to *her*? If this would be possible, we would never see the effects of condition B with referential antecedents, contrary to fact. This led to the postulation of what one may call a "traffic rule," reflecting an economy principle: *If a particular interpretation is ruled out by the grammar, this prohibition may not be bypassed* (see Reinhart, 1983; Grodzinsky and Reinhart, 1993; Reinhart, 2006; Reuland, 2011a, for discussion of this principle in various forms). In short, there are sentences where a binding and a coreference construal potentially compete, and when the binding dependency is rejected by the grammar, the coreference alternative is not considered. This is indicative of an economy ranking: *grammar* < *discourse*, reflected in Reinhart's Rule I and its successors[8].

The notion of an economy ranking plays an even more crucial role in the Primitives of Binding (PoB) model developed in Reuland (2001, 2011a), where the conditions on binding are derived from more elementary properties of the grammatical system. In its simplest form this economy measure is based on the assumption that the language system as whole is "lazy" and prefers to minimize the number of cross-modular steps, as in (15b) with one cross-modular step less than (15a)—i.e., in (15a) more information needs to be transferred from the grammar system to the interpretational system than in (15b).

The dependencies discussed so far were established in the "translation procedure" from syntactic representations to the interpretation system. But dependencies can also be pre-encoded by morpho-syntactic means. Quite characteristically, morpho-syntactic dependencies are obligatory. Whereas *him* in (14c) cannot depend for its interpretation on *every criminal*, nothing prevents it from being interpreted as some individual in the associated discourse. This is different from what we see with anaphors, like English *himself*, Dutch *zich(zelf)*, etc. Such expressions *must* be bound (at least in the core cases, see Reinhart and Reuland, 1993). Moreover, they must be bound in a very local domain, as illustrated for English by the contrast in (18), here again represented with the index notation:

(18) a. *Alice$_i$ expected [the King to admire herself$_i$]
b. Alice expected [the Queen$_i$ to admire herself$_i$]

In (18a) *Alice* is too far away from *herself* to serve as its antecedent, whereas *the King* is not a suitable antecedent due to a gender mismatch. As a result the sentence is ungrammatical. In (18b) *the Queen* is near enough, matches in features, and hence, binds *herself*. This is one of the main patterns captured by Condition A of the Canonical Binding Theory (CBT, Chomsky, 1981, 1986):

(19) An anaphor is bound in its governing category (=roughly, the domain of its nearest subject).

A characteristic property of the CBT is that it was based on a mix of syntactic and semantic properties. The notion of governing category is syntactic, the notion of binding itself is semantic, and the notion of an index—one of its key ingredients—was of a hybrid syntactic-semantic nature. This made it highly problematic as an ingredient of an explanatory theory (see Reinhart, 1983, for an initial discussion, and Reuland, 2011b, for a systematic overview of the problems with indices).

Minimalist approaches to grammatical structure (Chomsky, 1995, and subsequent work) introduced a strict separation between morpho-syntax and the interpretive system. Indices are not morpho-syntactic objects, hence, it was concluded, they have no place in syntax. Consequently, whatever there is syntactic in the binding conditions—such as locality—has to be derived with purely syntactic means. The means to do so in syntax are limited, just Movement and Agree (feature checking and valuation). This necessitated a thorough rethinking of binding and the binding conditions. A specific proposal to implement this was developed in Reuland (2001), and elaborated in Reuland (2011a). For reasons of space we will limit the discussion here to a few key issues, starting with condition A of the CBT.

In short, in Reuland (2011a) the locality property of *himself* is shown to follow from the semantic fact that *self* is an inherently reflexive relational noun. Given this property, *self* reflexivizes the predicate of which *himself* is an argument by—covert—head movement onto the verb (that is, it is interpreted as a reflexivizing operator). As we independently know, head-movement is strictly local (Travis, 1984). Hence, the locality of *himself* follows from

[7]For expository purposes we will ignore the subsequent modifications and explanations of Condition B in Reinhart and Reuland (1993); Reuland (2001, 2011a).

[8]Rule I: Intrasentential Coreference (Grodzinsky and Reinhart, 1993):
NP A cannot corefer with NP B if replacing A with C, C a variable *A*-bound by B, yields an indistinguishable interpretation.

the locality of head-movement. Thus, the relevant aspect of the representation of (18b) is as in (20):

(20) Alice expected [the Queen to SELF-admire her(self)][9]
….. the Queen (λx. (x admires SELF (x))

The upshot is, then, twofold. First, the interpretation of *himself*/*herself* involves a purely syntactic movement operation. Second, it is not just a matter of *himself*/*herself* 'looking for an antecedent' and being valued by the latter, but the process crucially involves the reflexivization of the predicate.

Indeed, there is independent experimental evidence that the processing of SELF-anaphors involves the verb in addition to whatever properties the antecedent may contribute. For example, Manika (2014), and Manika et al. (2014) using an information-theoretic approach (see Kostić, 1991, 1995 and subsequent work) show that the interpretation of a referentially dependent lexical item like Dutch *zichzelf* is modulated by the complexity of the verb—as quantified by the inflectional entropy of its paradigm—indicating that the interpretation of *zichzelf* involves an operation on the verb itself[10].

Also the binding of simplex anaphors like Dutch *zich* in (21) is encoded in the syntax (where SE stands for *simplex element anaphor*):

(21) *De klimmer* voelde [*zich* wegglijden]
The climber felt [SE slip away]

Here the encoding is brought about by the operation Agree: *zich* is deficient for gender and number, and is valued by Agree copying these features from the antecedent onto *zich*. The fact that (22) is ill-formed, again follows from economy (this contributes to deriving the canonical Condition B).

(22) **De klimmer* voelde [*hem* wegglijden]
The climber felt [him slip away]

To account for the fact that (22) is ruled out we apply the same logic as in the case of Rule I. The anaphoric dependency between *de klimmer* and *zich* in (21) can be encoded in syntax by Agree, but now consider the case where *hem* is selected, as in (22). Since *hem* is fully specified, it has no empty cells. Consequently, valuing it in the syntax by Agree is not an option. Hence, *zich* wins. But, crucially, this can only work if syntax cannot be bypassed by a derivation in which *hem* is directly interpreted as a bound variable by applying (12). So, syntax has to be considered before semantic binding can apply, and if syntax rejects the derivation, this is final. Since a syntactic operation such as Agree operates locally, we see this competition only when the dependent element is within the Agree domain of the element it is to depend on and not when it is further away[11]. Consequently, we arrive at the economy ranking in (23).

(23) syntax < semantics < discourse

We have by now prepared the ground for a discussion of the second assumption of good-enough interpretations: Are, deep—grammatical—operations indeed more costly for the processor than shallow-TD operations (i.e., in contrast to the economy ranking as depicted above)?

THE SHALLOW-TD ADVANTAGE ASSUMPTION: THE ISSUE OF ECONOMY

Interestingly, the issue of economy has received quite a bit of attention in the experimental literature, though not from the perspective sketched in the current contribution.

The Economy of Syntax

As is well-known, research on language acquisition shows an asymmetry between the performance of young children on condition A as compared to condition B. For instance, Chien and Wexler (1991) explored the question of whether children know Principles A and B from the outset or not. Their experiments show that children correctly require *local* antecedents for reflexives (Principle A) early on, whereas they are significantly delayed in disallowing *local* antecedents for pronouns (Principle B). As argued in Grodzinsky and Reinhart (1993) the computations involving the correct application of condition B are more costly than those involved in condition A. From the present perspective this indicates that the syntactic mode of encoding is indeed the least costly[12].

Although it is generally assumed in the psycholinguistic literature that condition A is a syntactic condition, it may be good to point out that in the PoB system condition A, as it is

[9]This analysis entails that locality is not an *intrinsic* property of *himself* qua being an anaphor. In fact, in positions from which *self* cannot move, there is no locality effect. This is what explains the fact that in (i), where *herself* is contained in a coordinate structure from which movement is prohibited there is no locality effect and *herself* can be bound by *Alice* (Reinhart and Reuland, 1991; Reuland, 2011a) despite the latter's distance:
(i) *Alice* was happy that the King invited the Rabbit and *herself* for tea

[10]This necessitates a rethinking of the conception of binding in the experimental literature, where it is mostly assumed that binding is just a matter of the anaphor looking for an antecedent.

[11]A simple illustration is provided by 1st person plural pronouns in Brazilian Portuguese. It has two forms: *nós*, which is both formally and semantically 1st person plural, and *a gente*, which is formally 3rd person singular, but semantically 1st person plural. *Nós* is free to semantically bind *a gente* and vice versa, but not when they are too close. In that case Agree causes a syntactic feature clash. But crucially, this clash cannot be bypassed by immediately going to the semantics. Note that this exposition is highly simplified. See Reuland (2011a) for the details.

[12]As brought up by one of the reviewers, since Chien and Wexler (1991) and Grodzinsky and Reinhart (1993) there has been considerable discussion about the status of the Delay of Principle B effect. For this discussion two issues must be distinguished. First, there is the question of whether there is a delay in the proper interpretations of pronominals at all. Second, as argued by both Chien and Wexler, and Grodzinsky and Reinhart, the delay shows up primarily with referential antecedents and not with quantificational antecedents. Elbourne (2005) expresses concerns about the adequacy of the experimental designs in these and subsequent studies that argue for such a difference between referential and quantificational antecedents. Conroy et al. (2009) present a number of new experiments that also call this contrast into question, coupled with an extensive overview of experiments discussed in the literature. Summarizing these contributions, it is clear that more factors have to be controlled for than previously assumed, not only involving the design but also the morpho-syntactic composition of the pronominal elements being studied (e.g., Baauw, 2002; Hartman et al., unpublished manuscript). However, even so, one can still maintain that children are more susceptible to Principle B violations than adults. Moreover, none of the literature cited calls into question that children behave quite adult-like with respect to condition A. Consequently, the general claim in the main text is not at issue.

reinterpreted, is indeed a purely syntactic operation (of course with semantic consequences). Hence, it is an ideal testing ground for the shallow vs. deep processing issue.

There are a number of experiments reported in the literature that test the status of condition A. Crucial for the present discussion, their results indicate that condition A applies early in the time course of processing and is, in addition, very robust. To illustrate this, in a well-known study Sturt (2003) carried out two eye-tracking experiments measuring (mis)match effects in sentences such as (24):

(24) *Jonathan/Jennifer* was pretty worried at the City Hospital. *He/She* remembered that *the surgeon* had pricked *himself/∗herself* with a used syringe needle. There should be an investigation soon.

In all the conditions only one character was structurally available (i.e., *the surgeon*, a profession with a stereotypically male gender). The distracting character (i.e., *Jonathan* or *Jennifer*) was highly prominent in the preceding discourse, yet not accessible as an antecedent for the reflexive. The results showed that if the reflexive and structurally available antecedent differed in gender, this immediately slowed down the reading process. Moreover, at this point during processing the distracting character (i.e., *Jonathan/Jennifer*) did not influence the resolution process. This suggests that the language system first attempts to link the reflexive to an antecedent that is structurally available—which will immediately fail when there is a gender-mismatch. In a follow-up experiment Sturt modulated the relative position of the distractor in the sentence, but the same conclusion was supported.

These finding are on a par with several other studies adopting a wide range of methodologies. For example, in an ERP experiment where the participants processed sentences such as (25), Xiang et al. (2009) investigate "intrusion effects" of potential, but non-commanding antecedents that appear—intrude—on the path between the SELF-anaphor and its antecedent.

(25) a. Congruent
The tough soldier that Fred treated in the military hospital introduced himself to all the nurses.
b. Intrusive
The tough soldier that Katie treated in the military hospital introduced herself to all the nurses.
c. Incongruent
The tough soldier that Fred treated in the military hospital introduced herself to all the nurses.

Furthermore, they compared these conditions to paired conditions in a second ERP-experiment in which intruders were present on the path between Negative Polarity Items and their licensers. Although they did find intrusion effects in the latter case, no significant intrusion effects were obtained in the case of the SELF-anaphor conditions as presented above (i.e., the ERP-waveforms revealed no difference between condition b and c). They concluded that during reflexive binding, syntactic constraints appeared to prevent intrusive antecedents from influencing the initial stages of anaphor resolution. In our view this points toward an early and robust application of the syntactic process establishing the dependency.

As a final example, Cunnings and Felser (2013) investigated the processing of SELF-anaphors in English, using the eye-tracking methodology. In their experiments they compared the performance of low working memory span with high working memory span readers. Here we will focus on one experiment—their Experiment 2—in which they measured the effect of a linearly intervening—but inaccessible antecedent (due to lack of c-command) using sentences as in (26):

(26) *James/Helen* has worked at the army hospital for years. *The soldier* that *he/she* treated on the ward wounded *himself/∗herself* while on duty in the Far East. Life must be difficult when you are in the army.

If Principle A would reflect a processing-based constraint this would lead to a different prediction than if it were a purely syntactic constraint. In the former case, particularly lower span readers may initially attempt to keep referential dependencies as short as possible. If so, main effects of the inaccessible antecedent should initially be observed. Higher span readers, on the other hand, would be less likely to find the creation of longer anaphoric dependencies difficult. It was found that for both lower and higher span readers the online application of Principle A could not be reduced to a (shallow-TD) memory-friendly "least effort" strategy of keeping anaphoric dependencies as short as possible[13]. All in all, the joint results of the two experiments they reported support, as they put it, a growing body of evidence showing that binding Principle A applies early during sentence processing to help guide reflexive anaphor resolution (e.g., Nicol and Swinney, 1989; Felser et al., 2009; Felser and Cunnings, 2012; Xiang et al., 2009; but see Badecker and Straub, 2002, for some conflicting evidence; see Dillon, 2014, for an excellent overview of all the relevant results).

Hence, a preferential position of syntactic encoding with respect to other strategies of anaphora resolution is warranted, which is in line with the PoB model (but not predicted by other approaches to binding). Or to put it slightly differently, a deep syntactic operation like binding of a SELF-anaphor is less costly for the processor than shallower operations, in contrast to what the "shallowness" approach predicts. In fact, this already is sufficient to establish our main point. There is *no clear support for a shallowness advantage*. Rather the opposite is the case: for the human processor *deep syntactic computations are preferred over shallow-TD interpretation processes*.

However, it will nevertheless be important to also assess the other members of the economy hierarchy as formulated in the PoB model: binding and coreference. This is what we will do next.

The Economy of Binding and Coreference

A well-known instantiation of the (economy) contrast between binding and coreference, introduced in (23) above, shows up in the interpretation of sentences with VP ellipsis, as in (27):

[13]Note that such a general least effort principle is highly implausible on other grounds, given the existence of long-distance anaphors in many languages, for instance, in Scandinavian.

(27) John fed his cat and Peter did too

Before we elaborate on this contrast in terms of economy, however, some facts and assumptions on VP-ellipsis should be discussed. First of all, it is clear that the second conjunct is about Peter feeding a cat, rather than about him combing a dog. This, uncontroversially, is a fact any theory of language will have to capture. A common idea is that for interpretation to obtain, the content of the VP in the second conjunct must somehow be recovered from the preceding context. As a first go one may assume a copying operation, as in (28).

(28) John fed his cat and Peter did <feed his cat> too.

As one can see, this gives rise to a puzzle, since the elided (i.e., covert) pronominal *his* in the second conjunct is ambiguous. More specifically, the interpretation of the full sentence can be either that John fed John's cat and *Peter fed Peter's* cat, as in (29a), or that John fed John's cat and that *Peter also fed John's cat*, as in (29b):

(29) a. John (λx. (x fed x's cat)) & Peter (λx. (x fed x's cat))
b. John (λx. (x fed a's cat) & a=J) & Peter (λx. (x fed a's cat) & a=J)

In (29a) *his* is interpreted as a variable, x, A-bound by Peter. This is what is generally referred to as the *bound variable* (BV), or "sloppy"[14] interpretation. In (29b), however, *his* is interpreted as a constant, here represented as *a*, which can receive the value of any individual in the discourse including John. That is, the occurrences of *his* in both conjuncts are *coreferential* (COR), yielding a "strict" interpretation.

Interestingly, the human processor is sensitive to this difference, and more importantly, it is a consistent finding in offline studies that in the interpretation of ambiguous VP-ellipses, BV-based interpretations are preferred over COR interpretations (see Frazier and Clifton, 2000, for an overview). This "preference" is reflected in the fact that typical subjects show longer reading times on COR in self-paced reading experiments (reported in Frazier and Clifton). In another experiment on the interpretation of VP ellipses that involved subjects with agrammatism, these subjects performed 80% correct on BV interpretations, but at chance on COR interpretations (Vasic et al., 2006). Curiously, then, what might seem to be the less sophisticated—more shallow—procedure, is the one that comes out as more costly in this case as well.

On the basis of such findings Frazier and Clifton (elaborating Reinhart, 1983; Avrutin, 1994, 1999) propose the following thesis as a hypothesis worth exploring:

(30) LF only/first hypothesis:
Bound-variable interpretations are preferred because the perceiver need only consult the LF representation (not the discourse representation) in order to identify the bound-variable analysis of the sentence.

[14]We use this term since it is so entrenched in the literature. But note that the "sloppy interpretation" is the one that does require grammatical operations. So, this is the one that is not shallow.

In order to do so they carry out a number of exploratory experiments and conclude that the hypothesis, though compatible with some of their results, is too problematic to be maintained.

However, as discussed by Frazier and Clifton (see also Koornneef, 2008; Koornneef et al., 2011) their results should be interpreted with some care, due to limitations of the experimental design and the statistical evaluation. In order to obtain more dependable results, subsequently, a number of full-size experiments using a more sensitive methodology were carried out, reported in Koornneef (2008, 2010), and Koornneef et al. (2006, 2011). Since the case is illustrative of the need to take theoretical advances into account we will briefly discuss Frazier and Clifton's interpretation of their findings before turning to the experiments of Koornneef and his colleagues.

One of the problems Frazier and Clifton note is of a *theoretical* nature. As they observed, a BV-preference also obtains across sentence boundaries, as in (31) (Experiment 1b). According to Frazier and Clifton this is incompatible with the nature of LF operations. That is, one would expect a grammatical operation like VP-copying to be limited to the domain of a sentence.

(31) Sarah left her boyfriend in May. Tina did [leave her boyfriend] too.

The other problem is *empirical* in nature. The choice between variable binding and coreference also shows up in the interpretation of *only*-sentences, illustrated in (32). Here it concerns the interpretation of the pronominal *he* in the complement clause of *think*. And again the pronoun shows an ambiguity. However, contrary to VP-ellipsis, Frazier and Clifton find a preference for a COR interpretation instead of the BV interpretation.

(32) Only Alfred thinks (that) he is a good cook.
a. Only Alfred thinks that Alfred is a good cook. (COR)
b. The only person who thinks of himself as a good cook is Alfred.
(BV).

On the basis of these findings, Frazier and Clifton conclude that the LF-only hypothesis (and equivalents) cannot be maintained. This, however, leaves a puzzle. Why would the case of VP-ellipsis be different from the *only*-case and what conclusions should we draw about the language processing system? Let's first address the theoretical issue Frazier and Clifton raise.

Theoretical Issue: What Mechanism Underlies Ellipsis?

The mechanism originally assumed in the literature on VP-ellipsis since Hankamer and Sag (1976) involved a copying operation (see Elbourne, 2008, for an overview and references). If so we would have to assume that the empty VP in the second sentence in (33a)—indicated by Δ—would be filled by a syntactic operation applying across sentences.

(33) a. Sarah left her boyfriend in May. Tina did Δ too.
b. Sarah (λx. (x left x's boyfriend)). Tina (λx. (x left x's boyfriend)) too.

This, Frazier and Clifton feel, violates the generally accepted idea that grammatical operations are limited to the sentential domain. Therefore, Δ cannot be interpreted by a grammatical copying operation. The question is, then, what kind of mechanism, is involved.

In recent years, however, independent evidence has been found that the theory of ellipsis should allow for greater flexibility (Merchant, 2001, 2008; Elbourne, 2008). This is illustrated by cases like (34) (Elbourne, 2008):

(34) Saskia, being a competitive type, has managed to acquire all the skills that Maaike and Brigitte possess. Maaike dances. Brigitte sings. Saskia does Δ too.

Here, Δ can be interpreted as the combined property of singing and dancing. In order to account for these and a variety of other cases, Elbourne proposes that ellipsis sites have internal—unpronounced—syntactic structure and are to be analyzed as silent "definite descriptions." In line with this, (33a) would be represented as (35), where the label TheP indicates that the complement of *did* is such a silent definite description (perhaps superfluously, we also indicate the silence by strike-through).

(35) Sarah left her boyfriend in May. Tina did [$_{\text{TheP}}$ ~~leave her boyfriend~~] too.

Then, to interpret the VP-ellipsis, the parser must somehow access the context (in this case "Sarah left her boyfriend in May") to retrieve the values for the constituent parts of TheP. Elbourne provides an elegant, yet fairly extensive and technical implementation whose details are beyond the scope of our present contribution. Relevant here is that, as he shows, the interpretation of the ellipsis site does not depend on a sentence-grammar "copy-and-paste operation," but rather reflects how a pronominal picks up its reference. That is, the elided VPs are treated as null pronouns, and under anybody's account, pronouns are able to pick up values from the preceding context. Hence, the relevant difference with the LF copying account is that under Elbourne's approach there is no theoretical reason to expect the context for the interpretation of VP-ellipsis to be limited to the same sentence.

What does the above mean for the explanation of a BV preference in VP-ellipsis like (34) in which the interpretation of the section "Tina did too" depends on retrieving information from a previous sentence? In fact, given that Elbourne's account obviates the same-sentence constraint, the same mechanisms are at work as in (28) where the elided site and the context clause *are* part of the same sentence. To illustrate this, in (35) the parser retrieves either "leave x's boyfriend" as value for the TheP (i.e., the preferred BV interpretation), or alternatively, it picks up "leave Sarah's boyfriend" as a COR alternative. More specifically, just like in the classic examples of VP-ellipsis—in which the ellipsis and context clause are part of the same sentence—any preference for a dependency type in the first sentence will be inherited by the second sentence in (35). No additional stipulations are necessary and in fact the theoretical problem as described by Frazier and Clifton does not arise—which illustrates yet again the fact that it is important to keep reassessing the interpretation of experimental results in view of theoretical advances[15].

Empirical Issue: Interpretational Preferences in Only-sentences[17]

In addition to a theoretical problem for the BV-preference in VP-ellipsis, Frazier and Clifton also report an empirical problem for so-called *only*-sentences. In order to understand what is at stake in *only*-sentences, consider again the pattern in (32), repeated here with additional material:

(36) Only Alfred thinks (that) he is a good cook.
  a. Only Alfred thinks that Alfred is a good cook (COR)
     Only Alfred (x thinks Alfred is a good cook)
  b. The only person who thinks of himself as a good cook is Alfred. (BV)
     Only Alfred (x thinks that x is a good cook)

Frazier and Clifton conducted a questionnaire study, which shows a strong preference for the (36a) interpretation among the respondents. However, there is a caveat about such off-line studies. They reflect an end-result, but don't give insight in the process itself. As it is, if we wish to interpret their results two questions come up. First, is it just a matter of BV vs. COR, or do other factors play a role? Second, what kind of information does

[15] One might wonder if perhaps even a simpler mechanism might work, namely a preference for an antecedent that is as local as possible. This, however, would not derive the parallelism the construction shows. One of the available options is a "3rd party" reading, as for instance in *John* loves his$_1$ *cat and Peter does love his*$_2$ *cat too*, where *his*$_1$ could be Charles given a suitable context. If *his*$_1$ is Charles, *his*$_2$ has to be as well. This shows that there is a dependency between the two occurrences of *his* that has to be represented in the licensing mechanism.

[17] Frazier and Clifton also discuss another empirical puzzle, based on their experiment 1a, a self-paced reading experiment. In this experiment they compare VP ellipsis internal to a sentence with VP ellipsis across sentences. Sentences (a) and (b) are neutral in the sense that they are easily compatible both with a BV and a COR interpretation, whereas (c) and (d) are biased in favor of a COR interpretation.

a. John thinks it's a good idea to shave himself before he goes to sleep and Andy does too. (Bound/Neutral, one sentence)
b. John thinks it's a good idea to shave himself before he goes to sleep. Andy does too. (Bound/Neutral, two sentences)
c. John thinks it's a good idea to shave himself before he goes to sleep and Anne does too. (Coreferential, one sentence)
d. John thinks it's a good idea to shave himself before he goes to sleep. Anne does too. (Coreferential, two sentences)

The puzzle this experiment raises is that the BV advantage seems to disappear across a sentence boundary as in the (b) and (d) cases. If so, this would suggest that whatever one sees in VP ellipsis is not the manifestation of a unified phenomenon. As already noted, the interpretation of their results is not entirely clear-cut due to the limitations of their design. In this case another complication arises.

In the contrast between these sentence types three factors are involved: i. Reflexivization of shave by himself; ii. Control: assigning a value to PRO in PRO to shave himself; iii. The interpretation of the implicit argument of good idea for x (PRO to shave himself) as either Andy, or John (assuming Anne to be ruled out due to the feature mismatch with himself). The latter constitutes a crucial independent factor, which should have been controlled for, in order for a proper interpretation of this result to be possible. The experiment, then, appears to bear on the interpretation of implicit arguments, rather than on VP ellipsis and the LF-first hypothesis directly.

the language processor have to draw together, to obtain either a BV or a COR interpretation in sentences with *only?*

For a proper understanding of these issues at least the following crucial fact should be taken into account: Across both interpretations the fact that Alfred is happy about his own cooking remains constant. Yet, a full interpretation requires the representation of some sort of "hidden" reference set consisting of everybody but Alfred, or in other words the *contrast set* (e.g., Rooth, 1985). The contrast set, implicitly introduced through the use of the term *only*, behaves differently in a BV reading than in a COR reading: whereas in the BV reading each individual member of the set is not that happy about his own cooking, the contrast set in the COR reading consists of members who think that Alfred's cooking is not very good. Given this, a possible additional factor in a BV or COR preference is how well the hidden contrast set fits the context overall.

Thus, a factor to take into account is that, possibly, the hidden set of the COR reading in the sentences tested by Frazier and Clifton just fits the context better. In fact, Frazier and Clifton presented their sentences without an explicit context. But, in order to interpret *only*-sentences, participants will have to set up a context. Thus, the question is what context they construe.

Crain and Steedman (1985) propose a *Principle of Referential Success,* reflecting that people choose the reading with the fewest "open ends." In view of this, it may well be the case that a strict interpretation is chosen more often in "*only* Alfred thinks he is a good cook" because it is more likely that the sentence is talking about Alfred's cooking, which is explicitly mentioned, than about the cooking of the "entire world." Hence, the lack of context could very well bias participants to a COR interpretation regardless of whether the language processor initially prefers a BV reading or not. It is therefore crucial to properly investigate the role of context, and, where necessary, control for its effects.

In summary, Frazier and Clifton (2000) reported both a theoretical problem and an empirical problem for the LF-only hypothesis—which incorporates the BV preference. We have shown that the theoretical problem with VP-ellipsis is in fact not problematic according to the most recent insights of linguistic theories. The second problem (a COR preference in only-sentences), we argued, required further testing. More specifically, as we will discuss in the next section, it generated the following hypotheses in (37) and a series of experiments testing them (e.g., Koornneef et al., 2006, 2011; Koornneef, 2008, 2010; Cunnings et al., 2014).

(37) Hypotheses
- The language processor initially prefers a BV interpretation.
- Context may then lead to a COR interpretation.
- This (mental) backtracking should be visible in the time course of the process.

Tracking the Time Course of Anaphora Resolution

The hypotheses presented in (37), and the issues raised by Frazier and Clifton regarding sentences containing the only-operator, were addressed by Koornneef et al. (2011) in a questionnaire (to assess the final interpretation of the participants) and an eye-tracking experiment (to track the mental processes preceding this final interpretation). In their study Dutch university students read a series of short texts in 4 versions about 2 story characters of the same gender (e.g., *Lisa* and *Anouk*, see ex. 38).

(38) Example of BV-biased/only-sentence condition
(S1)Lisa en Anouk zijn dol op de muziekzender MTV. (S2) Zij konden hun geluk niet op toen zij mee mochten doen aan het programma "Pimp My Room," waarin hun kamers werden opgeknapt. *(S3) Alleen Lisa vindt dat haar gepimpte kamer klasse heeft.* (S4) Smaken verschillen nu eenmaal.
"(S1). Lisa and Anouk love the music channel MTV. (S2) They were very happy when they were selected for the show 'Pimp My Room,' in which their rooms were redecorated. *(S3) Only Lisa thinks that her pimped room has a touch of class.* (S4) Oh well, each to his own taste."

Each story contained a critical third sentence (S3) that was ambiguous between a sloppy (BV) and strict (COR) interpretation. Moreover, two factors were manipulated in the stimuli. First, the critical sentence was an ambiguous only-sentence (e.g., "Only Lisa thinks that her pimped room has a touch of class.") or, alternatively, an ambiguous VP-ellipsis sentence (e.g., "Lisa thinks that her pimped room has a touch of class, but Anouk does not"). Second, by providing background information in the second sentence about both story characters ("Lisa and Anouk were very happy…") or, alternatively, about only one story character ("Lisa was very happy…"), the context either favored a BV interpretation or a COR interpretation of the ambiguous critical sentence, respectively.

The results of the questionnaire experiment, in which the participants presented their final interpretation of the ambiguous sentence (in addition to providing ratings of story-plausibility and -difficulty) showed that, while using a relatively, simple manipulation and exactly the same critical sentence, readers were more easily biased toward a BV interpretation than toward a COR interpretation. Moreover, contrary to the findings of Frazier and Clifton the context manipulation in the second sentence affected the interpretation of the only-sentences and ellipsis-sentences in the exact same way. Hence, these finding are consistent with the idea that the interpretation of the referential ambiguity in only-sentences and VP-ellipses is driven by the same constraints, which preferable single out a BV interpretation.

The eye-tracking data of the reading experiment of Koornneef et al. (2011) confirmed and extended these results. First of all, the stories in which the interpretation of the ambiguous sentence was biased toward a BV interpretation elicited shorter first pass reading times in the critical VP-ellipsis sections than the stories biased toward a COR interpretation[18]. Furthermore, the reading times for the second sentence (i.e., the sentence that contained the biasing information) also revealed a clear contrast between the COR- and BV-biased stories. In this case the second-pass durations—indicative of re-analysis and repair—were much longer for the COR-biased stories. Interestingly,

[18]Note that these results confirmed the findings of the self-paced reading experiments reported by Frazier and Clifton (2000). Hence, across methodologies and languages there is evidence that readers prefer to assign a sloppy identity to ambiguous elliptic structures.

this was observed for ellipsis- and only-sentences alike, which again suggests that the preference for BV interpretations is not restricted to ellipses, but a general property of the parser.

In all, the results of the offline questionnaire and in particular the online eye-tracking experiment were consistent with the hypotheses as formulated in (37). That is, the readers initially preferred a BV reading, since BV reflected the cheaper option in the processing hierarchy. However, when the larger context forced a COR reading instead, readers reanalyzed the story to change their initial BV reading into the more suitable COR reading. This (mental) backtracking surfaced in the eye-tracking data as longer first-pass reading times near the elided section of the ellipsis sentences and longer second-pass reading times at the biasing second sentence.

In a similar eye-tracking study examining the interplay between BV and COR, Koornneef et al. (2006) showed that the preference of the parser for BV dependencies generalizes beyond ambiguous ellipsis- and only-sentences. They observed that in sentences like (39) containing a quantified antecedent "iedere arbeider" (every worker) in a c-commanding position and a proper name "Paul" in a non-commanding position, readers more easily connected the ambiguous pronoun to the former than to the latter—even when the context preceding the critical sentence clearly mandated the COR reading in which "hij" (he) equaled "Paul."

(39) *Iedere arbeider* die zag dat *Paul* bijna geen energie meer had, vond het heel erg fijn dat *hij* wat eerder naar huis mocht vanmiddag.
"Every worker who noticed that Paul was running out of energy, thought it was very nice that he could go home early this afternoon."

In a more recent eye-tracking study, however, Cunnings et al. (2014) addressed some weaknesses in the stimuli of Koornneef et al. (2006) and failed to replicate the preference for quantified c-commanding antecedents over non-c-commanding proper names. More specifically, in the most relevant experiment of their study (i.e., Experiment 1) Cunnings et al. embedded sentences like (40) in a short discourse and manipulated the gender of the critical pronoun and the preceding proper name[19].

(40) *Every soldier* who knew that *James/Helen* was watching was convinced that *he/she* should wave as the parade passed.

At the critical pronoun and the region immediately following the pronoun they observed longer re-reading and total reading times when the proper name antecedent mismatched in gender with the pronoun. These results, according to Cunnings et al., indicated that readers preferred to connect the pronoun to the linearly closer, yet non-c-commanding antecedent. This would be inconsistent with the PoB framework, since "it fails to support the hypothesis that variable binding relations are computed before coreference assignment."

Although we agree with Cunnings et al. that these results do not provide strong evidence in favor of the PoB approach we disagree with the claim that the results are inconsistent with the approach, for the following reasons. First, in the experiment of Cunnings et al. the individuals [James/Helen in (40)] were not introduced previously—note that this was controlled for in the Koornneef et al. study (2006; see for a detailed discussion Koornneef, 2008). Therefore it is not unlikely that the readers were trying to get further information after the topic shift in the story, and thus tempted to consider a subsequent pronominal as a source of such information. This would be consistent with the fact that the reported differences show up in so-called "later" eye-tracking measures only. Which brings us to a second and arguably more important issue. That is, since the reading time differences become visible in later eye-tracking measures only, the non-c-commanding proper name does not seem to impact the interpretive costs of the pronoun immediately. Hence, instead of ruling out an early preference for BV dependencies over COR dependencies, the findings of Cunnings et al. indicate that COR distractors can influence the interpretive system during later stages of processing—i.e., not unlike the defeasible filter model concerning Principle A (e.g., Sturt, 2003). Crucially, this would be compatible with the PoB approach in which the choice between variable binding and coreference for an ambiguous pronoun is intrinsically free (e.g., Koornneef, 2008).

In all, we do not fully agree with the conclusions as presented by Cunnings et al. (2014), and hence, we maintain our position that there is sufficient evidence for a BV preference—and no convincing evidence against it. Hence, with respect to the good-enough approach (e.g., Ferreira, 2003; Karimi and Ferreira, 2015), the focus of our current contribution, we state that the empirical studies examining bound vs. coreferential dependencies confirm and extend our previous conclusion, where we reported that grammatical operations (such as binding of a SELF-anaphor) are less burdensome for the processor than shallower operations. Again in contrast to what the good-enough approach predicts, the experiments discussed above show that the same holds for binding of a pronominal; the deep variable binding algorithm is less costly than—and preferred over—the shallow top-down driven operation of coreference[20].

Before we present our final assessment of the good-enough approach in the domain of anaphoric dependencies, however, we should address some interesting suggestions of Cunnings et al. (2014) as to how their results can be related to more general architectural issues. First, they observe that a recurrent issue, highly relevant for the bound variable vs. coreferential (or grammatical vs. extra-grammatical) distinction, is the role of structure-based vs. unconstrained cue-based memory retrieval mechanisms (see e.g., Dillon, 2014; Jäger et al., 2015a, for recent overviews of this issue). Second (and somewhat related), they suggest that their results are more easily explained with a uni-modular approach as in Heim (2007), than with the multi-modular architecture assumed in the PoB model. These two architectural issues will be addressed in more detail below.

[19]The interpretation of their Experiment 2 is not entirely straightforward due to the presence of two c-commanding potential antecedents, but the findings of this experiment seem to be consistent with the predictions of the PoB model.

[20]Quite interestingly, many of the facts discussed in Karimi and Ferreira (2015) are consistent with the idea that there is a cost associated with accessing discourse. The essence of shallow-R processing of anaphoric dependencies appears to consist of foregoing or postponing the access to discourse, leaving pronominals unvalued.

Structure-Based vs. Unconstrained Cue-Based Retrieval

The PoB economy ranking in relation to shallow vs. deep processing is by no means the only issue that arises in the field of anaphor processing. For example, by now an important recurrent issue—although to some extent orthogonal to the economy issue—is what kind of retrieval mechanism promotes anaphor resolution. More specifically, based on a growing body of literature, Cunnings et al. (2014) distinguish two theoretically plausible ways in which the antecedent of a linguistic element can be retrieved from (working) memory. As a first possibility, a serial search mechanism is proposed in which the text representation is searched in a step-by-step manner until the proper antecedent for an anaphor has been located. A qualitatively different search (or retrieval) mechanism is based on the idea of a content-addressable memory (CAM) architecture (Lewis et al., 2006). In the latter type of memory systems, previously stored information can be accessed directly by the use of certain features as retrieval cues.

Cunnings et al. (2014; see also Jäger et al., 2015a,b) make the interesting conjecture that a specific instantiation of a serial search mechanism could be a structure-based retrieval mechanism in which syntactic tree-configurational information (e.g., c-command) guides the retrieval process. That is, in these type of systems "the priority in which antecedents are retrieved is dependent upon their relative position in the search path" (pp. 42) which would be compatible with an architecture assuming a BV preference. In contrast, CAM-like, unconstrained cue-based retrieval assumes that all available cues (e.g., gender, number, person, animacy, etc.) are used immediately (and in parallel) to retrieve an anaphor's antecedent. This system allows for more flexibility as structural constrains do not have a privileged status and, hence, COR interpretations of (reflexive) pronominals are also considered immediately—i.e., not subsequent to BV interpretations.

Cunnings et al. (2014) claim that the results of their eye-tracking experiments favor the latter cue-based approach, as recency (or linear proximity) of the antecedent seemed to guide the resolution process of a pronoun, rather than the structural notion of c-command. Indirectly, then, one could state that there is no solid experimental evidence to maintain a distinction between variable binding and coreference (cf. our discussion on uni-modular vs. multi-modular architectures below). Moreover, it would imply that the same cue-based memory mechanisms underlying the construction of a range of other (syntactic) dependencies—such as filler-gap dependencies (McElree et al., 2003), subject-verb dependencies (Van Dyke and Lewis, 2003; Van Dyke and McElree, 2006; Van Dyke, 2011, 2007; Wagers et al., 2009; Dillon et al., 2013), the licensing of negative-polarity items (Vasishth et al., 2008) and verb-phrase ellipsis (Martin and McElree, 2008)—are responsible for determining the proper antecedent for (reflexive) pronominals.

Whether cue-based memory retrieval, however, is indeed the most valid way to describe anaphoric processing is hotly debated still. For example, Dillon (2014) shows in a very systematic overview that reflexives are relatively immune to so-called retrieval interference, a property that would set them apart from superficially similar syntactic dependencies like subject–verb agreement. This conclusion in turn, is disputed by Jäger et al. (2015a) who conducted reading time experiments on German and Swedish reflexives, and *did* observe occurrences of retrieval interference as predicted by the cue-based approach—and as they claim, not by the structure-based approach.

Hence, at this point in time we are simply not in the position to single out a unique framework as the correct approach. In fact, in the case of anaphora it might well be true that both types of memory retrieval systems are somehow involved. For one thing, although binding dependencies are often discussed in terms of c-command, this certainly does not entail that the formation of logical form representations should be considered to be blind to cues such gender and number. Hence a possible, and in fact very plausible, outcome is that the antecedents for (bound) pronouns are determined by means of a system that combines structure- and cue-based search algorithms, with their respective roles depending on timing. For instance, one might expect intrusion effects at a stage before the final structure is established. In all, the precise nature of the interplay between c-command vs. morpho-syntactic cues is an important issue that must be left for future research (but note that coding a tree-configurational relation as a cue for a CAM-like system is not as straightforward as coding gender and number; see Jäger et al., 2015a, footnote 4).

Albeit in a different way, this latter question also surfaces in the second architectural issue raised by Cunnings et al. (2014). That is, incorporating c-command as a "normal" cue in a CAM retrieval system, or alternatively, setting it apart as a qualitatively different cue, can ultimately be interpreted as a debate on uni-modular vs. multi-modular approaches to anaphor resolution.

Uni-Modular vs. Multi-Modular Architectures

A very fundamental issue raised by Cunnings et al. (2014), concerns the (uni)-modular architecture of the anaphoric system. That is, in contrast to the PoB framework (in which at least three different modules/algorithms are assumed to underlie anaphora interpretation) they follow Heim (2007) who, they claim, puts forward a uni-modular approach. However, we feel that their interpretation of Heim's proposal on uni-modularity is less straightforward than they assume.

First, Heim's discussion is limited to condition B, and the status of Reinhart's Rule I. It does not address condition A, which uncontroversially is syntactic. So, even if Heim's endeavor works for condition B, binding theory as a whole would still minimally be "bi-modular."

Second, Heim does not include the interpretation of proper names and other referential expressions in her discussion. But, even in her system, one must assume that these are directly interpreted as some individual in the discourse—but of course, relative to context. This interpretation strategy, however, must also be available for certain uses of pronouns. Just like we can start a story with *Helen was watching the parade with a feeling of disgust. Suddenly* ….where we are introducing a discourse individual and slowly building a character while reading on, we can start a story with *She was watching the parade with disgust. Suddenly…* and again we will be introducing a discourse individual and slowly building a character. It seems to us that

there is no independent ground to treat the reference assignment differently in these cases. If so, not all cases of pronominal interpretation will fall under the binding strategy Heim proposes. Hence, whatever the division of labor in other cases, no truly uni-modular model for this domain will result in the end.

Heim doesn't discuss this issue. But if one looks carefully, one sees that what she achieves is tantamount to building Reinhart's Rule I into the binding conditions. Given that she set out to retain the core of Reinhart's insight, it is not surprising, then, that it surfaces in the details of the formulation of condition B. In fact what her system does is generalize over the "worst case scenario." The difference between binding and co-valuation shows up in the explicit role of context in the latter, but not in the former. This is interesting by itself, since from a processing perspective, this would make it quite unexpected for co-valuation to require fewer resources than binding. But it also shows that the core of the contrast between binding and co-valuation is in fact retained in her system.

Note furthermore that Heim's unification program is based on the idea that condition B is essentially semantic. However, as shown in Volkova and Reuland (2014), this idea cannot be maintained in view of languages with locally bound pronominals. Such cross-linguistic variation shows that there must be a syntactic component in condition B (see Reuland, 2011a, and Volkova and Reuland, 2014, for further evidence that condition B is in fact not a unified phenomenon). Pronoun resolution in such languages [as for instance Frisian, or (Tegi) Khanty] has not yet been studied experimentally to our knowledge. Such experiments could shed further light on the way interpretive dependencies are processed, and more specifically, on the contrasting economy rankings and its relation to shallow and deep processing as proposed in the good-enough and PoB frameworks.

This brings us back to the issue we started out with, and in fact to a conclusion[21]

CONCLUSION

As part of our more general goal of reassessing the interpretation of experimental results in view of the ongoing advances made in theoretical linguistics and psycholinguistics, the main focus of the current contribution was to evaluate the core assumptions of the good-enough framework as proposed by Ferreira and colleagues (e.g., Ferreira, 2003; Ferreira and Patson, 2007; Karimi and Ferreira, 2015). We structured our discussion around a recent elaboration of the good-enough approach (Karimi and Ferreira, 2015) in which an explicit distinction is being made between "deep" bottom-up syntactic algorithms and "shallow" top-down semantic/discourse operations. Crucially, given the presumed complexity of syntactic algorithms, the latter type of (extra-grammatical) heuristics should be preferred, thereby inducing good-enough representations of an utterance or text.

As it turned out, one of the key-notions in the discussion had to be reassessed. That is, we proposed that one must make a distinction between *shallow-TD* processing as a top-down process, and *shallow-R* processing as involving a reduced input (see e.g., Stewart et al., 2007). Taking this into account, the conclusion in terms of the *shallow equivalence* and the *shallow advantage* assumptions (cf. Principle 1 and 2 in Karimi and Ferreira, 2015) as formulated at the outset of this contribution are straightforward and simple. First, in the domain of anaphoric dependencies the equivalence assumption does not hold. There are binding dependencies whose interpretation cannot even be approximated by shallow-TD procedures. Second, and perhaps for current purposes more importantly, we reviewed a variety of experiments bearing on a purported shallow-TD advantage. None of the experiments provided support for such an advantage. Rather the opposite is the case: in the domain of anaphoric dependencies deep algorithmic computations are preferred over shallow-TD interpretational processes. Such a preference not only shows up in the comparison between syntax and what one may broadly call the interpretive system, but also *within* the latter system, i.e., between deep, structure-based (variable binding), and shallower context-based (coreference) interpretive procedures.

There is one important proviso: as becomes clear from the discussion (e.g., regarding Heim, 2007) context-based interpretive procedures may in fact require more computation than meets the eye. Hence, properly considered, they may not be as shallow as they prima facie appear to be. Perhaps, then, they are more costly because they, at least in some cases, require more sophisticated computations. But if this is so, this casts doubt on the very idea that there are truly shallow procedures. Such shallow procedures may well be no more than illusory effects that arise if some material is not admitted into the buffer. Therefore, we submit the bold claim that, until proponents of the existence of shallow procedures offer precise and falsifiable descriptions, Occam's razor requires us to treat them as just that: illusions.

AUTHOR CONTRIBUTIONS

All authors listed, have made substantial, direct, and intellectual contribution to the work, and approved it for publication.

FUNDING

This work was supported by an NWO (Netherlands Organization for Scientific Research) Veni grant [grant number 275-89-012] awarded to AK.

ACKNOWLEDGMENTS

We are very much indebted to the organizers and participants of the GLOW workshop on the Timing of Grammar

[21]Many further interesting issues about the processing of interpretive dependencies arise. One factor that sets the processing of pronominals apart from the processing of SELF-anaphors, is that pronominals don't have to be bound, whereas SELF-anaphors in non-exempt positions and simplex anaphors must be bound. In argument positions SELF-anaphors and bound pronominals are in complementary distribution, but not in locative and directional PPs. It would be interesting to investigate the effect of such non-complementarity. Also non-local binding of simplex anaphors in Dutch, German, and Mainland Scandinavian languages raises interesting issues. They must be bound within the sentence – although their domain varies. Especially in Scandinavian they allow a choice of antecedents, and in the non-local domain they are not in complementary distribution with bound pronominals. The question is, then, how precisely these factors show up in the processing of these elements.

.

(Potsdam 2012) for their comments and to the editors for their patience. We are very grateful to the reviewers for their careful and constructive comments which stimulated us to considerably sharpen our argumentation. Above all we would like to thank Loes Koring for reading and commenting on an early draft, and extensive conversations that provided the impetus for choosing this particular focus.

REFERENCES

Avrutin, S. (1994). *Psycholinguistic Investigations in the Theory of Reference*. Ph.D. thesis, MIT. Cambridge, MA.

Avrutin, S. (1999). *Development of the Syntax-Discourse Interface*. Dordrecht: Kluwer.

Baauw, S. (2002). *Grammatical Features and the Acquisition of Reference: A Comparative Study of Dutch and Spanish*. New York, NY: Routledge.

Badecker, W., and Straub, K. (2002). The processing role of structural constraints on the interpretation of pronouns and anaphors. *J. Exp. Psychol.* 28. 748–769. doi: 10.1037/0278-7393.28.4.748

Bosch, P. (1980). "The modes of pronominal reference and their constraints," in *Pronouns and Anaphora. Papers from the Parasession on Anaphora of the 16th Regional Meeting of the Chicago Linguistic Society*, eds J. Kreiman and A. Ojeda (Chicago, IL: Chicago Linguistic Society, University of Chicago), 64–78.

Bosch, P. (1983). *Agreement and Anaphora - A Study of the Roles of Pronouns in Discourse and Syntax*. London; New York, NY: Academic Press.

Carter, D. (1985). *A Shallow Processing Approach to Anaphor Resolution, TR-88.* Cambridge: University of Cambridge Computer Laboratory.

Chien, Y., and Wexler, K. (1991). Children's knowledge of locality conditions in binding as evidence for the modularity of syntax and pragmatics. *Lang. Acquis.* 1, 225–295.

Chomsky, N. (1981). *Government and Binding*. Dordrecht: Foris.

Chomsky, N. (1986). *Knowledge of Language*. New York, NY: Praeger.

Chomsky, N. (1995). *The Minimalist Program*. Cambridge, MA: MIT Press.

Conroy, A., Takahashi., E., Lidz, J., and Phillips, C. (2009). Equal treatment for all antecedents: how children succeed with Principle B. *Linguist. Inq.* 40, 446–486. doi: 10.1162/ling.2009.40.3.446

Crain, S., and Steedman, M. (1985). "On not being led up to garden path: the use of context by the psychological parser," in *Natural Language Parsing*, eds D. Dowty, L. Karttunnen, and A. Zwicky (Cambridge, UK: Cambridge University Press), 320–358.

Cunnings, I., and Felser, C. (2013). The role of working memory in the processing of reflexives. *Lang. Cognit. Process.* 28, 188–219. doi: 10.1080/01690965.2010.548391

Cunnings, I., Patterson., C., and Felser, C. (2014). Variable binding and coreference in sentence comprehension: evidence from eye movements. *J. Mem. Lang.* 71, 39–56 doi: 10.1016/j.jml.2013.10.001

Dąbrowska, E. (2012). Different speakers, different grammars: Individual differences in native language attainment. *Linguist. Approaches Biling.* 2, 219–253. doi: 10.1075/lab.2.3.01dab

De Lange, J. (2008). *Article Omission in Headlines and Child Language*: *A Processing Approach*. Utrecht: LOT International Dissertation Series.

Dillon, B. (2014). Syntactic memory in the comprehension of reflexive dependencies: an overview. *Lang. Linguist. Compass* 8, 171–187. doi: 10.1111/lnc3.12075

Dillon, B., Mishler, A., Sloggett, S., and Phillips, C. (2013). Contrasting intrusion profiles for agreement and anaphora: experimental and modeling evidence. *J. Mem. Lang.* 69, 85–103. doi: 10.1016/j.jml.2013.04.003

Elbourne, P. (2005). On the acquisition of Principle B. *Linguist. Inq.* 36, 333–365. doi: 10.1162/0024389054396908

Elbourne, P. (2008). Ellipsis sites as definite descriptions. *Linguist. Inq.* 39, 191–220. doi: 10.1162/ling.2008.39.2.191

Felser, C., and Cunnings, I. (2012). Processing reflexives in English as a second language: The role of structural and discourse-level constraints. *Appl. Psycholinguist.* 33, 571–603. doi: 10.1017/S0142716411000488

Felser, C., Sato, M., and Bertenshaw, N. (2009). The on-line application of binding Principle A in English as a second language. *Bilingualism Lang. Cognit.* 12, 485–502.

Ferreira, F. (2003). The misinterpretation of noncanonical sentences. *Cogn. Psychol.* 47, 164–203 doi: 10.1016/S0010-0285(03)00005-7

Ferreira, F., and Patson, N. D. (2007). The "good enough" approach to language comprehension. *Lang. Linguist. Compass* 1, 71-83. doi: 10.1111/j.1749-818X.2007.00007.x

Frazier, L., and Clifton, C. (2000). On bound variable interpretations: the LF-only hypothesis. *J. Psycholinguist. Res.* 29, 125–139. doi: 10.1023/A:1005136826534

Gigerenzer, G. (2000). *Adaptive Thinking: Rationality in the Real World*. New York, NY: Oxford University Press.

Gigerenzer, G., Todd, P. M., and A. B. C. Research Group (1999). *Simple Heuristics that Make us Smart*. New York, NY: Oxford University Press.

Grodzinsky, Y. (1995). A restrictive theory of agrammatic comprehension. *Brain Lang.* 51, 26–51 doi: 10.1006/brln.1995.1039

Grodzinsky, Y., and Reinhart, T. (1993). The Innateness of binding and coreference.Linguist. Inq. 24, 69–101.

Hankamer, J., and Sag, I. (1976). Deep and surface anaphora. *Linguist. Inq.* 7, 391–426.

Heim, I. (1982). *The Semantics of Definite and Indefinite Noun Phrases*. Ph.D. thesis, UMass, Amherst.

Heim, I. (2007). "Forks in the Road to Rule I. In NELS 38," in *Proceedings of the Thirty-Eigth Annual meeting of the North East Linguistic Society* (Ottawa: KS).

Jäger, L. A., Benz, L., Roeser, J., Dillon, B. W., and Vasishth, S. (2015a). Teasing apart retrieval and encoding interference in the processing of anaphors. *Front. Psychol.* 6:506. doi: 10.3389/fpsyg.2015.00506

Jäger, L. A., Engelmann, F., and Vasishth, S. (2015b). Retrieval interference in reflexive processing: experimental evidence from Mandarin, and computational modeling. *Front. Psychol.* 6:617. doi: 10.3389/fpsyg.2015.00617

Karimi, H., and Ferreira, F. (2015). Good-enough linguistic representations and online cognitive equilibrium in language processing. *Q. J. Exp. Psychol.* doi: 10.1080/17470218.2015.1053951. [Epub ahead of print].

Koornneef, A. (2010). "Looking at anaphora: the psychology reality of the primitives of binding model," in *The Linguistic Enterprise: From Knowledge of Language to Knowledge of Linguistics*, eds T. Everaert, H. Lentz, Ø. de Mulder, A. Nilsen, and A. Zondervan (Amsterdam: John Benjamins), 141–166.

Koornneef, A. W. (2008). *Eye-Catching Anaphora*. Utrecht: LOT International Dissertation Series.

Koornneef, A. W., Avrutin, S., Wijnen, F., and, Reuland, E. (2011). Tracking the preference for boound-variable dependencies in ambiguous ellipses and only-structures in experiments at the interfaces. *Syntax Semant.* 37. doi: 10.1163/9781780523750_004

Koornneef, A. W., Wijnen, F. N. K., and Reuland, E. J. (2006). "Towards a modular approach to anaphor resolution," in *Ambiguity in Anaphora Workshop Proceedings*, eds R. Artstein and M. Poesio (Malaga: ESSLI), 65–72.

Koring, L., Mak, P., and Reuland, E. (2012). The time course of argument reactivation revealed: using the visual world paradigm. *Cognition* 123, 361–379. doi: 10.1016/j.cognition.2012.02.011

Kostić, A. (1991). Informational approach to processing inflected morphology: standard data reconsidered. *Psychol. Res.* 53, 62–70. doi: 10.1007/BF00867333

Kostić, A. (1995). "Informational load constraints on processing inflected morphology," in *Morphological Aspects of Language Processing*, ed L. Feldman (Hillsdale, NJ: Erlbaum), 317–345.

Kuperberg, G. R. (2007). Neural mechanisms of language comprehension: Challenges to syntax. *Brain Res.* 1146, 23–49. doi: 10.1016/j.brainres.2006.12.063

Lewis, R. L., Vasishth, S., and Van Dyke, J. A. (2006). Computational principles of working memory in sentence comprehension. *Trends Cognit. Sci.* 10, 44–54. doi: 10.1016/j.tics.2006.08.007

Manika, S. (2014). *Understanding bit by bit: Information theory and the Role of Inflections in Sentence Processing*. Utrecht: LOT International Dissertation Series.

Manika, S., Avrutin, S., and Reuland, E. (2014). "The effect of verb's inflectional entropy on the processing of reflexive objects," in *Selected Papers from the 4th UK Cognitive Linguistics Conference,* eds G. Rundblad, A. Tytus,

O. Knapton and C. Tang (London: UK Cognitive Linguistics Association), 213–227.

Martin, A. E., and McElree, B. (2008). A content-addressable pointer mechanism underlies comprehension of verb-phrase ellipsis. *J. Mem. Lang.* 58, 879–906. doi: 10.1016/j.jml.2007.06.010

McElree,B., Foraker, S., and Dyer, L. (2003). Memory structures that subserve sentence comprehension. *J. Mem. Lang.* 48, 67–91. doi: 10.1016/S0749-596X(02)00515-6

Merchant, J. (2001). *The Syntax of Silence*. Oxford: Oxford University Press.

Merchant, J. (2008). "Variable island repair under ellipsis," in *Topics in Ellipsis*, eds Kyle Johnson (Cambridge, UK: Cambridge University Press), 132–153.

Nicol, J., and Swinney, D. (1989). The role of structure in coreference assignment during sentence comprehension. *J. Psycholinguist. Res.* 18, 5–19. doi: 10.1007/BF01069043

Partee, B. H. (1978). "Bound variables and other anaphors," in *Theoretical Issues in Natural Language Processing 2 (TINLAP-2),* ed David L. Waltz (Urbana, IL: University of Illinois), 79–85 (Reprinted in Partee, Barbara H. 2004. Compositionality in Formal Semantics: Selected Papers by Barbara H. Partee. Oxford: Blackwell Publishing, 110–121). Available online at: https://udrive.oit.umass.edu/partee/Partee78TINLAP.pdf

Reinhart, T. (1983). *Anaphora and Semantic Interpretation*. London: Croom Helm.

Reinhart, T. (2002). The theta system: an overview. *Theor. Linguist.* 28. 229–290. doi: 10.1515/thli.28.3.229

Reinhart, T. (2006). *Interface Strategies: Reference Set Computation*. Cambridge, MA: MIT Press.

Reinhart, T., and Reuland, E. (1991). "Anaphors and logophors: an argument structure perspective," in *Long Distance Anaphora*, eds J. Koster and E. Reuland (Cambridge: Cambridge University Press), 283–321.

Reinhart, T., and Reuland, E. (1993). Reflexivity. *Linguist. Inq.* 24, 657–720.

Reinhart, T., and Siloni, T. (2005). The lexicon-syntax parameter: reflexivization and other arity operations. *Linguist. Inq.* 36, 389–436. doi: 10.1162/0024389054396881

Reuland, E. (2001). Primitives of binding. *Linguist. Inq.* 32, 439–492. doi: 10.1162/002438901750372522

Reuland, E. (2011a). *Anaphora and Language Design*. Cambridge, MA: MIT Press.

Reuland, E. (2011b). "Syntax and interpretation systems: how is their labour divided," in *The Oxford Handbook of Linguistic Minimalism*, eds Boeckx and Cedric (Oxford: Oxford University Press), 377–395.

Reuland, E. (2012). How linguistic performance may differ across speakers. *Linguist. Approaches Bilingualism* 2, 287–292. doi: 10.1075/lab.2.3.09reu

Rooth, M. (1985). *Association with Focus*. Ph.D. thesis, UMass, Amherst.

Slobin, D. I. (1966). Grammatical transformations and sentence comprehension in childhood and adulthood. *J. Verbal Learn. Verbal Behav.* 5, 219–227. doi: 10.1016/S0022-5371(66)80023-3

Stewart., A. J., Holler, J., and Kidd, E. (2007). Shallow processing of ambiguous pronouns: evidence for delay. *Q. J. Exp. Psychol.* 60, 1680–1696. doi: 10.1080/17470210601160807

Sturt, P. (2003). The time-course of the application of binding constraints in reference resolution. *J. Mem. Lang.* 48, 542–562. doi: 10.1016/S0749-596X(02)00536-3

Townsend, D. J., and Bever, T. G. (2001). *Sentence Comprehension: The Integration of Habits and Rules*. Cambridge, MA: MIT Press.

Travis, L. (1984). *Parameters and Effects of Word Order Variation*. Ph.D. thesis, MIT.

Van Dyke, J. A. (2007). Interference effects from grammatically unavailable constituents during sentence processing. *J. Exp. Psychol. Learn. Mem. Cognit.* 33, 407–430. doi: 10.1037/0278-7393.33.2.407

VanDyke, J. A. (2011). Cue-dependent interference in comprehension. *J. Mem. Lang.* 65, 247–263. doi: 10.1016/j.jml.2011.05.002

Van Dyke, J., and Lewis, R. L. (2003). Distinguishing effects of structure and decay on attachment and repair: a cue-based parsing account of recovery from misanalysed ambiguities. *J. Mem. Lang.* 49, 285–316. doi: 10.1016/S0749-596X(03)00081-0

Van Dyke, J. A. and McElree, B. (2006). Retrieval interference in sentence comprehension. *J. Mem. Lang.* 55, 157–166. doi: 10.1016/j.jml.2006.03.007

Vasic, N., Avrutin, S., and Ruigendijk, E. (2006). Interpretation of pronouns in VP-ellipsis constructions in Dutch Broca's and Wernicke's aphasia. *Brain Lang.* 96, 191–206. doi: 10.1016/j.bandl.2005.04.003

Vasishth, S., Brüessow, S., Lewis, R. L., and Drenhaus, H. (2008). Processing polarity: how the ungrammatical intrudes on the grammatical. *Cognit. Sci.* 32, 685–712. doi: 10.1080/03640210802066865

Volkova, A., and Reuland, E. (2014). Reflexivity without reflexives? *Linguist. Rev.* 31, 587–633. doi: 10.1515/tlr-2014-0012

Von der Malsburg, T., and Vasishth, S. (2013). Scanpaths reveal syntactic underspecification and reanalysis strategies. *Lang. Cogn. Process.* 28, 1545–1578. doi: 10.1080/01690965.2012.728232

Wagers, M. W., Lau, E. F., and Phillips, C. (2009). Agreement attraction in comprehension: representations and processes. *J. Mem. Lang.* 61, 206–237. doi: 10.1016/j.jml.2009.04.002

Xiang, M., Dillon, B., and Phillips, C. (2009). Illusory licensing effects across dependency types: ERP evidence. *Brain Lang.* 108, 40–55. doi: 10.1016/j.bandl.2008.10.002

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer, CP, and handling Editor declared their shared affiliation, and the handling Editor states that the process nevertheless met the standards of a fair and objective review.

Copyright © 2016 Koornneef and Reuland. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 30 September 2016
doi: 10.3389/fpsyg.2016.01434

The Localization of Long-Distance Dependency Components: Integrating the Focal-lesion and Neuroimaging Record

Maria M. Piñango[1,2]*, *Emily Finn*[1,2], *Cheryl Lacadie*[2] and *R. Todd Constable*[2]

[1] Language and Brain Lab, Department of Linguistics, Yale University, New Haven, CT, USA, [2] Interdepartmental Neuroscience Program, Magnetic Resonance Research Center, Yale University, New Haven, CT, USA

In the sentence "The captain who the sailor greeted is tall," the connection between the relative pronoun and the object position of *greeted* represents a long-distance dependency (LDD), necessary for the interpretation of "the captain" as the individual being greeted. Whereas the lesion-based record shows preferential involvement of only the left inferior frontal (LIF) cortex, associated with Broca's aphasia, during real-time comprehension of LDDs, the neuroimaging record shows *additional* involvement of the left posterior superior temporal (LPST) and lower parietal cortices, which are associated with Wernicke's aphasia. We test the hypothesis that this localization incongruence emerges from an interaction of memory and linguistic constraints involved in the real-time implementation of these dependencies and which had not been previously isolated. Capitalizing on a long-standing psycholinguistic understanding of LDDs as the workings of an active filler, we distinguish two linguistically defined mechanisms: *GAP-search*, triggered by the retrieval of the relative pronoun, and *GAP-completion*, triggered by the retrieval of the embedded verb. Each mechanism is hypothesized to have distinct memory demands and given their distinct linguistic import, potentially distinct brain correlates. Using fMRI, we isolate the two mechanisms by analyzing their relevant sentential segments as separate events. We manipulate LDD-presence/absence and *GAP-search* type (direct/indirect) reflecting the absence/presence of intervening islands. Results show a *direct GAP-search*—LIF cortex correlation that crucially excludes the LPST cortex. Notably, *indirect GAP-search* recruitment is confined to supplementary-motor and lower-parietal cortex indicating that *GAP* presence alone is not enough to engage predictive functions in the LIF cortex. Finally, *GAP-completion* shows recruitment implicating *the dorsal pathway* including: the supplementary motor cortex, left supramarginal cortex, precuneus, and anterior/dorsal cingulate. Altogether, the results are consistent with previous findings connecting *GAP-search*, as we define it, to the LIF cortex. They are not consistent with an involvement of the LPST cortex in any of the two mechanisms, and therefore support the view that the LPST cortex is not crucial to LDD implementation. Finally, results support neurocognitive architectures that involve the dorsal pathway in LDD resolution and that distinguish the memory commitments of the LIF cortex as sensitive to specific language-dependent constraints beyond phrase-structure building considerations.

Keywords: left inferior frontal cortex, Broca's and Wernicke's aphasia, supplementary motor area, precuneus, long-distance dependencies, sentence comprehension, working memory, attention

OPEN ACCESS

Edited by:
Colin Phillips,
University of Maryland, USA

Reviewed by:
William Matchin,
University of Maryland, USA
Jonathan Brennan,
University of Michigan, USA

***Correspondence:**
Maria M. Piñango
maria.pinango@yale.edu

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: *02 July 2016*
Accepted: *07 September 2016*
Published: *30 September 2016*

Citation:
Piñango MM, Finn E, Lacadie C and Constable RT (2016) The Localization of Long-Distance Dependency Components: Integrating the Focal-lesion and Neuroimaging Record. Front. Psychol. 7:1434. doi: 10.3389/fpsyg.2016.01434

1. INTRODUCTION

A long-distance or filler-gap dependency (LDD) is a syntactico-semantic relation between a pronominal element and a syntactically licensed position, or GAP, in an embedded clause. The LDD is thus the linguistic device that allows the pronominal element to be interpreted within the embedded clause. In the English sentence "*The captain$_k$ [who$_{k/j}$ the sailor predicted that the weather would frighten (GAP)$_j$] smiled.*" the LDD is the connection between the relative pronoun and the object position of *frighten*, to which the semantic role of *frightenee* is assigned. LDDs have traditionally provided a window to explore the interaction between lexico-semantic and syntactic mechanisms involved in sentence composition, and have thus represented a rich space for neurolinguistic and psycholinguistic investigation. In LDDs, these mechanisms are specifically observed in the interpretation of the relative pronoun both as the object of the embedded verb (e.g., *the frightenee*) and as the coreferent to the head noun antecedent (e.g., *The captain*), mechanisms that are presumably grounded not only in fundamental properties of sentence composition such as argument structure licensing and discourse linking but also in the neurological properties of the linguistic subsystems that support those properties (e.g., Frazier et al., 1983; Frazier and Clifton, 1989; Grodzinsky, 1989; Swinney et al., 1989; Swinney and Zurif, 1995; Gibson, 1998; Grodzinsky, 2000; Phillips, 2003; Avrutin, 2006).

From a neurolinguistic perspective, LDD implementation also allows us to investigate how the interaction between sentence composition and memory should be understood, as well as what the cortical distribution of this interaction should be. Interpretation of the relative pronoun is, after all, expected to place significant demands on the memory system: the pronoun must be held in memory while the intervening syntactic and semantic material is parsed (in the present case "*that the weather would*"). The presence of intervening material taxes the processing system (e.g., King and Kutas, 1995; Cooke et al., 2002; Fiebach et al., 2002; Santi and Grodzinsky, 2012; Santi et al., 2015) and is subject to aging effects (Zurif et al., 1995). So, understanding the cortical distribution of these dependencies gives us insight into the basic commitments that any neurocognitive model of language must allow with respect to sentence composition in addition to the interactions of sentence composition with other components of cognition, most notably memory.

The record on LDD comprehension reveals a long-standing incongruence regarding the language processing commitments of the left inferior frontal (LIF) cortex: lesion studies show that in contrast to Wernicke's patients and patients with lesions in the right hemisphere homolog of Broca's area, Broca's patients fail to implement LDDs in a normal fashion during real-time comprehension. Specifically, these subjects fail to show normal implementation of the "GAP-filling" effect: the reactivation of the antecedent (i.e., the entity coreferent with the relative pronoun) at the position of the GAP (e.g., Zurif et al., 1993; Swinney et al., 1996; Grodzinsky et al., 1999; Grodzinsky, 2000; Burkhardt et al., 2003; Love et al., 2008). Given the localization value of Broca's and Wernicke's aphasia, this pattern of performance is taken to indicate that LDDs demand the workings of the LIF cortex and, crucially, do not depend on the workings of the left posterior superior temporal (LPST) cortex. By contrast, neuroimaging work has shown equal engagement of the LIF cortex *and* the LPST cortex for the implementation of the same dependencies (e.g., Stromswold et al., 1996; Cooke et al., 2002; Fiebach et al., 2002; Ben-Shachar et al., 2003, 2004; Friederici et al., 2003; Grodzinsky and Friederici, 2006; Santi and Grodzinsky, 2008).

We take both sets of results– lesion- and neuroimaging-based– to be valid and on that basis propose that together they provide complementary observations about LDDs and the neurocognitive resources that support them. Specifically, we hypothesize that one crucial property of LDD implementation–*GAP-search*–relies on the workings of the LIF cortex, as the lesion-based record shows. This leaves open the question of the role of the LPST cortex reported in the neuroimaging record. In this respect we test the hypothesis that such LPST cortical recruitment would not be connectable to the implementation of *GAP-search*; and may be instead implicated in *GAP-completion*, a local, lexically-driven process fundamental to all sentence composition. To this end, we isolate the neurocognitive factors underpinning LDD comprehension on the basis of an analysis of relative pronouns that connects to parallel, incremental left-to-right structure-building mechanisms with potential neurocognitive relevance. Using fMRI, we examine the timing and cortical commitments of the interaction of these mechanisms. We conclude with a discussion of the implications of these findings for the lesion vs. imaging "mismatch," and in the context of current neurocognitive models for our understanding of the LIF cortex as a "language" area.

1.1. The Structural and Processing Properties of Long-Distance Dependencies

The purpose of this section is to present the linguistic structure for long-distance dependencies (LDDs) that supports their real-time processing implementation. This structure is therefore the basis for the definitions of the processing mechanisms of *GAP-search* and *GAP-completion*, which operationalize the dependency in neurocognitive terms[1]. In English, long-distance dependencies prototypically emerge in relative clause and *wh*-question formation. In the case of relative clauses, they involve three main elements: the antecedent, the relative pronoun, and the GAP. The antecedent is the denotation of the head noun

[1]This linguistic description captures the consensus among a variety of syntactic approaches, e.g., Government Binding/Minimalism, Head-Driven Phrase Structure Grammar, Lexical Functional Grammar, and Simpler Syntax, among others, that LDDs are grounded on two organizational properties of language: (1) the possibility to "package" the semantic and syntactic local conditions of the relative pronoun as lexicalized content in the form of subcategorization and/or selectional restrictions, and (2) the possibility of a GAP, a phonologically empty lexico-syntactic entity whose purpose is to instantiate the lexical requirements of the embedded verb; requirements that are expressed in the form of argument structure and subcategorization specifications. These are fundamental and widely accepted properties of the language system. The description presented here is therefore compatible with any representational analysis of relative pronouns that incorporates these two properties (see Culicover and Jackendoff, 2005, for extensive discussion of the syntax-semantics interactions in LDDs and the assumptions that lead the various approaches in question to favor one specific implementation over another).

of the noun phrase containing the relative clause [*captain* in (1) below]. The RELPRO (which may be phonologically empty in English) is the entity that semantically links the antecedent and the GAP [*who* in (1) below]. The RELPRO occupies what we would call a "non-canonical" position, a position that does not receive direct semantic role assignment by a predicate, and therefore does not receive direct interpretation with respect to the proposition associated with the embedded clause. This interpretation is provided instead through the dependency it forms with the GAP. The GAP, in turn, is a hypothesized phonologically empty syntactically valid place-holder of the "displaced" relative pronoun which receives a semantic role by virtue of its grammatical function within the embedded clause. (1) below illustrates the relation between the GAP to which the semantic role of "experiencer" is assigned and the denotation of the head noun *captain* (the antecedent):

(1) The captain$_{antecedent}$ [who the sailor predicted that the weather would frighten ("the captain")$_{GAP}$] turned back to port.

The relation between the antecedent and the GAP is mediated by the relative pronoun (RELPRO). The RELPRO holds a coreference relation with the antecedent. And it is this coreference relation between the RELPRO and the antecedent that allows the antecedent to be interpreted as a participant in the proposition associated with the embedded clause, i.e., *the sailor predicted that the weather would frighten* ***the captain***. Establishing an LDD therefore means connecting, on the one hand, the antecedent and the RELPRO and, on the other, the RELPRO and the GAP. These two distinct links are identified by the (shared) indices in (2) below:

(2) The captain$_k$ [who$_{k/j}$ the sailor predicted that [the weather would frighten (GAP)$_j$]] turned back to port.

As can be seen, LDDs contain syntactic (construal of the relative pronoun as a grammatical relation in a"noncanonical" position) and lexico-semantic (semantic role assignment) mechanisms which are categorically distinct, and consequently subject to at least partially independent principles of composition. They also involve pronoun interpretation (the establishment of coreference between the RELPRO and the antecedent), which, at least for processing purposes, is identified as a discourse process (e.g., Grodzinsky et al., 1991; Avrutin, 1999; Piñango and Burkhardt, 2005). We take these mechanisms to be encoded in the lexical representation of the RELPRO itself as syntactic, discourse, and semantic selectional requirements respectively. The proposed representation is presented in (3) below:

(3) RELPRO "who"

{Syn:	[$_{NP}$ N$_{head}$	[$_{CP}$ **NP**$_{k/j/w}$	[$_{IP}$ NP$_j$	[V NP$_w$]]]]}
{Sem:	[argument$_k$	[**pronoun**$_{k/j/w}$	[predicate	[argument$_j$, argument$_w$]]]]}
{Discourse:	[antecedent$_k$	[**PRO**$_{k/j/w}$	[SUBJ$_j$	[OBJ$_w$]]]]}
{Phon:	[hu] }			

The representation in (3) specifies the syntactic, discourse, and lexico-semantic environments in which the RELPRO *who* may be licensed, thus capturing the main properties of its linguistic distribution in English. Retrieval of a RELPRO during comprehension therefore means the retrieval of this lexical *composite* with all the mutually constraining algorithms that determine the environment of its realization. In this way the lexical entry itself makes explicit the possible predictions by the parser regarding preceding and crucially, incoming lexical material.

This description thus represents the relevant lexico-syntactic characterization that we take to underlie both the filler-gap effect (e.g., Crain and Fodor, 1985; Stowe, 1986; Swinney et al., 1988; Frazier and Flores d'Arcais, 1989; MacDonald, 1989; McElree and Bever, 1989; Nicol and Swinney, 1989; Fodor, 1995) and its corresponding psycholinguistic generalization, the *Active Filler Hypothesis* (Frazier and Clifton, 1989). Specifically, in this linguistic articulation, the GAP is simply the realization of a coindexation relation between the relative pronoun and a phonologically unsupported [NP+semantic argument+grammatical relation] "triplet" in the embedded IP. The Active Filler Strategy therefore emerges as the implementation of the search to satisfy the RELPRO's requirements[2]. We conjecture that the explicitness of this lexically "packaged" parallel, multi-layer structure is what gives the LDD its seemingly unified processing implementation, what informs the parser as to the syntactic constituents where it can/cannot find a GAP (e.g., Stowe, 1986), and what so powerfully drives the RELPRO (the filler) to hypothesize a GAP even in constructions where it will ultimately be disallowed (e.g., Frazier et al., 1983; Hickok, 1993).

Having made explicit the necessary linguistic and psycholinguistic considerations, we turn to other non-linguistic real-time implementation requirements, specifically, memory requirements. We observe that there are in principle three "inflection points" in the LDD processing: the signaling by RELPRO retrieval that a GAP is incoming, the search for the GAP, and the actual instantiation of the GAP; that is, the point in the composition of the embedded clause where the RELPRO requirements are met (i.e., the GAP). We reason that whereas the antecedent-RELPRO coreference relation and GAP instantiation are unambiguous and local, the instantiation of the search for the GAP is, by contrast, multiply ambiguous due to the availability of multiple potential GAP positions that the RELPRO can be coindexed with and that are associated with all the possible grammatical relations in the embedded clause. This inherent ambiguity is presumably what forces the processor to closely track the syntactic and semantic structure of the incoming embedded clause until the GAP is reached, thus making it memory taxing. It is this basic difference what makes the gap search process a clearer candidate

[2]The index alignment shown across constituents in the syntactic, semantic, and discourse layers makes explicit the observation by most linguistic frameworks of a robust correlation between syntactic category/position, grammatical relation, and

for the probing of cortically localizable real-time linguistic processes.

On this basis, we articulate the LDD into two linguistically distinct stages, the search process itself vs. the licensing point of the GAP. These stages are in turn operationalizable as two mechanisms distinguishable by their differing memory demands. Those mechanisms are:

(i) *GAP-search*: triggered by the retrieval of the RELPRO. It is the language composition process where memory resources are maximally taxed: Upon retrieval of the RELPRO its lexico-syntactic requirements must be satisfied all while the phrase structure and semantic representations of the embedded-clause are being composed [e.g., *who*$_{k/j}$ *the sailor predicted that the weather*, in (2)]. *GAP-search* is the mechanism that effectively implements the Active Filler Strategy: the RELPRO's lexically-driven search within the embedded clause in order to meet its lexico-syntactic requirements.

(ii) *GAP-completion*: triggered by the retrieval of the embedded verbal predicate. It is the process whereby the RELPRO's lexico-syntactic requirements are satisfied. In (2) this process takes place when the embedded verb is retrieved: the earliest point at which the embedded predicate (e.g., *frighten*) can license the object grammatical relation/NP structure and assign to the RELPRO the corresponding semantic role (e.g., *frighten-ee*). This not only completes the interpretation of the RELPRO *who* within the embedded clause but "grounds it," as it were, into the composition of the rest of the embedded clause. It allows the interpretation of the coreferring antecedent (the denotation of the matrix subject head, *captain*) as a participant in the embedded proposition's semantic representation. Crucially, this process, like *GAP-search*, is compositional and therefore expected to require memory resources beyond lexical retrieval. However, given the locality of its resolution, the amount of memory resources *GAP-completion* demands should be significantly less than those demanded by *GAP-search*.

Here, we hypothesize that given their respective linguistic properties and correlated memory demands, these two mechanisms are potentially neurologically dissociable in a way that could shed light on the neurocognitive incongruence at issue. Notably, this kind of processing analysis finds direct support in previous findings by Phillips et al. (2005). That report presents two distinct electrophysiological components associated with long-distance dependency comprehension: a sustained anterior negativity subsequent to the initiation of the *wh*-dependency and a late posterior positivity (P600) associated with the completion of the dependency. We take that pattern to represent the electrophysiological correlates of *GAP-search* and *GAP-completion* respectively and thus take them as initial support for the analytical approach adopted here.

[2]semantic relation, such that if a predicate licenses, say, an agent argument, this argument will bear the subject function, which in English can be associated with NP category and SPEC;IP position (e.g., Chomsky, 1965, 1981; Bresnan, 1982, 2001; Fillmore, 1988; Goldberg, 1995; Van Valin and LaPolla, 1997; Culicover and Jackendoff, 2005).

Most crucially for our present purposes however, a closer look at the fMRI record also suggests the potential viability of this dissociation. We turn to that record directly below.

1.2. LDDs and the LIF Cortex in fMRI: Previous Experimental Record

In this section we discuss previous neuroimaging work that has also targeted either *GAP-search* or *GAP-completion* as we define them here in connection to the workings of the LIFG. Our search through the record was constrained by the requirement that the given report target one, the other, or both mechanisms in question as unified phenomena. The conclusions from that work together with the lesion-based evidence constitute the basis for the specific localizational predictions that we test[3]. Of the large body of neuroimaging work on LDD comprehension, four reports specifically deal with *GAP-search* as we have defined it: (Santi and Grodzinsky, 2007, 2010, 2012) and Matchin et al. (2014). Interestingly, we found no previous work on LDD comprehension targeting *GAP-completion*. In line with the focal lesion evidence these four reports converge on the observation that at least *GAP-search*, as we have defined it here, preferentially recruits the workings of the the LIF cortex. This is what unites them. In what follows we discuss for each of the reports the specifics of how these observations came to be.

Santi and Grodzinsky (2007) connect LDDs to the LIF cortex exclusively through what they call a "distance" effect. They test two phenomena. The one at issue involves object relatives in three conditions: one-NP embedded subject, two-NP embedded subject, and three-NP embedded subject. Crucially, these added NPs are irrelevant to the structure of the RELPRO-GAP dependency itself as the NPs have been added to the embedded subject phrase. Their function in the experimental design is to add material (specifically NP material which is syntactically identical to the RELPRO) between the RELPRO and the object-GAP. This material does not add to the complexity of the LDD but does increase the linear distance between the RELPRO and the GAP. In so doing, it increases the amount of structure the parser must build in order to get to the GAP. Such increase is coupled with an increase in number of nominals (one to three). Santi and Grodzinsky (2007)'s results show recruitment

[3]Our selectional criteria, necessary for our localizational purposes, had the unintended consequence of filtering out reports that have otherwise been valuable for our understanding of LDD processing. Fiebach et al. (2005), for example, connect (non-canonical) *GAP-search* to the LIF cortex, but report activation in other areas as well. For their Long > Short (obj.) contrast, they report in addition to the LIF cortex, right inferior frontal (RIF) cortex, junction of the left precentral sulcus, bilateral STS, MTG (21/22) and the left thalamus. By contrast, for their Long > Short (subj.) contrast, no LIFG is reported. Instead, they report activation in the bilateral inferior portion and left superior portion of the parieto-occipital sulcus (BA 17/30 and BA 7 respectively). So, this report relates the LIF cortex to *GAP-search* but not in a unified manner. Similarly, the results published in Makuuchi et al. (2009) address LDDs but are not directly relatable to our present objectives. Whereas they do report LIF cortex activation in connection to comprehension of double-center embedded clauses vs. single-embedded clauses akin to that reported by Santi and Grodzinsky (2007) and Fiebach et al. (2005), their report is based on a region of interest analysis exclusively, and not on a whole brain analysis. Whereas this approach makes sense given their specific interest in the internal articulation of the LIF cortex and not on localizing LDDs components, it prevents us from concluding whether the association they found targeted specifically the LIF cortex.

of the LIF cortex in the three vs. two nominal increment. We see this manipulation as addressing *GAP-search* as we have defined it (to the exclusion of *GAP-completion*) because in the three-NP condition, the minimal difference was the increase in distance between the RELPRO and the GAP, and this greater distance had to be tracked in order for the parser to get to the GAP[4].

More recently, Santi and Grodzinsky report in two separate papers, 2010 and 2012, an association between the LIF cortex and LDD processing which, given their respective designs, again target *GAP-search* to the exclusion of *GAP-completion*. Whereas in Santi and Grodzinsky (2010) the manipulation involves a comparison between *GAP-search* and embedding, connecting only *GAP-search* to the LIF cortex, Santi and Grodzinsky (2012) distinguishes general *dependency* from *predictability*, the ability of the parser to predict the need for a GAP. Their results show that *predictability* not *dependency* correlates with the LIF cortex effect, focused on BA 45[5].

Finally, Matchin et al. (2014) test the hypothesis that the LIF cortex supports a more general "antecedent-variable" dependency function, thus allowing the possibility to consider *GAP-search* as a member of a larger family of "search"-based processes. Such a hypothesis predicts an LIF cortex preferential activation for *pronoun-antecedent* relations (i.e., backward anaphora) which, like RELPRO-based LDDs, contain as a "variable" an element with an incomplete referential interpretation (pronoun) which must actively look for an "antecedent," the entity with which it must corefer. As with Santi and Grodzinsky (2007), the experimental design of Matchin et al. (2014) targets the *GAP-search* portion of the pronoun-dependency, as we have defined it. Their results show that only the subtractions involving backward anaphora (and not the RELPRO-based LDDs) yielded LIF cortex activation. And for these there was, in addition, activation in the right MTG, STC, bilateral SMA, bilateral occipital activation, and left STS. So, even though the observation is clearly made that the LIF cortex participates in predictive searches similar to *GAP-search*, it is also the case that other cortical regions also participate in this process, rendering the specific contribution of the LIF cortex in the processing of this kind of LDD inconclusive. This said, the presence of LIF cortex activation in this fairly different kind of dependency is suggestive of a deeper processing commonality, which so far has not been fully explored in the neuroimaging literature, and is one that we think may be captured by the generality of the *GAP-search* mechanism[6].

In sum, whereas the vast majority of fMRI research involving LDDs correlate them to cortical regions beyond the LIF cortex, some do provide exclusive or close to exclusive correlation with LIF cortex. Those that do, target *GAP-search* as we have defined it. By contrast, *GAP-completion*, the other major LDD mechanism capturing the more general properties of LDD composition, remains less explored. In light of this, and in order to further understand the factors involved in the neurocognition of LDDs we ask the following questions: What is the neurocognitive relation between *GAP-search* and *GAP-completion*? Do they rely on the workings of overlapping brain regions? And, could we associate *GAP-completion* to the LPST cortex, thus directly addressing the lesion-neuroimaging incongruence? In addition, a new question is revealed: if the effects reported reflect *GAP-search*, why are they observed mainly in the context of *object*-relative GAPs? The specifics of the study seeking to address these question are presented directly below.

1.3. The Study: Determining the Neurological Underpinnings of LDDs

Our analysis above shows that LDD comprehension can be organized into at least two processing mechanisms. We propose here that the existence of this dual mechanism infrastructure and the differential memory resources that it demands is the source of the disparity regarding the cortical recruitment of LDD processing. Moreover, we propose that the reason it has not been detected before has been due to a limitation inherent to the traditional data-analysis approach used in the past. We thus propose that the cortical localizational incongruence is the result of the interaction of two factors: one linguistic and one methodological. The **linguistic factor** refers to the previous analyses which collapse *GAP-search* with RELPRO interpretation at the GAP position, *GAP-completion*, thus conflating processes with potentially distinct neurocognitive demands. The **methodological factor** refers to the traditional approach to data analysis in language-related fMRI whereby

[4]We do observe, though, that this association is not unambiguous. An alternative interpretation to these findings could be that the reported LIF cortex effect results instead from the composition of a more complex meaning structure associated with a semantically more informative embedded subject. In this scenario, preferential activation of the LIF cortex emerges not from greater LDD distance, but from the semantic demands of processing an incrementally more elaborate embedded subject in composition with the embedded transitive verb and its complement. Indeed, this kind of effect is connectable to a similar LIF cortex recruitment found by Husband et al. (2011) and Lai et al. (2014), who independently show LIF cortex involvement in connection, this time, to the processing of complement coercion (e.g., *The girl began the book* vs. *The girl wrote the book*), a phenomenon also described as involving "enrichment" of the semantic representation.

[5]In another related paper, Santi et al. (2015) test a distinction similar to the one reported in 2007. In addition to the NP category, they introduce CP as potential intervening category. Their results show that for both conditions together (CP+NP), there is, in addition to LIF cortex activation, RIF cortex activation, again correlated with distance. The novel comparison here is the joint results involving the CP condition which, as the authors point out, suggest that the syntactic category of the intervening material is not relevant to *GAP-search*, a conclusion that contrasts with previous findings regarding Broca's poor performance in CP production, and fMRI results showing CP processing in connection to the LIF cortex (Shetreet et al., 2009). As in the case of Santi and Grodzinsky (2007), we believe that their results warrant consideration of an alternative interpretation: the possibility that the increased cost contributed by the CP distance be due instead to the possible garden-path created by the absence of complementizer in the lower CP. In the sentence "I knew [which porter the neurosurgeon said] CP_2 [the resident liked GAP] CP_1" two possible structural paths are possible at CP_2. Specifically, the CP_2 verb "said" subcategorizes for both an NP and a CP. When an NP is suggested (due to the absence of the complementizer), the CP possibility is discarded. But this soon proves to be the wrong decision both on semantic grounds (the neurosurgeon said [the resident]$_{NP}$) and on syntactic grounds (∗the neurosurgeon said [the resident liked]$_{*NP}$). Once the parser gets the lower verb "liked," it must revise its original decision in favor of the CP option, consequently incurring a cost.

[6]We find this kind of comparison to be right-minded and useful also because it connects with independent work on the neurology of anaphora resolution which notably reports an impairment in pronoun and logophor resolution in Broca's patients (e.g., Grodzinsky et al., 1993; Avrutin, 1999; Pinango, 2003; Piñango and Burkhardt, 2005; Schumacher et al., 2010).

subtractions take place at the *sentence level*, an approach which, in this case, prevents finer-grained exploration of the intra-sentential components of the dependency.

We address the linguistic factor by testing constructions that vary the degrees of linguistic compositional demands and in doing so allow us to examine the two mechanisms separately. These compositional demands range from a condition where an LDD is not required, as in (4):

(4) The captain **believed** the sailor's prediction yesterday **that** the weather would frighten$_{no-gap}$ **the crew** and turned back to port. (Condition D)

to one where an LDD is required and the link between the RELPRO and the GAP is syntactically direct, as in (2) above repeated here as (5):

(5) The captain$_k$ [**who**$_{k/j}$ the sailor predicted that the weather would frighten (GAP)$_j$] turned back to port. (Condition A)

to one where the syntactic connection between the RELPRO and the GAP is not direct [i.e., the intervening syntactic constituent does not contain the predicate licensing the GAP (6)][7]:

(6) The captain$_k$ [**who**$_{k/j}$ [the sailor's prediction yesterday about the weather] had frightened$_{gap}$, turned back to port. (Condition B/C)

Comparing these conditions allows us to observe the extent to which the memory-language interaction is sensitive to actual compositional linguistic mechanisms, and if so, which ones and with what cortical implications. In this respect, (5) > (4) and (6) > (4) in particular allow us to assess the cortical resources that must be recruited as the processor actively searches for the GAP [(5) > (4)] vs. those which must be recruited during the composition of sentence structure which the processor "knows" cannot contain a GAP, as in [(6) > (4)] (see Stowe, 1986; Kluender, 1998, respectively, for early evidence of the sensitivity of the processor to island constraints, and of how, and in contrast to widespread assumptions in linguistics, islands could in fact result from the interaction of processing factors).

With these contrasts in place, we are able to discuss our approach to the examination of the role of memory in the long-distance dependency construction. We do this through a data analysis manipulation whereby the two hypothesized processing mechanisms, *GAP-search* and *GAP-completion* are analyzed as separate *events*. Specifically, we use an intra-sentential event-related subtraction approach whereby subtractions are performed over the relevant non-overlapping segments of the sentence (see Data Analysis section below for technical details). This, in combination with the minimal contrasts in the linguistic manipulation between conditions, presence/absence of GAP and presence/absence of direct antecedent-GAP link, allows us to isolate simple phrase-structure building from active *GAP-search* and from *GAP-completion*, respectively. The details of the experimental design and data analysis are presented directly below (see Lai et al., in press for a similar use of event-related design in the context of semantic composition).

2. THE STUDY: INVESTIGATING *GAP-SEARCH* AND *GAP-COMPLETION*

2.1. Materials

The study contained a total of four conditions (A, B, C, and D) with 60 sentences in each of the conditions. Sentences were constructed as matching quadruples, thus controlling for non-relevant lexico-semantic and syntactic factors. This resulted in a final script of 240 sentences (60 quadruples). Test sentences for Conditions A and B were directly modeled from Gibson and Warren (2004), which introduces the ± direct RELPRO-GAP link manipulation. A sample of a quadruple is presented in **Table 1** below. As can be seen, whereas the conditions differ in the relevant syntactic properties (e.g., verbal vs. nominal: "sailor predicted" vs. "sailor's prediction") they share all other main lexico-semantic components, thus ensuring that they were as close as possible in terms of number of words, word frequency, and sense co-occurrence. Given our interest in separating activation related to *GAP-search* from that related to *GAP-completion* our unit of analysis was the *Event* which was a segment of the sentence. Accordingly, condition matching had to be implemented especially at the event level. For matching (and data analysis) purposes then each sentence was construed in terms of three events which in **Table 2** are observable in the internal bracketing of the sentences: **Event 0** contains the material before the brackets including head noun and relative pronoun/verb, **Event 1** corresponding to *GAP-search* contains the material in bold within brackets; and **Event 2** corresponding to *GAP-completion* contains the material after the brackets. As can be seen, for Event 1, all conditions match in terms of number of words. For Event 2, condition D, the control condition has in addition three words corresponding to the object NP (two words) and the conjunction (one word). We note that as this is the control condition any extra activation associated with the three extra words would be eliminated in the subtraction process.

[7]We call this condition "indirect *GAP-search*" and not "island" for the following reason: the term *island* refers to the perspective of the "moved" constituent *before* it has moved. This perspective states that such constituent cannot "leave" the larger constituent in which it is base-generated. To be sure, indirect *GAP-search* is a direct consequence of "movement"; but movement itself is only a metaphor, it has no processing status (i.e., the processor never carries out the movement; it only deals with its consequence). By contrast, the term indirect *GAP-search* is meant to refer to the perspective of the processor (left-to-right incremental composition). For the processor, what matters regarding any type of island is whether upon encountering a given constituent, it can hypothesize that the GAP is to be found within that constituent. If it can, then that constituent is searched for potential GAP positions, if it cannot, then the processor "waits," as it were, for that (local) constituent to end in order to continue the search. It is this situation that gives rise to the indirectness we refer to: the GAP is incoming, but not in the (minimal) constituent under construction. The "indirect *GAP-search*" label thus allows us to separate the linguistic intricacies of islands, which go well beyond the condition tested here, from one well-attested processing consequence of them. The label "indirect" therefore speaks to the fact that the subcategorized CP is not provided within the local constituent directly after the relative pronoun. So, from the perspective of the parser an "island" is simply a constituent that is not subcategorized and therefore it is not expected to contain the GAP.

TABLE 1 | Four experimental conditions.

	Sentence	Condition
A	The captain, who **[the sailor predicted yesterday that the weather]** would frighten$_{gap}$,turned back to port.	*GAP-search*/direct and *GAP-completion*
B	The captain, who **[the sailor's prediction yesterday about the weather]** had frightened$_{gap}$,turned back to port.	*GAP-search*/indirect and *GAP-completion*
C	*The captain, who **[the sailor's prediction yesterday about the weather]** had frightened the crew, turned back to port.	*GAP-search*/indirect and *GAP-completion* violation
D	The captain believed **[the sailor's prediction yesterday that the weather]** would frighten the crew and turned back to port.	No GAP-search and No GAP-completion

TABLE 2 | Experimental conditions by events.

Condition	Event 0	Event 1 (± ***GAP-search***)	Event 2 (± ***GAP-completion***)
A	The politician who	the journalist claimed that the government report	had bothered$_{gap}$ is calling a press conference
B	The politician who	[the journalist's claim about the government report]$_{island}$	had bothered$_{gap}$ is calling a press conference
C	The politician who	[the journalist's claim about the government report]$_{island}$	had bothered the people is calling a press conference
D	The politician believed	the journalist's claim that the government report	had bothered the people and is calling a press conference

TABLE 3 | Planned subtractions by events: single subtractions.

Subtraction	Event 1	Event 2
A > D	*GAP-search*$_{direct}$	*GAP-completion*
B > D	*GAP-search*$_{indirect}$	
A > B	*GAP-search* (*GAP-search*$_{direct}$ – *GAP-search*$_{indirect}$)	–
B > C	–	*GAP-completion*

(For further description of the analysis approach see **Table 3** in the Data Analysis section).

Table 1 presents the conditions with their respective dependencies. Figure S1 in the Supplementary materials presents the corresponding syntactic structures (Note that for Conditions A vs. B/C, the different syntactic structures determine the nature of the link between the RELPRO and GAP: *direct* for A and *indirect* for B). Asterisk (∗) in Condition C signals ungrammaticality.

In addition, the A, B, and D conditions were pre-tested for acceptability using a five-point likert scale. This pre-test allowed us to ensure that even though D would be more acceptable than A and B, there would be no difference in acceptability between A and B conditions. And this is what planned comparisons show. As expected Condition D [$D_{mean} = 3.79$ ($SD = 0.5$)] was deemed significantly more acceptable than conditions A [$A_{mean} = 2.66$ ($SD = 0.5$) ($t = -4.05$, $p < 0.001$)] and B [$B_{mean} = 2.67$ ($SD = 0.6$) ($t = -4.1$, $p < 0.001$)]. Also as expected no statistical difference in acceptability between A and B was found ($t = -0.03$, $p = 0.48$). This was calculated on the basis of responses from a sample of 13 native English speakers from the Yale undergraduate population, the same population from which the fMRI participants were selected.

Comprehension questions followed all condition A, B, and D sentences. No questions followed condition C sentences as the kind of ungrammaticality in that condition makes it difficult to ask questions that have an unambiguous yes/no answer. This said, we note that the ungrammaticality in Condition C appears toward the end-of the sentence, crucially, at the *GAP-completion* segment sentence. So, subjects could not know during the first part of the sentence up to the embedded verb whether they were in the presence of a grammatical or ungrammatical sentence. This motivated them to pay attention to all sentences equally.

In addition, questions probed different combinations of the matrix subject, embedded subject, matrix verb, and embedded verb. This variability was introduced intentionally to motivate participants to pay attention throughout the sentence as opposed to specific features of the sentence. To further minimize strategizing, the assignment of a given question to a given sentence was random, so even if the participants could realize that the matrix/embedded subject nouns and the matrix/embedded verbs mattered, for any given sentence they could not predict what specific element would be queried. So, they had to pay attention to all components of the sentences equally. For a sentence like *The captain, who the sailor predicted yesterday that the weather would frighten, turned back toward port.*, subjects would get one of these possible questions:

i. Did the sailor predict that the weather would frighten the captain? (expected answer: Y)
ii. Did the captain predict that the weather would frighten the sailor? (expected answer: N)
iii. Did the captain turn back toward port? (expected answer: Y)
iv. Did the sailor turn back toward port? (expected answer: N)

Coming back to the experimental sentences, this is what each condition probes:

Condition A examines *GAP-search*, triggered at *who* and *GAP-completion*. The distance between the RELPRO and the GAP is expected to reveal the workings of the memory system in a situation where finding the GAP is expected, given the

absence of intervening islands, as compared to Condition D, the no-GAP condition, and Condition B, the island condition where the GAP is not expected *within* the local constituent.

Condition B also combines *GAP-search* (triggered at *who*) and *GAP-completion*. However, in contrast to Condition A, in Condition B the search for the GAP must bypass the embedded subject (which is an island). Bypassing the embedded subject means that the processor needs to wait for that NP constituent to end to find the GAP. That is what the B>D contrast is intended to reveal.

For both A>D and B>D contrasts there is a clear interaction with the memory system in connection to *GAP-search*. So similarity in recruitment is expected. A difference in recruitment (A>D) and (B>D) would then be interpreted as a difference in the *quality* of the interaction with respect to *GAP-search*, one where the processor is not actively looking for the GAP (B>D), vs. one where it is (A>D).

Condition C is identical to Condition B except that the GAP position has been filled with an additional NP, which renders the sentence ungrammatical. The motivation for this condition focuses on the possible distinct cortical recruitment associated with *GAP-completion*. If, as we hypothesize, *GAP-completion* has distinct neurological commitments from *GAP-search*, this process will be observed as a unique activation pattern when comparing Conditions B>C, as these two conditions differ only with respect to the *GAP-completion* factor. B>C thus effectively brings us the closest to observing the preferential recruitment for *GAP-completion* alone.

Condition D represents the control condition. It has the same number of words and constituents as the Condition A and B counterparts, thus equally requiring full phrase-structure building and semantic composition. It lacks a long-distance dependency, so it is expected to tax the memory system the least in comparison to Conditions A or B.

2.2. Design

Each subject was presented with the 240-sentence script containing the 4 conditions, A, B, C, and D (60 items per condition). No additional fillers were included in the script. All 240 sentences were distributed in a pseudo-random fashion in 10 separate runs of 24 sentences each. The four experimental conditions were distributed in a counterbalanced fashion within each run such that no two sentences of the same quadruple would be included in the same run. Each subject was presented with a unique order of runs. So, in the end no two subjects saw the exact same sentence presentation order.

Each sentence presented had a maximum of 22 words. Each word in the sentence was visually presented at 500 ms per word. The 500 ms/word pace was chosen out of a variety of timings previously considered because it was the one that optimized ease of reading, speed, and accuracy in the comprehension of the sentence.

For 180 (75%) of the sentences, a query (yes/no question) about the sentence just read was presented for 4000 ms. The ISIs within and between (sentence+query) items were each 500 ms for a total of 16 s per item. Accordingly, the total time per run was 6 min 24 s (16 s × 24 sentences).

2.3. Procedure

The pre-scanning practice session was designed to familiarize the participants not only with the general procedure in the scanner but also with the length of the experimental sentences. In this practice session each participant was exposed to long embedded sentences similar to the ones they would be encountering in the study and at the same reading pace: one word at a time, paced at 500 ms per word, presented at the center of the screen and followed by a comprehension question.

Participants were instructed to read the sentences silently in the most natural way possible. To facilitate this, sentences were presented with punctuation marks (commas) supporting a native prosodic contour. Responses to the queried sentences were recorded with a yes/no button box. The total duration of the functional component of the study was about an hour, and the total duration of the testing session was 90 min.

2.4. Participants

Fifteen native speakers of English (8 female and 7 male) between the ages of 18 and 22 participated in this study. All except for one subject were right handed with normal or corrected-to-normal vision. By their own report, none had suffered a concussion nor were they under treatment for a neurological or psychological condition. All participants gave written informed consent in accordance with the guidelines set by the Yale University Human Subjects Committee and were compensated for their participation.

2.5. Data Acquisition

Head positioning in the magnet was standardized using the canthomeatal landmarks. In the scanner, cushions inside the head coil were used to reduce head movement and headphones were used to dampen the scanner noise and to communicate with participants. Conventional T1-weighted spin-echo sagittal anatomical images were acquired for slice localization using a 1.5T whole body imaging system with a quadrature head coil (Siemens, Erlangen, Germany). After a 3-plane localizer and a multiple-slice sagittal localizer, 28 T-1 weighted axial slices (TR = 485 ms; TE = 11 ms; bandwidth = 130 Hz/pixel; FA = 90°; slice thickness = 5 mm; FOV = 200 × 200 mm; matrix = 256 × 256) were obtained using flash spin-echo imaging parallel to the anterior and posterior commissure (AC–PC). Ten functional data series were then acquired with a single-shot gradient-echo echo planar imaging (EPI) sequence (TR = 2000 ms; TE = 30 ms; bandwidth = 1735 Hz/pixel; FA = 80°; slice thickness = 5 mm; FOV = 220 × 220 mm; matrix = 64 × 64; with 196 measurements) with same slice localizations as the T-1 anatomical. Stimuli were projected onto a semi-transparent screen at the head of the bore, viewed by the subject via a mirror mounted on the head coil. At the end of the functional imaging, a high resolution 3D Magnetization Prepared Rapid Gradient Echo (MPRAGE) sequence (TR = 24 ms; TE = 4.66 ms; bandwidth = 130 Hz/pixel; FA = 45°; slice thickness = 1.3 mm; FOV = 340 ×

340 mm; matrix = 256 × 256) was used to acquire sagittal images for multi-subject registration.

2.6. Data Analysis

All data were converted from Digital Imaging and Communication in Medicine (DICOM) format to analyze format using XMedCon (Nolfe et al., 2003). During the conversion process, the first three images at the beginning of each of the eight functional series were discarded to enable the signal to achieve steady-state equilibrium between radio frequency pulsing and relaxation leaving 193 images per slice per trial for analysis. Functional images were realigned (motion-corrected) with the Statistical Parametric Mapping 5 algorithm (www.fil.ion.ucl.ac.uk/spm/software/spm5) for three translational directions (x, y, or z) and three possible rotations (pitch, yaw or roll). Trials with linear motion that had a displacement in excess of 1.5 mm or rotation in excess of 2 degrees were rejected.

Individual subject data were analyzed using a General Linear Model (GLM) on each voxel in the entire brain volume with regressors specific for each task. For each of the four sentence types (A, B, C, D) there were four regressors (shown in **Table 2**): **Event 0** = onset of the first word up to the offset of "that/about," **Event 1, GAP-search** = onset of subject of relative/complement clause up to offset of word before lowest embedded verb; **Event 2, GAP-completion** = onset of lowest embedded verb up to end of the sentence, **Question** = onset of comprehension question up to the end of the question. We account for the hemodynamic delay within the General Linear Model used which includes the waver hemodynamic response function (hrf) from the AFNI software.

The resulting beta images for each task were spatially smoothed with a 6 mm Gaussian kernel to account for variations in the location of activation across subjects. The output maps were normalized beta-maps, which were in the acquired space (3.438 × 3.438 × 5 mm).

To take these data into a common reference space, three registrations were calculated within the Yale BioImage Suite software package (www.bioimagesuite.org, Papademetris et al., 2006). The first registration performs a linear registration between the individual subject raw functional image and that subject's 2D anatomical image. The 2D anatomical image is then linearly registered to the individual's 3D anatomical image. The 3D differs from the 2D in that it has a 1 × 1 × 1 mm resolution whereas the 2D z-dimension is set by slice-thickness and its x-y dimensions are set by voxel size. Finally, a non-linear registration is computed between the individual 3D anatomical image and a reference 3D image. The reference brain used was the Colin27 Brain (Holmes et al., 1998) which is in Montreal Neurological Institute (MNI) space (Evans et al., 1992) and is commonly applied in SPM and other software packages. All three registrations were applied sequentially to the individual normalized beta-maps to bring all data into the common reference space.

Data were corrected for multiple comparisons by spatial extent of contiguous suprathresholded individual voxels at an experiment-wise $p < 0.05$. In a Monte Carlo simulation within the AFNI software package and using a smoothing kernel of 6 mm and a connection radius of 6.97 mm on 3.44 × 3.44 × 5 mm voxels, it was determined that an activation volume of 197 original voxels (5319 microliters) satisfied the $p < 0.05$ threshold. Clusters were created for each of the four subtractions. Each cluster was identified with a region label, and then associated with additional numeral labels corresponding to Brodmann areas. Regional labels were assigned using the Yale Brodmann Area Atlas which is defined on the Colin27 Brain at 1 mm resolution.

2.7. Predictions

Table 3 presents the planned single subtractions isolating the two mechanisms in question and corresponding to the two (intrasentential) events: Event 1 and Event 2. Event 1-related subtractions target *GAP-search* and *direct vs. indirect GAP-search*: the correlates of a lexically driven search for the GAP in two contexts, direct vs. indirect, above and beyond phrase-structure building considerations. Event 2-related subtractions target *GAP-completion*: the satisfaction of the syntactic and lexico-semantic requirements of the RELPRO as comprehension unfolds. In addition, a series of *double subtractions* and three conjunction analyses were also performed to show whether or not any of the potential effects observed could be viewed as tapping a common cognitive process and if so which one. The specific double subtractions and conjunction analyses are presented further below in connection to the corresponding general predictions[8].

If *GAP-search*-which takes place during Event 1- and *GAP-completion*-which takes place during Event 2- place compositionally distinct linguistic demands with presumably different memory load implications, then they are likely to have distinct cortical recruitment commitments. The existence of distinct cortical recruitment is in turn hypothesized to be the root of the lesion-based/neuroimaging incongruence regarding LDD implementation. This distinction in recruitment should be observed between the two events across the relevant conditions (e.g., **Conditions A and B** vs. **Condition D** during Event 1 and **Conditions A and D** vs. **Condition D** during Event 2). Specifically:

2.7.1. Prediction for *GAP-search*: $GAP\text{-}search_{direct}$ and $GAP\text{-}search_{indirect}$

If the LIF cortex supports *GAP-search*, regardless of whether it locally leads to a GAP position or not, both the $GAP\text{-}search_{direct}$ and $GAP\text{-}search_{indirect}$ conditions (**Condition A**, Event 1 and **Condition B**, Event 1, respectively) should elicit the same pattern when a no-GAP condition is subtracted (**Condition D**, Event 1).

If, by contrast, the brain distinguishes between the situation where the memory system is actively *participating* in the *GAP-search* process, rather than simply *supporting* the phrase structure composition that happens to involve this process, we should observe a divergence in activation. In this case we expect that at least $GAP\text{-}search_{direct}$ - the condition that has been previously reported to be vulnerable in Broca's aphasia, is correlated with LIF cortex activation.

[8]We thank a reviewer for calling our attention to the importance of these two second-order analyses which as will be seen strengthened the quality of the evidence overall.

Three double subtraction analyses (1) *GAP-search* $A_1>D_1$ vs. *GAP-completion* $A_2>D_2$, (2) *GAP-search* $B_1>D_1$ vs. *GAP-completion* $B_2>D_2$ and (3) *GAP-search*$_{direct}$ $A_1>D_1$ vs. *GAP-search*$_{indirect}$ $B_1>D_1$ and one conjunction analysis *GAP-search*$_{direct}$ $A_1>D_1$ and *GAP-search*$_{indirect}$ $B_1>D_1$ are relevant for this prediction. The first two double subtractions test LIF cortex sensitivity to *GAP-search* once activation associated with *GAP-completion* has been eliminated. The third double subtraction and the conjunction analysis allows us to see the extent to which *GAP-search*$_{direct}$ and *GAP-search*$_{indirect}$ have common activation.

2.7.2. Prediction for *GAP-completion*

Our analysis confers *GAP-completion* a subordinate role in LDD composition as it is a strictly local process connecting *GAP-search* to the ongoing composition of the sentence. In terms of cortical localization, we have seen that the previous neuroimaging record does not isolate it. By contrast, the focal-lesion record gives us an important clue as to *GAP-completion*'s potential cortical distribution: For Broca's patients, the reactivation of the GAP, presumably involving *GAP-completion*, is not simply absent, it is *abnormal*. The GAP-filling effect is absent right after the licensing verb, but visible around 500 ms later (e.g., Burkhardt et al., 2003; Love et al., 2008). On the basis of our analysis, we interpret this comprehension pattern as the manifestation of a dissociation between *GAP-search* and *GAP-completion* such that the latter is evidently impacted by, but is not crucially dependent on, the workings of the LIF cortex.

Completing this picture, the lesion-based evidence also tells us that Wernicke's patients are able to implement gap-filling in a timely manner. Yet, *in offline tasks* such as sentence-to-picture matching, these very patients show impaired comprehension not only of object relative clauses, but also of subject relative clauses and non-embedded agentive matrix clauses, a behavior that has traditionally been rooted to a lexically-based deficit, and that accordingly confers Wernicke's area a generalized *compositional* role with direct semantic implications (e.g., Caramazza and Zurif, 1976; Shapiro and Levine, 1990; Piñango and Zurif, 2001, 2015).

Combining these pieces we reason that if there is a connection between LDD composition and the LPST cortex at all, it should be neither in connection to *GAP-search* nor to *GAP-completion* specifically, but in connection to a more general compositional process, involving the coupling of morphosyntactic and semantic composition, of which *GAP-completion* is but one manifestation. So, the localization prediction for *GAP-completion* is exploratory: *GAP-completion*—targeted in three Event 2-related comparisons (1) **Condition B** vs. **Condition C**, (2) **Condition B** vs. **Condition D**, and (3) **Condition A** vs. **Condition D**—should not activate the LIF nor the LPST cortices. But it should show an activation pattern that is instead neuroanatomically *connectable* to both LIF and LPST cortex associated with Broca's and Wernicke's aphasia respectively.

In terms of the double subtraction and conjunction analyses associated with this prediction, our objective is to determine whether or not, despite arising from different contrasts, the three activation patterns predicted to reveal *GAP-completion* indeed manifest the same preferential recruitment. Specifically, we compare $A_2>D_2$ and $B_2>D_2$ to each other and crucially to $B_2>C_2$, and look at how they differ (subtractions) and what cortical recruitment they have in common (conjunction).

In the strongest form of the prediction, if $A_2>D_2$ and $B_2>D_2$ are targeting the same process, subtracting them from each other and from $B_2>C_2$ should result in no difference. By the same token if the three subtractions are revealing the same cognitive process, the conjunction analysis with all three subtractions $A_2>D_2$ and $B_2>D_2$ and $B_2>C_2$ should show a high degree in overlap, one that is coherent with the single subtraction results.

3. RESULTS

3.1. Behavioral Task

Results from the post-sentential questions show an average accuracy rate of 87.6%, which was distributed across conditions as follows: Condition A: 90.57% (29.24), Condition B: 87.47% (33.1), Condition D: 84.82% (35.9). A mixed-model analysis revealed a marginally significant effect of condition [Chi-square $= 7.59$, ($df = 2$) $p = 0.083$]. Pairwise comparisons revealed a significant difference in A vs. D ($p = 0.001$) and a marginally significant difference in A vs. B ($p = 0.08$). There was no difference between B vs. D ($p = 0.1$) (all results corrected for multiple comparisons).

Given that all conditions show an accuracy rate higher than 80%, we interpret the A vs. D difference as the result of lapses in attention due to the relatively undemanding nature of the D condition which allowed the subjects to lose concentration and in turn miss some of the comprehension questions.

3.2. Isolating *GAP-search*: *GAP-search*$_{direct}$ and *GAP-search*$_{indirect}$

Figure 1A shows the pattern of activation, presented in radiological format, for the $A_1>D_1$ subtraction (Conditions A and D, Event 1). Two main regions of preferential activation are observed: the first one involves left BAs 45, 44, 47, 22 (inferior, medial), 38, and insula. The second one involves posterior cingulate, left primary and association cortex, and BAs 7 and 31 (both bilateral). **Figure 1B** shows the $B_1>D_1$ contrast. Interestingly, this pattern of activation appears as a non-overlapping recruitment involving one region connecting bilateral BA 6 (medial superior), bilateral BA 8, and bilateral BA 32 and right BA 24. **Table 4** below shows the significant differential volume by region for each of these comparisons.

The first double subtraction ($A_1>D_1$ vs. $A_2>D_2$) (**Figure 2**, **Table 5** below) shows a pattern almost identical to the one yielded by the original single subtraction: left BAs 47, 46, 45, 44, and 38, medial BA 7, BAs 17, 18, and 19, and the cerebellum. The second and third double subtractions ($B_1>D_1$ vs. $B_2>D_2$) and ($A_1>D_1$ vs. $B_1>D_1$) by contrast yielded no significant activation.

Finally, the conjunction analysis counterpart comparing direct vs. indirect search showed an empty intersect. This analysis which, crucially, is based only on the corrected maps, tells us that for this comparison the stronger more reliable activation is in terms of the *differences* in preferential activation between the two contrasts. This supports the possibility that for Event 1, any privileged association is not between LIF cortex and *GAP-search* but between LIF cortex and *GAP-search*$_{direct}$.

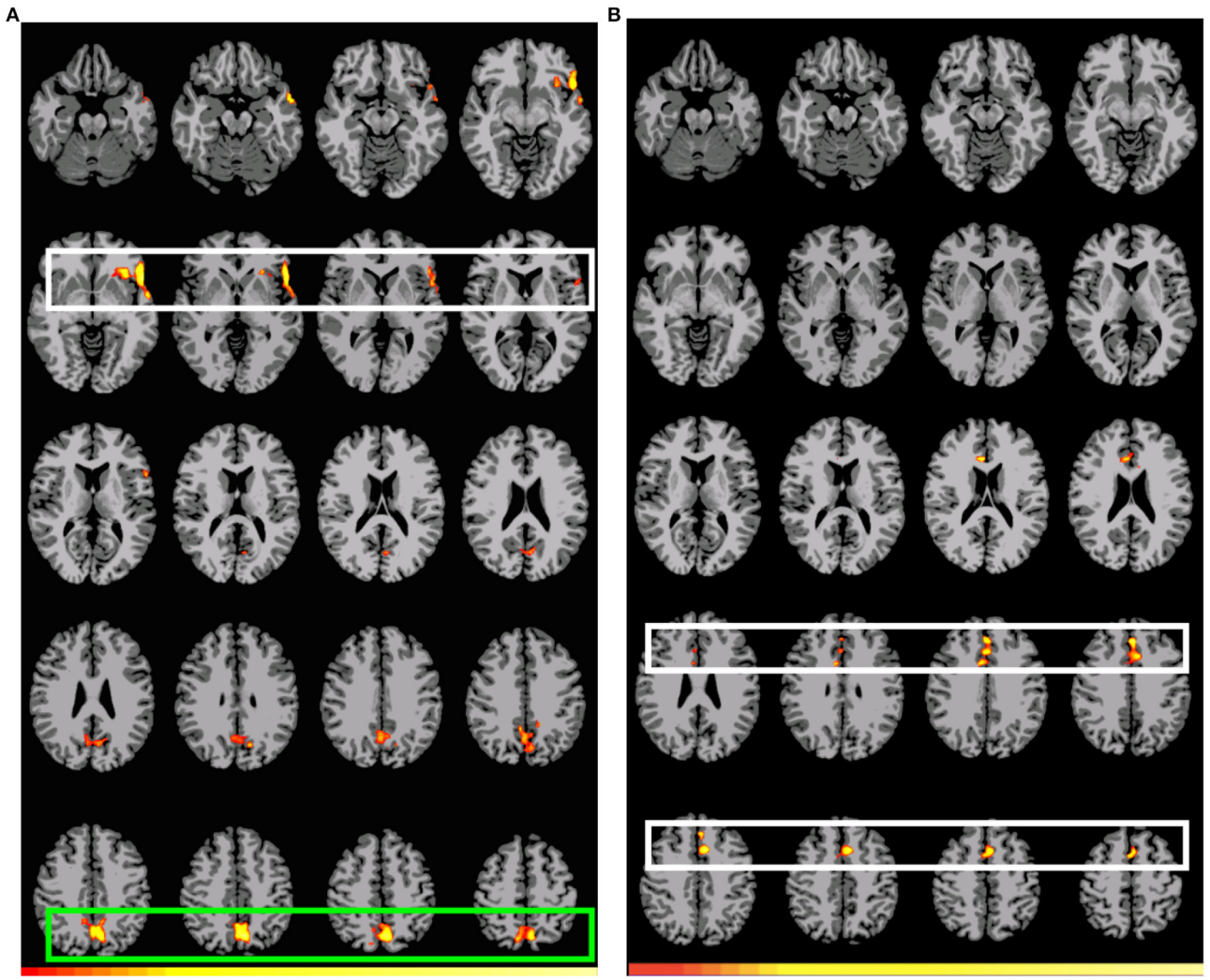


FIGURE 1 | Preferential activation for both *GAP-search* (Event 1) subtractions. Images are shown corrected at $p < 0.05$ in radiological format (LH is on the right). **(A)** *(Direct) GAP-search* (Event 1) subtraction: $A_1 > D_1$: "the journalist claimed that the government report" > "the journalist's claim that the government report." White: BAs 45, 44, and 47. Green: Posterior cingulate and sensory association cortex. Low thresholds ($p < 0.05$; $t = 2.14$) are indicated by red, while high thresholds ($p < 0.000007$; $t = 6.94$) are indicated by yellow. **(B)** *(Indirect) GAP-search* (Event 1) subtraction: $B_1 > D_1$: "the journalist's claim about the government report" > "the journalist's claim that the government report." White: SMA activation. Only positive activation reported. Low thresholds ($p < 0.05$; $t = 2.14$) are indicated by red, while high thresholds ($p < 0.00002$; $t = 6.42$) are indicated by yellow.

3.3. Isolating *GAP-completion*

Table 6 (**Figure 3**) shows the pattern of activation for all three contrasts involving *GAP-completion*: **Condition B > Condition C** (Event 2), **Condition A > Condition D** (Event 2), and **Condition B > Condition D** (Event 2), respectively. We interpret them together because the pattern of activation they each give rise to is by our hypothesis reflecting the same *GAP-completion* process. We present them separately because each emerges from different surface-level subtractions: a legitimately filled gap vs. an illegitimately filled gap ($B_2 > C_2$) and a filled gap vs. non-gap ($A_2 > D_2$ and $B_2 > D_2$).[9] Moreover, those segments come from different (non-local) sentential contexts (A and B, respectively). We reason that if *GAP-completion* is an isolable process, it should yield a similar activation pattern regardless

[9]A reviewer asks us about the meaningfulness of the B>C (event 2) comparison. The results from the B>C (event 2) subtraction reflect the preferential activation triggered by B (event 2) without the material that is represented by C (event 2). We infer that this preferential activation includes any process involved in the interpretation of the GAP that survives the violation which, by our analysis, includes the interpretation of the GAP by means of the dependency formation. Indeed, we take this to be the substance of the residual of the subtraction. This contrasts with the violation condition, where the GAP position has been independently filled thus preventing dependency from being completed. As we discuss here, what is interesting of this subtraction, B>C (event 2), is that despite emerging from a subtraction by a violation, the resulting pattern is comparable to the others in Event 2, particularly in terms of the SMA and parietal activation pattern, which the other two non-violation based contrasts also show.

TABLE 4 | Significant differential volumes by region for *GAP-search* subtractions.

(Direct) GAP-search (Event 1) subtraction, $A_1>D_1$				
Region	**Volume (mm^3)**	**Mean *T*-value**	**Max *T*-value**	**Max MNI Coords. (x, y, z)**
Region 1: left BA 45, left BA 44, left BA 47, left BA 22 (inferior, medial), left BA 38, insula	5571	2.94	5.83	−54, 18, −3
Region 2: posterior cingulate, left visual cortex (primary and association), bilateral BA 7 and 31	9516	2.82	6.94	−9, −66, 45
(Indirect) GAP-search (Event 1) subtraction, $B_1>D_1$				
Region	**Volume (mm^3)**	**Mean *T*-value**	**Max *T*-value**	**Max MNI Coords. (x, y, z)**
Right BA 24, bilateral BA 6 (medial superior), bilateral BA 8, bilateral BA 32	6262	2.80	5.56	−6, 15, 45

TABLE 5 | Significant differential volumes by region for *GAP-search* (Event 1) double subtraction ($A_1>D_1$) vs. ($A_2>D_2$).

Region	Volume (mm^3)	Mean *T*-value	Max *T*-value	Max MNI Coords. (x, y, z)
Region 1 left BA 47, left BA 46, left BA 45, left BA 44, left BA 38 (temporal pole)	9328	2.70	6.03	−51, 30, −6
Region 2: posterior cingulate, sensory (association)	7522	2.94	5.84	0, −63, 36
Region 3: cerebellum	7526	2.71	6.15	24, −69, −33

of non-local context. This is especially the case for $A_2>D_2$ and $B_2>D_2$ which share the same subtrahend.[10]

Figure 3 shows that there is indeed a very similar pattern of activation across the three subtractions. For all three contrasts, there are two main foci of preferential recruitment: BA 6 (left and bilateral) and visual cortex (primary/association). This said, type of subtraction also mattered: for the $A_2>D_2$ and $B_2>D_2$ contrasts, common preferential areas were revealed which did not emerge in the $B_2>C_2$ subtraction: anterior cingulate, BA 7 (precuneus), and BA 32. In addition, $B_2>D_2$ revealed activation of left BA 40. Finally, none of the contrasts showed overlap with BA 44 or BA 45–regions that were observed in the *GAP-search* condition. This finding was further confirmed in the double subtraction and conjunction analyses (see below). All results (from both Events 1 and 2) are summarized in **Table 7** below.

The two *GAP-completion*-related double subtractions yielded interesting results. We predicted that if all Event 2 subtractions are targeting the same process, subtracting one from the other should result in no difference. And indeed that is what we found for $A_2>D_2$ vs. $B_2>D_2$. When these two conditions were compared to $B_2>C_2$ a difference was observed not in terms of localization but in terms of *volume* of activation. As **Table 8** (see also **Figure 4**) shows, the activation pattern observed for these single and corresponding double subtractions is almost identical.

[10] The activation in blue associated with the violation C segment, C_2 vs. B_2, is addressed in the subsection "Other Patterns."

What we observe in the double-subtraction is a change in the volume for the SMA and which goes from 17,585 in the single $B_2>C_2$ subtraction down to 11,479 when subtracted by $A_2>D_2$ and to 13624 when subtracted by $B_2>D_2$. We compare these double subtractions to one where the *GAP-search* counterpart is subtracted: $B_2>C_2$ vs. $B_1>C_1$. We reason that if the previous two double subtractions are reflecting *GAP-completion* their results should converge with this one which isolates *GAP-completion* from *GAP-search*. And that is what we find. These results are summarized in **Table 9**. Finally, the conjunction analysis confirms these findings by showing again not only the primary and association visual cortex and connected posterior cortex, but crucially, BA 6 as a main area of overlap. These results are summarized in **Table 10** and shown in **Figure 5**.

3.4. Activation Beyond *GAP-search* and *GAP-completion*: Discourse-Composition and GAP Violation

In the Event 2 contrast, an additional pattern of activation is observed which results from the inverse subtraction $C_2>B_2$ and which is associated with a GAP violation. (The violation is caused by an expected GAP that already appears filled.) This contrast was not part of the main question the study seeks to address, but in light of the other results, it reveals a very interesting pattern which we believe is connectable to our main question. The $C_2>B_2$ segment, which reflects the violation proper, recruits no LIF, LPST, or parietal cortices. Instead, it recruits the *right hemisphere* BAs 45 and 46 and bilateral prefrontal cortex (BAs 9 and 10).

This pattern is interesting because it reflects cortical recruitment beyond the traditional language areas, suggesting that its impact is outside language composition strictly speaking. Indeed in connection to this observation a reviewer points out, correctly in our view, that this pattern of activation lines up with the so-called default mode network (DMN); a network traditionally associated with resting states or situations where subjects are left to carry out "undirected" thinking. Consequently, the reviewer suggests, these could be an indication that the parser most likely has simply halted the comprehension process.

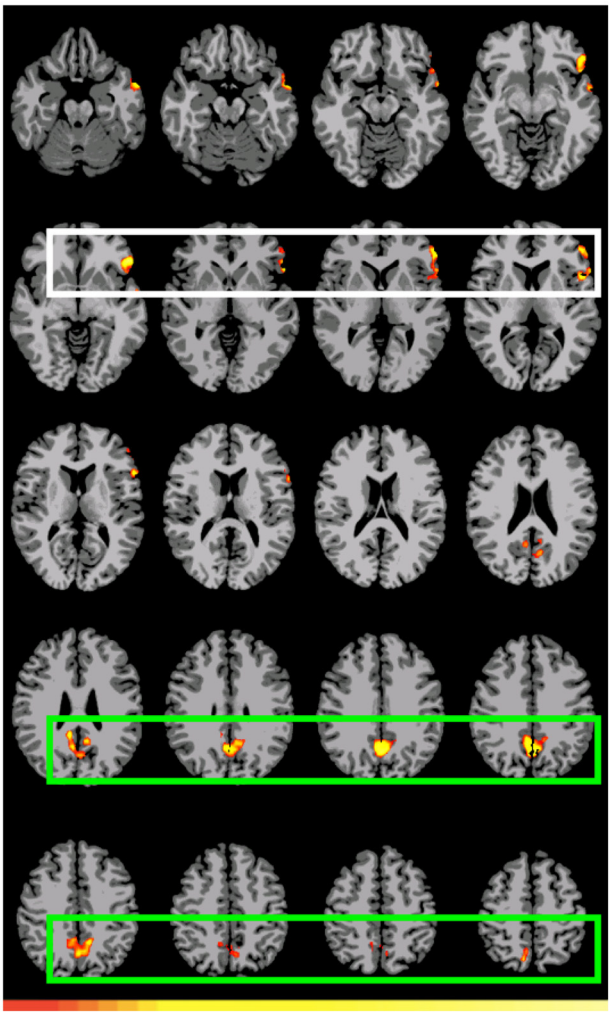


FIGURE 2 | Preferential activation for *(direct) GAP-search* (Event 1) double subtraction: (A_1>D_1) vs. (A_2>D_2). White: BAs 44, 45, 46, and 47. Green: Posterior cingulate and sensory association cortex. Images are shown corrected at $p < 0.05$ in radiological format (LH is on the right). Only positive activation reported. Low thresholds ($p < 0.05$; $t = 2.14$) are indicated by red, while high thresholds ($p < 0.000005$; $t = 7.20$) are indicated by yellow.

We agree with the reviewer that to the extent that we do not fully know the impact of ungrammaticality in the process of comprehension, the possibility remains that faced with ungrammaticality, the comprehension system stops tracking linguistic composition altogether, thus allowing the mind to direct thought away from the utterance in question. This said, we would like to propose an alternative interpretation which is connectable with our present aims: that the pattern of preferential activation observed, partially overlapping with the default mode network, directly reflects the specific *discourse-based* nature of the violation in Condition C; a possibility that complements the recruitment pattern involved in gap-search/completion. On our analysis, the violation in Condition C is caused by the inability of the parser to integrate the composed meanings of the embedded and matrix clauses. These clauses are each independently syntactically and semantically well-formed yet cannot be linked with each other. The ill-formedness is caused by the requirement that *GAP-completion* apply at a point in the sentence where it is not allowed to. *GAP-completion* is the process where the referent associated with the antecedent finds an interpretation as a participant in the semantic representation associated with the embedded clause, thus linking the proposition denoted by the embedded clause with that of the matrix clause. In the ungrammatical utterance, *The politician who the journalist's claim about the government report had bothered* ***the people*** *is calling a press conference*, *GAP-completion* cannot take place because the GAP is already occupied by another NP (*the people*). Consequently, not only is the antecedent (*the politician*) left without a (necessary) interpretation within the embedded clause, but a new and unexpected semantic interpretation (involving the participant *the people*) has been introduced, which is locally plausible but cannot be connected with the meaning of the matrix clause. These two locally coherent segments (matrix clause: *the politician is calling a press conference* and embedded clause: *the journalist's claim about the government report had bothered the people*) result in compositionally conflicting linguistic representations, which in turn yield a meaning incoherence for the sentence as a whole (i.e., two mutually exclusive individuals "the politician" and "the people" must be licensed as the experiencer of "bother"). The meaning of the embedded clause (containing the new participant) can no longer be incorporated into the meaning of the matrix clause (containing the antecedent). This incoherence cannot be resolved not because there is no one plausible interpretation to be obtained, but because there is one too many plausible interpretations.

We propose that the comprehension system is sensitive to this situation and it is the conflict that it represents what underlies the activation pattern observed for C>B event 2. This would suggest in turn that the thrust of the violation lies on higher level meaning-based structure, even though the violation itself is triggered by a local syntactico-semantic misstep[11].

If this were the case, it would make the non-linguistic regions in question relevant for language comprehension processes involving contextualization or integration of composed meaning. Early support for this possibility is found in fMRI reports suggesting a correlation of relevant right-hemisphere cortical areas with notions such as "discourse" level composition (e.g., Costello and Warrington, 1989; Devlin et al., 2003) and "aboutness" (Bornkessel-Schlesewsky et al., 2012). Specifically relevant to the DNM is the work on fMRI patterns relating the DMN to social cognition processes, in particular those connecting middle frontal cortex with theory of mind processes (see Mars et al., 2012 for a meta-analysis of this body of work in connection also to DMN processes in non-human primates). As noted, this interpretation is not intended to apply to syntactic violations across the board, but to activation

[11]This interpretation rests on a very specific assumption about the parser. The assumption is that the parser will attempt to build an interpretation even in the face of partial incoherence in the input as is the case in condition C. Yet as a reviewer correctly points out this assumption is not necessarily settled in the literature.

TABLE 6 | Significant differential volumes by region for all *GAP-completion* (Event 2) contrasts.

Contrast	Region	Vol. (mm^3)	Mean *T*-value	Max *T*-value	Max MNI (x, y, z)
$B_2>C_2$	Region 1: BA 6 (left, medial, bleeding into left BA 44), left primary motor/sensory	17,585	2.80	5.56	−39, −18, 57
	Region 2: visual cortex (primary/association)	28,691	2.80	5.56	15, −87, −6
	Region 3: caudate and putamen	7820	2.80	5.56	18, 21, 6
$A_2>D_2$	Region 1: BA 6 (bilateral, mostly medial/superior, bleeding into BA 8), anterior and dorsal cingulate, left primary motor/sensory	19,052	2.98	7.61	−24, −9, 57
	Region 2: BA 7, visual cortex (primary/association)	45,564	3.13	8.56	−18, −69, −9
	Region 3: caudate, left putamen, left BA 47, left insula	5829	2.63	4.51	−30, 21, 0
$B_2>D_2$	Connected region: BA 6 (bilateral, medial, superior), BA 8, anterior and dorsal anterior cingulate (right), left primary motor/sensory, BA 7 (bleeding into BA 39), left BA 40, visual cortex (primary/association)	72,530	2.97	9.07	−12, −87, −3

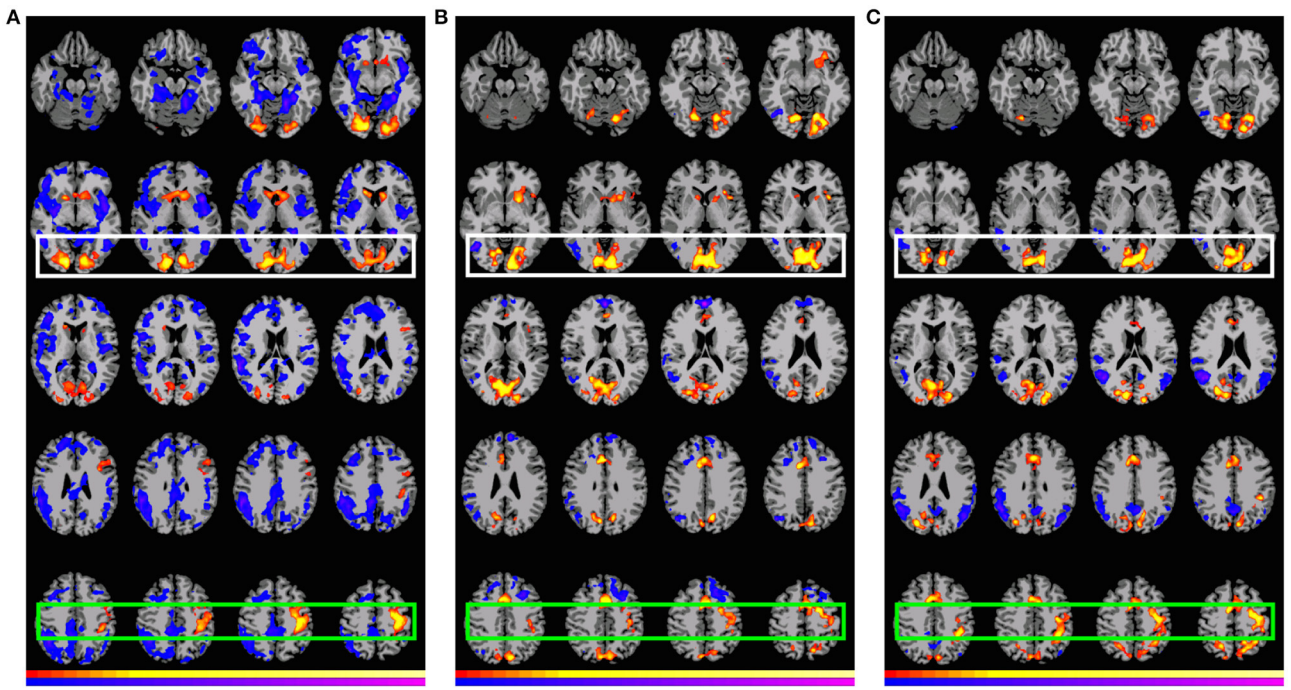


FIGURE 3 | Preferential (positive and negative) activation for *GAP-completion* (Event 2) subtractions. White: Visual cortex (association and primary), Green: SMA and parietal activation. Images are shown corrected at $p < 0.05$ in radiological format (LH is on the right). For **(A)**, low thresholds ($p < 0.05$; $t = 2.14$) are indicated by red and blue, while high thresholds ($p < 0.000000005$; $t = 12.6$) are indicated by yellow and purple. For **(B)**, low thresholds ($p < 0.05$; $t = 2.14$) are indicated by red and blue, while high thresholds ($p < 0.0000006$; $t = 8.6$) are indicated by yellow and purple. For **(C)**, low thresholds ($p < 0.05$; $t = 2.14$) are indicated by red (positive activation) and blue (negative activation), while high thresholds ($p < 0.0000003$; $t = 9.1$) are indicated by yellow (positive activation) and purple (negative activation). **(A)** $B_2>C_2$: "had bothered$_{gap}$ is calling a press conference." > "had bothered the people is calling a press conference." **(B)** $A_2>D_2$: "had bothered$_{gap}$ is calling a press conference." > "had bothered the people and is calling a press conference." **(C)** $B_2>D_2$: "had bothered$_{gap}$ is calling a press conference." > "had bothered the people and is calling a press conference."

patterns where the violation results in a larger discourse incoherence such as that created by a "doubly-filled" argument position. (For a more general discussion about brain patterns and violations, see Embick et al., 2000; Friederici et al., 2003).

4. DISCUSSION

Past neuroimaging work has shown that even though long-distance dependencies seem to recruit the workings of the LIF cortex, they also recruit the workings of the LPST cortex

TABLE 7 | Summary of cortical recruitment by Events: Single subtractions.

Subtraction	Event 1: GAP-search	Event 2: GAP-completion
B > C	–	BA 6 (lateral), BA 44 (edge), visual cortex (BAs 17, 18, 19), caudate and putamen. Activation is also observed in primary motor and primary sensory cortex.
A > D	Region 1: left BA 45, left BA 44, left BA 47, insula, left BA 22 (inferior, medial), BA 38 (temporal pole) Region 2: posterior cingulate, left primary visual cortex, left visual association cortex, bilateral BA 7, BA 31	Region 1: BA 6 (bilateral, mostly medial/superior, bleeding into BA 8), anterior and dorsal cingulate, left primary motor/sensory Region 2: caudate/putamen (left), BA 47 (left), insula (left) Region 3: BA 7 (bilateral), primary visual/association cortex (bilateral medial)
B > D	BA 6 (medial superior), BA 8, BA 24 (right), and BA 32.	BA 6 (bilateral, medial, superior), BA 8, anterior and dorsal anterior cingulate (right), left primary motor/sensory, BA 7 (bleeding into BA 39), left BA 40, visual cortex (primary/association)

TABLE 8 | Significant differential volumes by region for all *GAP-completion* (Event 2) double subtractions.

Contrast	Region	Vol. (mm^3)	Mean *T*-value	Max *T*-value	Max MNI (x, y, z)
(B_2>C_2) vs. (B_2>D_2)	Region 1: left BA 6, primary motor/sensory	13264	3.11	7.18	−33, −18, 60
	Region 2: right visual cortex (primary/association)	10579	3.67	12.07	15, −87, −6
(B_2>C_2) vs. (B_1>C_1)	Region 1 left BA 6, primary motor/sensory, BA 8	18634	3.19	10.06	−39, −27, 57
	Region 2: visual cortex (primary/association)	35814	3.42	12.05	15, −87, −6
	Region 3: caudate, putamen, thalamus	11848	2.68	5.12	18, 21, 12
(B_2>C_2) vs. (A_2>D_2)	Region 1: left BA 6, primary motor/sensory	11479	2.97	6.29	−51, −18, 51
	Region 2: right visual cortex (primary/association)	9155	3.35	8.91	18, −84, −12

and surrounding areas (e.g., Cooke et al., 2002; Fiebach et al., 2002; Amunts et al., 2004; Fiebach et al., 2005; Grodzinsky and Friederici, 2006; Santi et al., 2015). Moreover, while the lexical role of the LPST cortex has been well documented (see Wise et al., 2001; Hickok and Poeppel, 2004, 2007 for proposals regarding the role of the various subcomponents of the LPST cortex in long-term phonological encoding), no conclusive explanation has been given for why this area should be recruited in the instantiation of these dependencies. At the same time, whereas Wernicke's patients (with damage involving the left posterior temporal cortex, including parts of the angular and supramarginal gyri) show across-the-board impaired sentence comprehension including constructions containing dependencies, they are indistinguishable from matching controls in their ability to exhibit the gap-filling effect, thus indicating that whatever their linguistic impairment, it does not seem to involve *GAP-search* or *GAP-completion per se*.

Indeed, Wernicke's performance has been seen to reflect the capacity to implement the basic syntactic mechanics of the dependency, but showing, offline, an inability to put this knowledge to use, presumably due to an inability to properly access the necessary lexico-semantic information that makes the dependency meaningful (e.g., Caramazza and Zurif, 1976; Shapiro and Levine, 1990; see Piñango and Zurif, 2015 for a summary of the main findings). By contrast, Broca's patients, while unable to properly implement these dependencies (e.g., Zurif et al., 1993, 1994; Burkhardt et al., 2003; Love et al., 2008), show, offline, a selective pattern of impairment whereby canonical (subject) relative clauses result in above-chance performance and non-canonical (object) relative clauses reliably result in poor (chance-level) comprehension, a pattern of performance that appears to be linguistic in nature. So whereas the neuroimaging evidence tells us the brain regions that could be potentially participating in the implementation of the dependencies, the lesion-based evidence tells us of the possibility of an *asymmetry* in their participation.

The analysis of LDDs that we present here provides the basis for a potential reconciliation of these two sets of seemingly conflicting observations by invoking organizing principles that could give rise to such an asymmetry. Specifically, the model captures the main linguistic components of a dependency (phrase structure building, argument structure licensing, and pronoun resolution) as selectional/subcategorization constraints on the relative pronoun that separate the process of searching for the environment of argument licensing within the sentence (*GAP-search*) from the actual argument licensing (*GAP-completion*).

In the remainder of this section we discuss the specific activation patterns observed in connection to the hypothesized functional distinctions.

4.1. *GAP-search*: $GAP\text{-}search_{direct}$ vs. $GAP\text{-}search_{indirect}$

The hypothesis that LIF cortex is sensitive to *GAP-search* independently of the internal articulation of the dependency (direct vs. indirect) was not borne out. To the extent that *GAP-search* was reliably associated the LIF cortex it

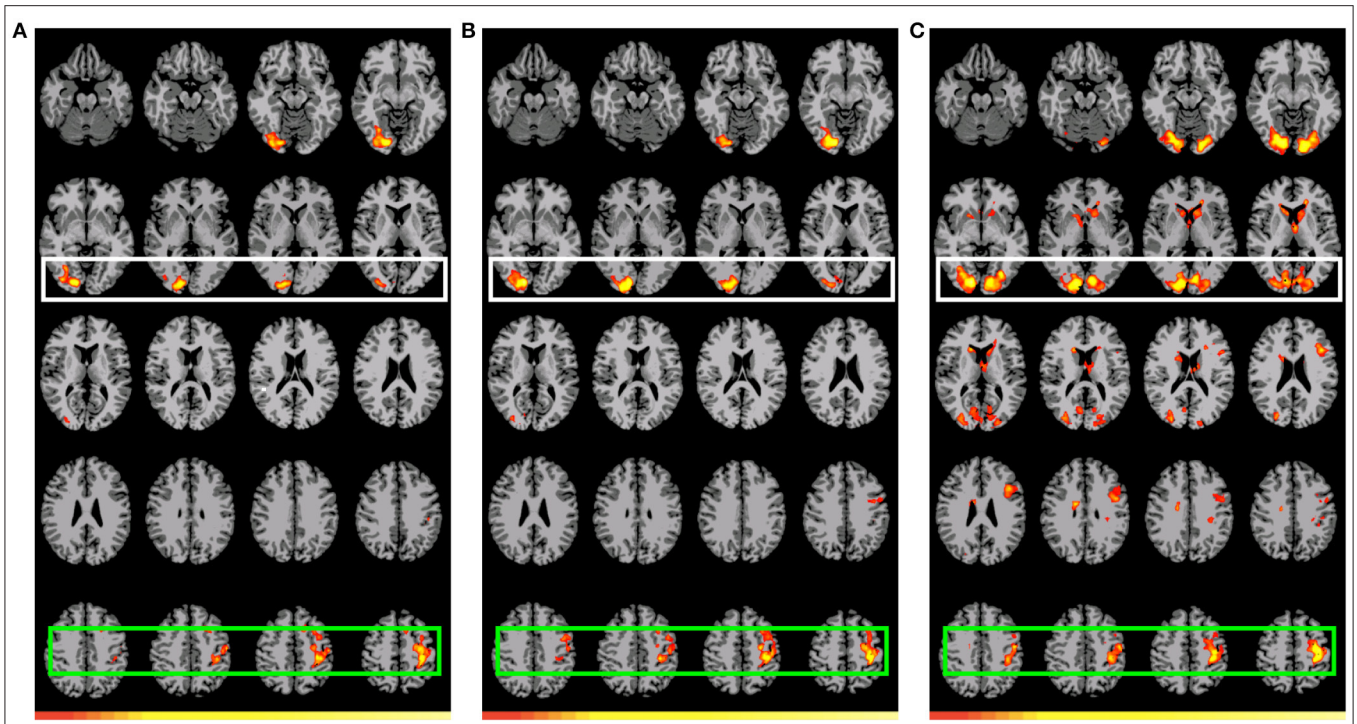


FIGURE 4 | Preferential positive activation for *GAP-completion* (Event 2) double subtractions. White: Visual cortex (association and primary), Green: SMA and parietal activation. Images are shown corrected at $p < 0.05$ in radiological format (LH is on the right). For **(A)**, low thresholds ($p < 0.05$; $t = 2.14$) are indicated by red, while high thresholds ($p < 0.00000001$; $t = 11.8$) are indicated by yellow. For **(B)**, low thresholds ($p < 0.05$; $t = 2.14$) are indicated by red, while high thresholds ($p < 0.000000009$; $t = 12.1$) are indicated by yellow. For **(C)**, low thresholds ($p < 0.05$; $t = 2.14$) are indicated by red, while high thresholds ($p < 0.000000009$; $t = 12.1$) are indicated by yellow. **(A)** ($B_2 > C_2$) vs. ($A_2 > D_2$): Residual preferential activation for *GAP-completion*. **(B)** ($B_2 > C_2$) vs. ($B_2 > D_2$): Residual referential activation for *GAP-completion*. **(C)** ($B_2 > C_2$) vs. ($B_1 > C_1$): Preferential activation for *GAP-completion* > preferential activation for *GAP-search*.

TABLE 9 | Summary of cortical recruitment for all *GAP-completion* (Event 2) double subtractions.

Subtraction	Event 2: *GAP-completion*
($B_2 > C_2$) vs. ($A_2 > D_2$)	Region 1: left BA 6, primary motor/sensory cortex. Region 2: right primary/association visual cortex.
($B_2 > C_2$) vs. ($B_2 > D_2$)	Region 1: left BA 6, primary motor/sensory cortex. Region 2: right primary/association visual cortex.
($B_2 > C_2$) vs. ($B_1 > C_1$)	Region 1: left BA 6, primary motor/sensory cortex, BA 8. Region 2: primary/association visual cortex. Region 3: caudate, putamen, thalamus

TABLE 10 | Significant differential volumes by region for conjunction of *GAP-completion* (Event 2) subtractions , ($A_2 > D_2$) + ($B_2 > C_2$) + ($B_2 > D_2$).

Region	Volume (mm^3)	Mean *T*-value	Max *T*-value	Max MNI Coords. (x, y, z)
Region 1 BA 6, motor (primary, supplementary), sensory (primary), anterior cingulate	5399	X	X	−2, −82, 5
Region 2: right visual (association), right angular/supramarginal gyri, fusiform	4422	X	X	49, −55, 19
Region 3: BA7, BA19, visual (primary, association), sensory (association)	15,869	X	X	−33, −16, 54

was only in connection to the *direct* condition (single and double subtractions). Within this pattern of activation two connected regions were involved: region 1 included BAs 45, 44, 47, bordering with the left insula and left temporal pole (anterior BAs 22 and 38). A second associated region connecting primary and associate visual cortex and BA7 and BA31 were also preferentially recruited. This second region of activation is interesting for two reasons; (1) it appears in $A_1 > D_1$ but not in the $B_1 > D_1$ contrast, and this is relevant because it involves the participation of BA7, a cortical region previously connected to CP embedding, precisely the kind of composition present in A_1 and absent in B_1, and (2) it

continues to appear in connection to *GAP-completion* for both $A_2>D_2$ and $B_2>D_2$ contrasts, thus suggesting that this area is sensitive to general composition such as that involved in *gap-completion*[12].

The results from the $A_1>D_1$ vs. $A_2>D_2$ double subtraction, support the importance of the LIFG for *GAP-search*$_{direct}$, an observation that replicates previous findings both from neuroimaging and lesion-studies. Those results further indicate that this cortical recruitment may at least be partly distinct from the cortical recruitment of *GAP-completion*.

Results also show that when *GAP-search* encounters a linguistic "obstacle"—as in Condition B (Event 1) *GAP-search*$_{indirect}$, and revealed in the B > D (Event 1) contrast - a different preferential activation pattern emerges involving BA 6 (medial superior), BA 8, right BA 24, and BA 32. At the same time, results from the double-subtraction $A_1>D_1$ vs. $B_1>D_1$ reveal no preferential activation suggesting that these two conditions are also very similar. So, in light of the ambiguous statistical results, we offer an interpretation constrained by previous neuroimaging and lesion-based observations. We propose here that these two sets of results indicate there may not be a categorical distinction between the cortical regions engaged in *GAP-search*$_{direct}$ vs. those engaged in *GAP-search*$_{indirect}$, instead the two reflect different patterns of activation within what is ultimately the same cortical network.

We thus interpret the LIF cortex preferential activation associated with *GAP-search*$_{direct}$ as resulting from an *interaction* of two factors involved in LDD resolution: (a) the prediction of a GAP, and (b) the possibility that the GAP be found within the syntactic and semantic contexts immediately after the RELPRO, that is, when nothing in the unfolding syntactic and semantic structure prevents the licensing of the GAP. These findings would thus represent independent neurological support for the existence of an active-filler (Clifton and Frazier, 1989; Frazier and Clifton, 1989; Fodor, 1995) that, crucially, is sensitive to the details of the linguistic context of the relative pronoun independently of the length of the dependency (Phillips et al., 2005).

Indeed, we take this pattern to reflect not necessarily a difference in search but a difference in quality of the search: when the parser is forced to use memory resources outside of the implementation of any specific linguistic mechanism -the delay caused by the parser's recognition that the expected GAP is *not* to be found in the current local constituent- those resources are recruited from cortical regions, most relevant BA 6 (SMA), which have been previously identified as participatory for language composition. The combined *GAP-search* pattern of results (direct plus indirect) would thus be reflecting the workings of two functional foci of the same linguistic *network*.

Support for this view is the observation that the LIF cortex and SMA have been traditionally connected, particularly in the focal-lesion literature (e.g., Benson, 1985; Tonkonogy, 1986; Vignolo, 1988; Naeser et al., 1989; Alexander et al., 1990; Goodglass, 1993). This would mean in turn that the LIF cortex is sensitive to the expedient resolution of the dependency, which will only happen when such resolution is allowed by the local linguistic context. If it is not, then the preferential activation shifts (or reduces) to pre-SMA–all, however, within the same pathway.

This interpretation is consistent with Santi and Grodzinsky (2012) regarding the connection between "prediction" and the LIF cortex. Yet, what our results show is that presence of "prediction" is not enough. For the LIF cortex to be fully engaged, it must continuously be tracking for "gap-viability" as composition unfolds[13].

Further elaborating on this issue, a reviewer suggests a perspective on the $B_1>D_1$ activation pattern that gives it a specific role—namely the suppression or inhibition of the direct *GAP-search* mechanism associated with the LIF cortex. In this view then, the monitoring action would presumably rely on the workings of the pre-SMA and in the situation where the *GAP-search* could not take place, due to the island, it would act on the LIF cortex to suppress or hold search activity. We agree that this possibility, though outside the scope of the present data, is interesting and consistent with all other roles independently attributed to the SMA (e.g., Schwartze et al., 2012). Moreover, it brings the debate not only to a discussion of networks but to the possible distinguishable roles that their individual components may play during real-time cognitive processing.

Indeed, we take the activation of the supplementary motor area (SMA) in the B>D (Event 1) contrast to be an important clue to the cortical recruitment of LDDs. Not only regarding *GAP-search*, but also *GAP-completion* as we will see below. Specifically, pre-SMA and SMA-proper (BA 6) have been independently shown to be involved in sensory-motor processing possibly manifested through a "gradient" in which sensory, nonsequential, suprasecond information is processed rostrally (recruiting pre-SMA cortex) while motoric, sequential, and subsecond information is processed more dorsally (Schwartze et al., 2012). Our present data are not fine-grained enough to reveal a dissociation between pre-SMA and SMA proper. However, the data do show the shared locus of activation to be on medial BA 6, suggesting the targeting of pre-SMA over SMA-proper. Such a locus would be consistent with the processing of non-motoric, non-sequential, suprasecond information such as that involved in the holding of the filler in memory, as it were,

[12]A reviewer asks us about our predictions for event 0. We note that no A/B > D difference was predicted and no difference was found at this segment. There are two reasons for this: (1) this early in the sentence, both A/B and D conditions show composition between the head noun and RELPRO, on the one case, and between the subject and the verb, on the other. Even though the nature of the composition that each carries is presumably different, we have no reason to expect that each will recruit visibly distinct cortical regions as a result. (2) Regarding the relative pronoun, even though it is true that by our definition *GAP-search* is triggered as soon as the RELPRO is retrieved, at this early point no structure has been built over which the search is to be carried out. So, even though *GAP-search* is triggered at event 0, it will not be visible until the embedded clause is beginning to be built. This is precisely what the event 1 contrast is intended to reveal.

[13]These findings also connect directly to the cause of the abnormally delayed gap-filling observed in Broca's real-time comprehension (e.g., Burkhardt et al., 2003, Love et al., 2008). The combined behavior of A>D and B>D suggests that Broca's impairment may not be rooted to a generalized problem in syntactic structure formation (brought about in turn to a slowing in lexical retrieval), as has been proposed, but instead to the inability to engage the filler in an active manner as composition progresses, that is, to keep track of the viability of the syntactico-semantic structure being built.

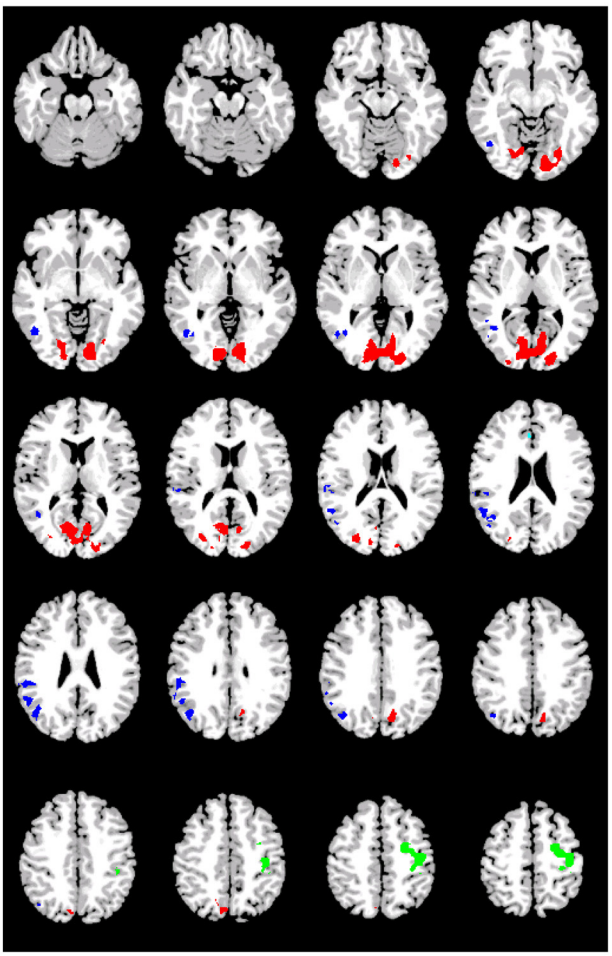


FIGURE 5 | Preferential activations for the conjunction of *GAP-completion* subtractions. Images are shown corrected at $p < 0.05$ in radiological format (LH is on the right). Green: BA6, primary motor and primary sensory connected with anterior cinculate (Acqua); Red: BA7, BA19, primary and association visual; Blue: right visual (association), right angular/ supramarginal gyri, and fusiform.

until the "GAP-unviable" segment has passed and active search can resume[14].

4.2. *GAP-completion*

Regarding *GAP-completion*, our findings from the simple subtractions show that this mechanism recruits the workings of a contiguous cortical region within the left fronto-parietal lobes (and non-overlapping with those associated with (direct) *GAP-search*) connecting supplementary motor area, precuneus, and portions of the left angular and supramarginal gyri and peristriate (BA 19). This observation is further supported by all relevant double subtraction and conjunction analyses. What emerges then is a coherent language "network," as all of these areas have been independently connected with related components of language processing. Most critically, they have been associated with lexically-driven composition, such as that involving subcategorization (Shetreet et al., 2009) and lexico-semantic selectional restrictions (e.g., Lai et al., 2014). Indeed, we conjecture that this pattern of preferential activation is part and parcel of the "Dorsal Stream" or "Dorsal Pathway" (Hickok and Poeppel, 2004, 2007; Friederici, 2009, 2012), which connects the frontal and left posterior cortices via the parietal lobe. To the extent that this network is seen to be involved in a mechanism such as *GAP-completion*, a mechanism that brings together syntactic, lexico-semantic, and discourse composition, it tells us that this cortical region is at least partly recruited during unification of interpretation. And this would also be consistent with a version of the Memory, Unification and Control model (e.g., Hagoort, 2005, 2014) whereby the true locus of *semantic* unification includes, most crucially, at least the pre-SMA. It is in this way that the LPST cortex is connected to LDD implementation: as a potential participating region in a larger network that supports real-time lexically-driven language composition which, by definition, also supports *GAP-completion*.

One additional advantage of the connection between *GAP-completion* with the dorsal pathway is that it affords a possible explanation for the long-standing observation regarding Conduction aphasia comprehension first reported in Caramazza and Zurif (1976). Specifically, Caramazza and Zurif (1976) report that patients with Conduction aphasia (a syndrome associated with damage to the arcuate fasciculus) exhibit *chance* performance in the comprehension of semantically reversible (object) relative-clauses. Such a pattern is indistinguishable from that shown for Broca's comprehension but claimed to emerge from different causes. Caramazza and Zurif (1976) further note that, like Broca's, the pattern shown by Conduction patients contrasts sharply with that exhibited by Wernicke's patients, who show performance that is not attributable to any one linguistic or processing factor. Here we reason that if *GAP-completion* is dependent on the workings of the dorsal pathway, presumably connected to the arcuate fasciculus, it explains why Conduction patients would be impaired in the interpretation of semantically reversible relative clauses, despite being able to carry out *GAP-search*[15]. In sum, we take the overall pattern accrued for all three Event 2, related double subtraction and conjunction contrasts to reflect components of this dorsal pathway, with BA 6 as a crucial area. This interpretation captures the normal-like performance by Wernicke's in online gap-filling constructions and suggests in turn that the LPST cortex activation from the imaging literature may not have been in connection to *GAP-search* proper.

In light of these findings, we are now able to address the questions posed in the introduction. What is the neurocognitive relation between *GAP-search* and *GAP-completion*? Answer: Their loci appears to be the LIF cortex and the (pre-)SMA, respectively. Do they rely on the workings of overlapping brain regions? Answer: The patterns we report show minimal overlap in recruitment. However, to the extent that at least the lower SMA

[14]Interestingly, BA 6 and supplementary motor cortex have both been associated to Broca's aphasia *and* Transcortical Motor Aphasia, indicating a potential connection between the two syndromes, whose functional implications are still not well understood (Naeser et al., 1989; Alexander et al., 1990).

[15]The authors thank Julius Fridiksson (p.c) for reminding us of this long-standing yet unexplained observation.

has been considered to be part of Broca's area, they are expected to functionally overlap. Our conjecture regarding the two areas [viable resolution (LIF cortex) vs. holding in memory (SMA)] is a proposal about how this overlap could take place. Can we associate *GAP-completion* to the LPST cortex, thus addressing the lesion-neuroimaging incongruence? Answer: we can if we understand Wernicke's area not as an isolated "language area" but as a part of a larger connectivity pathway "the dorsal stream" that connects Wernicke's area to the left fronto-parietal cortex including BA40, BA7 and the SMA. In line with the lesion-based literature, we conclude that LDD processing (defined in terms of *GAP-search* and *GAP-completion*) does not directly involve the preferential workings of Wernicke's area, but relies on areas that are functionally related to Wernicke's area.

Finally, if the effects reported reflect *GAP-search*, why are they observed mainly in the context of *object*-relative GAPs? Answer: what we find is that the effects reported reflect only an aspect of *GAP-search*, namely the requirement that the RELPRO be locally interpreted. But this only happens when *GAP-search* is being carried out over viable structure. So, it is as if the function of the LIF cortex is to monitor or keep track of the ability of the structure being composed to provide a GAP slot. In terms of our analysis, that amounts to keeping track of whether the selectional requirements of the RELPRO are being satisfied. As long as the composition signals that the GAP is incoming, the LIF cortex is fully engaged. From this perspective, then, the fact that this is observed mainly in object-GAP constructions is not a consequence of the grammatical feature *per se*, but of the fact that in these constructions, it takes longer for the RELPRO to be resolved as compared with subject-gap constructions, thus increasing the probability that the effect will be observed.

As a separate observation, our results also show that processing of memory-taxing sentential constructions (A and B) appear to systematically recruit the workings of the visual cortex (primary and association) areas (see Santi et al., 2015 and references therein for similar findings). We interpret this pattern separately for two reasons: (1) these areas are not traditionally associated with linguistic processing proper, and (2) this preferential activation was observed both during *direct GAP-search* and *GAP-completion*, suggesting that the areas in question are not showing sensitivity to a specific linguistic process.

In light of this, we connect these findings to independent observations regarding the visual system and linguistic load, particularly in relation to pupillometry measures (see Piquado et al., 2010 for a review and additional experimental evidence in relation to language processing load and the visual system). That observation has been shown not to be restricted to cognitive effort, but to extend even to physical effort (Zénon et al., 2014). Accordingly, we take the visual cortex activation pattern to reflect the increased attention (i.e., effort) that the implementation of the relevant linguistic tasks represents but whose source may not be strictly linguistic (see (Martínez et al., 1999; Posner and Gilbert, 1999; Petersen and Posner, 2012) for observations specifically regarding non-visually related attention load and its impact on the visual cortex). In this respect we note that the visual cortex activation was not observed during *indirect GAP-search* further supporting the possibility that during the building of structure that is non-viable for a GAP, no search is actually taking place. And this would make this segment of comprehension less cognitively taxing.

5. CONCLUSIONS: THE PRESENT RESULTS IN THE CONTEXT OF NEUROCOGNITIVE ARCHITECTURES

In this section, we connect our results to larger neurocognitive architecture models. In this respect, we consider three models which address syntactic and/or semantic composition, the sort presumably directly involved in *GAP-search* and *GAP-completion*. The first general observation is that whereas no one model accounts for the findings, each provides an insight into the larger pattern that the findings reflect. This gives us, then, the opportunity to focus on the common ground that each provides. This is what guides our discussion.

We start with Lau et al. (2008), who propose a model of semantic composition that could potentially involve LDD composition. In this model, the LIF cortex is connected to lexical retrieval. Interestingly our processing analysis of LDDs is lexically driven, and the key Event 1 contrast $A_1 > D_1$ does vary the presence of the relative pronoun. However, as we have seen, lexical retrieval differences alone do not account for the activation pattern: specifically, the results from $B_1 > D_1$, which also differ by the presence of the relative pronoun, do not show preferential LIF cortex activation. So, what is required in this model is a more precise treatment of the connection of lexical-retrieval to *GAP-search* in particular[16].

The second model we consider is Friederici (2012), which proposes that language composition, understood as the process of building a semantic representation through syntactic structure, recruits the workings of the LIF cortex. To the extent that *GAP-search* has been isolated from syntactic structure building through the subtraction process, the model predicts the LIF cortex will not be involved in this process, a prediction that is not supported by the evidence. For the same reasons, the model does successfully predict the absence of activation of BAs 44 and 45, particularly BA 44, in $B_1 > D_1$, which involves *GAP-search* but no hierarchical building. Friederici's (2012) model predicts no direct *GAP-search* in connection to the LIF cortex, because according to this model BAs 44 and 45 in particular are responsible for all syntactic structure building. Our results do not contradict this, but do point to the fact that BAs 44 and 45 must be additionally characterized as having specific compositional sensitivity, beyond generalized structure building. Finally, and as mentioned in the discussion, a most relevant aspect of Friederici's (2012) model (which also incorporates important insights from Hickok and Poeppel, 2004, 2007) involves the dorsal pathway, specifically,

[16]Relatedly, the more natural association with the workings of LDDs would be composition, since after all, the whole motivation for *GAP-search* is to compose the subject matrix nominal into the meaning of the embedded clause (and vice versa). But in the semantic model presented in Lau et al. (2008), this task is connected instead to the anterior temporal cortex (ATC) and angular gyrus, areas that are connectable instead to the ventral pathway and which, in our data, were only partly connected to *GAP-search* (in the form of the marginal activation of the temporal pole for $A_1 > D_1$).

Pathway I (also discussed in Friederici, 2009) and which connects the STG and BA 6 through the arcuate fasciculus. It is this pathway, we propose, that is responsible for compositional processes such as those represented by *GAP-completion*.

We reserve the end of this section for discussion of the Memory, Unification and Control (MUC) Model (e.g., Hagoort, 2005, 2014). To our knowledge, this is the only model that explicitly assumes lexically-driven processing and grammatical systems, a feature that our processing analysis of LDDs also assumes. The model also capitalizes on the notion of unification, which provides a processing-friendly approach to composition. Like the Friederici (2012) model, Hagoort's 2005, 2014 model proposes a divide within the LIF cortex separating BAs 45 and 47 from BAs 44 and 45 for semantic and syntactic unification/processing functions, respectively. Whereas our data do not speak to the functional articulation within the LIF cortex, they do reveal that both subregions can at least work in tandem, as in the case of the activation for direct *GAP-search*. This is a reasonable interpretation, given that direct *GAP-search* involves both semantic and syntactic computations. What is not clear at this point is how unification should be understood such that it will include *direct GAP-search* as a mechanism while simultaneously excluding *indirect GAP-search*; both processes that are on the one hand "dynamic" in nature, and on the other highly sensitive to the linguistic context of the GAP. Another pending question is the nature of the connection between the LIF cortex and SMA/lower parietal cortex. Under MUC, these two regions could be involved in the same larger processing network, and the SMA activation observed could be part of the dynamics of the network triggered in turn by the linguistic properties of the sentence. In this interpretation, LDDs allow us to localize not two regions, but a network with two foci reflected in these two mechanisms. Since our data cannot speak directly to this point, this proposal remains to be supported.

To conclude, the results presented here suggest a resolution of the imaging vs. lesion incongruence by showing the privileging of BAs 45, 44, and 47 (over BA 6 and parietal and parieto-temporal cortex, including the LPST cortex) in the process of *direct GAP-search* and by suggesting that the activation of LPST cortex reported in the neuroimaging literature is a manifestation of the workings of a network that supports other linguistic compositional processes associated instead with *GAP-completion*.

The results capture the inherent asymmetry between *GAP-search* and *GAP-completion* and explain why damage to the LIF cortex would dramatically impact the ability of the comprehension system to complete the dependency, even if the cortical regions involved in *GAP-completion* remained intact. By the same token, to the extent that the evidence presented here does *not* involve the left posterior superior temporal cortex at least directly, the results tell us why Wernicke's patients should not have issues in searching for and completing the GAP. Indeed, if our conjecture regarding the functional commitments of the SMA and the left lower parietal region (associated with *GAP-completion*) to compositional unification is correct, Wernicke's patients, who have been shown to have lexical retrieval problems, should not show problems in finding/completing the GAP but in unifying this information with the matrix clause into an interpretable string. Such a situation would lead to across-the board comprehension problems in these patients, a prediction that evidence from offline comprehension of these patients (in contrast to Broca's patients) consistently supports.

AUTHOR CONTRIBUTIONS

MP is responsible for the conception of the work and participated in each aspect of the project including experimental design, data acquisition and analysis, and drafting of the work. EF is responsible for stimuli generation and norming, subject recruitment, data acquisition, data analysis planning (e.g., timing file generation), interpretation and drafting of the work, and participated in the final approval of the version to be published. CL carried out all the data analysis and participated in the final approval of the version to be published. RC participated in the experimental design, data analysis and interpretation, drafting of the work, and in the final approval of the version to be published.

FUNDING

This project was funded by NSF BCS-0643266 awarded to MP and NSF-INSPIRE 1248100 awarded to MP, Ashwini Deo, Mokshay Madiman, and RC.

ACKNOWLEDGMENTS

We thank Jeetu Bhawnani and Hedy Sarofin for programming and technical support through the data acquisition stage. We also thank Emily Foster Hanson, Sara Sanchez-Alonso, and Muye Zhang for key assistance with aspects of the data analysis and data presentation. Finally, we are grateful to Edgar Zurif for much discussion on the gap-filling effect and the neurocognition of language composition; discussion that has directly impacted the approach taken here. All errors remain our own.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpsyg.2016.01434

REFERENCES

Alexander, M. P., Naeser, M. A., and Palumbo, C. (1990). Broca's area aphasias: aphasia after lesions including the frontal operculum. *Neurology* 40, 353–362. doi: 10.1212/WNL.40.2.353

Amunts, K., Weiss, P. H., Mohlberg, H., Pieperhoff, P., Eickhoff, S., Gurd, J. M., et al. (2004). Analysis of neural mechanisms underlying verbal fluency in cytoarchitectonically defined stereotaxic space: the roles of Brodmann areas 44 and 45. *Neuroimage* 22, 42–56. doi: 10.1016/j.neuroimage.2003.12.031

Avrutin, S. (1999). *Development of the Syntax-Discourse Interface*, Vol. 23. Norwell: Kluwer; Springer Science and Business Media. doi: 10.1007/978-94-017-1239-2

Avrutin, S. (2006). "Weak syntax," in *Broca's Region*, eds Y. Grodzinsky and L. Amunts (Norwell: Oxford University Press), 179–183. doi: 10.1093/acprof:oso/9780195177640.003.0004

Ben-Shachar, M., Hendler, T., Kahn, I., Ben-Bashat, D., and Grodzinsky, Y. (2003). The neural reality of syntactic transformations: evidence from functional magnetic resonance imaging. *Psychol. Sci.* 14, 433–440. doi: 10.1093/acprof:oso/9780195177640.003.0004

Ben-Shachar, M., Palti, D., and Grodzinsky, Y. (2004). Neural correlates of syntactic movement: converging evidence from two fMRI experiments. *NeuroImage* 21, 1320–1336. doi: 10.1016/j.neuroimage.2003.11.027

Benson, F. (1985). "Aphasia," in *Clinical Neuropsychology*, eds K. Heilman and E. Valenstein (New York, NY: Oxford University Press), 17–48.

Bornkessel-Schlesewsky, I., Grewe, T., Schlesewsky, M. (2012). Prominence vs. aboutness in sequencing: a functional distinction within the left inferior frontal gyrus. *Brain Lang.* 120, 96–107. doi: 10.1016/j.bandl.2010.06.004

Bresnan, J. (1982). *The Mental Representation of Grammatical Relations*, Vol. 1. Cambridge, MA: The MIT Press.

Bresnan, J. (2001). "The emergence of the unmarked pronoun," in *Optimality-Theoretic Syntax*, eds G. Legendre, S. Vikner, and J. Grimshaw (Oxford: Blackwell Publishers Ltd.), 45–50.

Burkhardt, P., Piñango, M. M., and Wong, K. (2003). The role of the anterior left hemisphere in real-time sentence comprehension: evidence from split intransitivity. *Brain Lang.* 86, 9–22. doi: 10.1016/S0093-934X(02)00526-6

Caramazza, A., and Zurif, E. (1976). Dissociation of algorithmic and heuristic processes in language comprehension: evidence from aphasia. *Brain Lang.* 3, 572–582. doi: 10.1016/0093-934X(76)90048-1

Chomsky, N. (1965). *Aspects of a Theory of Syntax*. Cambridge, MA: The MIT Press.

Chomsky, N. (1981). "Principles and parameters in syntactic theory," in *Explanation in Linguistics: The Logical Problem of Language Acquisition, Vol. 32, Lectures on Government and Binding: The Pisa Lectures* (Holland: Foris Publications), 75.

Clifton, C. Jr., and Frazier, L. (1989). "Comprehending sentences with long-distance dependencies" in *Linguistic Structure in Language Processing*, eds G. Carlson and M. Tanenhaus (Dordrecht: Kluwer), 273–317. doi: 10.1007/978-94-009-2729-2_8

Crain, S., and Fodor, J. D. (1985). "How can grammars help parsers?," in *Natural Language Parsing: Psychological, Computational Ad Theoretical Perspectives*, eds D. Dowty, L. Karttunen, and A. Zwicky (Cambridge: Cambridge University Press), 24–128.

Cooke, A., Zurif, E. B., DeVita, C., Alsop, D., Koenig, P., Detre, J., et al. (2002). Neural basis for sentence comprehension: grammatical and short-term memory components. *Hum. Brain Mapp.* 15, 80–94. doi: 10.1002/hbm.10006

Costello, A. L., and Warrington, E. K. (1989). Dynamic aphasia: the selective impairment of verbal planning. *Cortex* 25, 103–114. doi: 10.1016/S0010-9452(89)80010-3

Culicover, P. W., and Jackendoff, R. (2005). *Simpler Syntax*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780199271092.001.0001

Devlin, J. T., Matthews, P. M., and Rushworth, M. F. (2003). Semantic processing in the left inferior prefrontal cortex: a combined functional magnetic resonance imaging and transcranial magnetic stimulation study. *J. Cogn. Neurosci.* 15, 71–84. doi: 10.1162/089892903321107837

Embick, D., Marantz, A., Miyashita, Y., O'Neil, W., and Sakai, K. L. (2000). A syntactic specialization for Broca's area. *Proc. Natl. Acad. Sci. U.S.A.* 97, 6150–6154. doi: 10.1073/pnas.100098897

Evans, A. C., Marrett, S., Neelin, P., Collins, L., Worsley, K., Dai, W., et al. (1992). Anatomical mapping of functional activation in stereotactic coordinate space. *Neuroimage* 1, 43–53. doi: 10.1016/1053-8119(92)90006-9

Fiebach, C., Schlesewsky, M., and Friederici, A. (2002). Separating syntactic memory costs and syntactic integration costs during parsing: the processing of German wh-questions. *J. Mem. Lang.* 47, 250–272. doi: 10.1016/S0749-596X(02)00004-9

Fiebach, C., Schlewesky, M., Lohmann, G., von Cramon, D., and Friederici, A. (2005). Revisiting the role of Broca's area in sentence processing: syntactic integration versus syntactic working memory. *Hum. Brain Mapp.* 24, 79–91. doi: 10.1002/hbm.20070

Fillmore, C. J. (1988). "The mechanisms of "construction grammar," in *Annual Meeting of the Berkeley Linguistics Society*, Vol. 14 (Berkeley, CA), 35–55. doi: 10.3765/bls.v14i0.1794

Fodor, J. (1995). "Comprehending sentence structure," in *An Invitation to Cognitive Science: Language*, eds D. Osherson, L. Gleitman, and M. Liberman (Cambridge, MA: The MIT Press), 209–246.

Frazier, L., Clifton, C., and Randall, J. (1983). Filling gaps: decision principles and structure in sentence comprehension. *Cognition* 13, 187–222. doi: 10.1016/0010-0277(83)90022-7

Frazier, L., and Clifton, C. (1989). Successive cyclicity in the grammar and the parser. *Lang. Cogn. Process.* 4, 93–126. doi: 10.1080/01690968908406359

Frazier, L., and Flores d'Arcais, G. (1989). Filler-driven parsing: a study of gap filling in Dutch. *J. Mem. Lang.* 28, 331–344.

Friederici, A. D., Rüschemeyer, S. A., Hahne, A., and Fiebach, C. J. (2003). The role of left inferior frontal and superior temporal cortex in sentence comprehension: localizing syntactic and semantic processes. *Cereb. Cortex* 13, 170–177. doi: 10.1093/cercor/13.2.170

Friederici, A. D. (2009). Pathways to language: fiber tracts in the human brain. *Trends Cogn. Sci.* 13, 175–181. doi: 10.1016/j.tics.2009.01.001

Friederici, A. D. (2012). The cortical language circuit: from auditory perception to sentence comprehension. *Trends Cogn. Sci.* 16, 262–268. doi: 10.1016/j.tics.2012.04.001

Gibson, E. (1998). Linguistic complexity: locality of syntactic dependencies. *Cognition* 68, 1–76. doi: 10.1016/S0010-0277(98)00034-1

Gibson, E., and Warren, T. (2004). Reading-time evidence for intermediate linguistic structure in long-distance dependencies. *Syntax* 7, 55–78. doi: 10.1111/j.1368-0005.2004.00065.x

Goldberg, A. E. (1995). *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago: University of Chicago Press.

Goodglass, H. (1993). *Understanding Aphasia*. San Diego, CA: Academic Press.

Grodzinsky, Y. (1989). Agrammatic comprehension of relative clauses. *Brain Lang.* 37, 480–499. doi: 10.1016/0093-934X(89)90031-X

Grodzinsky, Y., Pierce, A., and Marakovitz, S. (1991). Neuropsychological reasons for a transformational analysis of verbal passive. *Nat. Lang. Linguist. Theory* 9, 431–453. doi: 10.1007/BF00135354

Grodzinsky, Y., Wexler, K., Chien, Y. C., Marakovitz, S., and Solomon, J. (1993). The breakdown of binding relations. *Brain Lang.* 45, 396–422. doi: 10.1006/brln.1993.1052

Grodzinsky, Y., Piñango, M. M., Zurif, E., and Drai, D. (1999). The critical role of group studies in neuropsychology: comprehension regularities in Broca's aphasia. *Brain Lang.* 67, 134–147. doi: 10.1006/brln.1999.2050

Grodzinsky, Y. (2000). The neurology of syntax: language use without Broca's area. *Behav. Brain Sci.* 23, 1–21. doi: 10.1017/S0140525X00002399

Grodzinsky, Y., and Friederici, A. D. (2006). Neuroimaging of syntax and syntactic processing. *Curr. Opin. Neurobiol.* 16, 240–246. doi: 10.1016/j.conb.2006.03.007

Hagoort, P. (2005). On Broca, brain, and binding: a new framework. *Trends Cogn. Sci.* 9, 416–423. doi: 10.1016/j.tics.2005.07.004

Hagoort, P. (2014). Nodes and networks in the neural architecture for language: Broca's region and beyond. *Curr. Opin. Neurobiol.* 28, 136–141. doi: 10.1016/j.conb.2014.07.013

Hickok, G. (1993). Parallel parsing: evidence from reactivation in garden-path sentences. *J. Psycholinguist. Res.* 22, 239–250.

Hickok, G., and Poeppel, D. (2004). Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. *Cognition* 92, 67–99. doi: 10.1016/j.cognition.2003.10.011

Hickok, G., and Poeppel, D. (2007). The cortical organization of speech processing. *Nat. Rev. Neurosci.* 8, 393–402. doi: 10.1038/nrn2113

Holmes, C. J., Hoge, R., Collins, L., Woods, R., Toga, A. W., and Evans, A. C. (1998). Enhancement of MR images using registration for signal averaging. *J. Comput. Assist. Tomogr.* 22, 324–333. doi: 10.1097/00004728-199803000-00032

Husband, E. M., Kelly, L. A., and Zhu, D. C. (2011). Using complement coercion to understand the neural basis of semantic composition: evidence from an fMRI study. *J. Cogn. Neurosci.* 23, 3254–3266. doi: 10.1162/jocn_a_00040

King, J., and Kutas, M. (1995). Who did what and when? Using word- and clause level ERPs to monitor working memory usage in reading. *J. Cogn. Neurosci.* 7, 376–395. doi: 10.1162/jocn.1995.7.3.376

Kluender, R. (1998). "On the distinction between strong and weak islands: a processing perspective," in *Syntax and Semantics 29: The Limits of Syntax*, eds P. Culicover and L. McNally (San Diego, CA: Academic Press), 241–279.

Lai, Y.-Y., Lacadie, C., Constable, T., Deo, A., and Piñango, M. M. (2014). "Complement coercion as the processing of aspectual verbs: evidence from self-paced reading and fMRI," in *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, eds P. Bello, M. S. Guarini, M. McShane, and B. Scassellati (Austin, TX: Cognitive Science Society), 2525–2530.

Lai, Y.-Y., Lacadie, C., Constable, T., Deo, A., and Piñango, M. M. (in press). "Complement coercion as the processing of aspectual verbs: evidence from self-paced reading and fMRI," in *Compositionality and Concepts in Linguistics and Psychology*, eds Y. Winter and J. Hampton (London: Springer).

Lau, E. F., Phillips, C., and Poeppel, D. (2008). A cortical network for semantics: (de)constructing the N400. *Nat. Rev. Neurosci.* 9, 920–933. doi: 10.1038/nrn2532

Love, T., Swinney, D., Walenski, M., and Zurif, E. (2008). How left inferior frontal cortex participates in syntactic processing: evidence from aphasia. *Brain Lang.* 107, 203–219. doi: 10.1016/j.bandl.2007.11.004

Martínez, A., Anllo-Vento, L., Sereno, M. I., Frank, L. R., Buxton, R. B., Dubowitz, D. J., et al. (1999). Involvement of striate and extrastriate visual cortical areas in spatial attention. *Nat. Neurosci.* 2, 364–369. doi: 10.1038/7274

Makuuchi, M., Bahlmann, J., Anwander, A., and Friederici, A. D. (2009). Segregating the core computational faculty of human language from working memory. *Proc. Natl. Acad. Sci. U.S.A.* 106, 8362–8367. doi: 10.1073/pnas.0810928106

Mars, R. B., Neubert, F., Noonan, M. P., Sallet, J., Toni, I., and Rushworth, M. F. S. (2012). On the relationship between the "default mode network" and the "social brain." *Front. Hum. Neurosci.* 6:189. doi: 10.3389/fnhum.2012.00189

MacDonald, M. C. (1989). Priming effects from gaps to antecedents. *Lang. Cogn. Process.* 4, 1–72. doi: 10.1080/01690968908406356

McElree, B., and Bever, T., (1989) The psychological reality of linguistically defined gaps. *J. Psycholinguist. Res.* 18, 21–35. doi: 10.1007/BF01069044

Matchin, W., Sprouse, J., and Hickok, G. (2014). A structural distance effect for backward anaphora in Broca's area: an fMRI study. *Brain Lang.* 138, 1–11. doi: 10.1016/j.bandl.2014.09.001

Naeser, M. A., Palumbo, C. D., Helm-Estabrooks, N., Stiassny-Eder, D., and Albert, M. L. (1989). Severe nonfluency in aphasia. *Brain* 112, 1–38. doi: 10.1093/brain/112.1.1

Nicol, J., and Swinney, D. (1989). The role of structure in coreference assignment during sentence comprehension. *J. Psycholinguist. Res.* 18, 5–19. doi: 10.1007/BF01069043

Nolfe, E. X., Voet, T., Jacobs, F., Dierckx, R., and Lemahieu, I. X. (2003). An open-source medical image conversion toolkit. *Eur. J. Nucl. Med.* 30(Suppl. 2):S246.

Papademetris, X., Jackowski, M. P., Rajeevan, N., DiStasio, M., Okuda, H., Constable, R. T., et al. (2006). BioImage suite: an integrated medical image analysis suite: an update. *Insight J.* 2006:209.

Petersen, S. E., and Posner, M. I. (2012). The attention system of the human brain: 20 years after. *Ann. Rev. Neurosci.* 35, 73–89. doi: 10.1146/annurev-neuro-062111-150525

Phillips, C. (2003). Linear order and constituency. *Linguist. Inq.* 34, 37–90. doi: 10.1162/002438903763255922

Phillips, C., Kazanina, N., and Abada, S. (2005). ERP effects of the processing of syntactic long-distance dependencies. *Cogn. Brain Res.* 22, 407–428. doi: 10.1016/j.cogbrainres.2004.09.012

Piñango, M. M., and Zurif, E. B. (2001). Semantic operations in aphasic comprehension: implications for the cortical organization of language. *Brain Lang.* 79, 297–308. doi: 10.1006/brln.2001.2492

Pinango, M. M. (2003). "Cortical reflections of two pronominal relations," in *Pronouns: Grammar and Representations*, eds H. Simon and H. Wiese (Berlin: Verlag), 232–252.

Piñango, M. M., and Burkhardt, P. (2005). "Pronominal interpretation and the syntax-discourse interface: real-time comprehension and neurological properties," in *Anaphora Processing: Linguistic, Cognitive and Computational Models*, eds A. Branco, T. McEnery, and R. Mitkoy (Amsterdam: John Benjamins Publishing Company), 221–238.

Piñango, M. M., and Zurif, E. (2015). "An evolving view of enriched semantic composition," in *Conceptual Semantics*, eds P. Csuri and I. Toivonen (Cambridge, MA: MIT Press), 167–186.

Piquado, T., Isaacowitz, D., and Wingfield, A. (2010). Pupillometry as a measure of cognitive effort in younger and older adults. *Psychophysiology* 47, 560–569. doi: 10.1111/j.1469-8986.2009.00947.x

Posner, M. I., and Gilbert, C. D. (1999). Attention and primary visual cortex. *Proc. Natl. Acad. Sci. U.S.A.* 96, 2585–2587. doi: 10.1073/pnas.96.6.2585

Santi, A., and Grodzinsky, Y. (2007). Working memory and syntax interact in Broca's area. *NeuroImage* 37, 8–17. doi: 10.1016/j.neuroimage.2007.04.047

Santi, A., and Grodzinsky, Y. (2008). The battle for Broca's region. *Trends Cogn. Sci.* 12, 474–480. doi: 10.1016/j.tics.2008.09.001

Santi, A., and Grodzinsky, Y. (2010). fMRI adaptation dissociates syntactic complexity dimensions *NeuroImage* 51, 1285–1293. doi: 10.1016/j.neuroimage.2010.03.034

Santi, A., and Grodzinsky, Y. (2012). Broca's area and sentence comprehension: a relationship parasitic on dependency, displacement or predictability? *Neuropsychologia* 50, 821–832. doi: 10.1016/j.neuropsychologia.2012.01.017

Santi, A., Friederici, A. D., Makuuchi, M., and Grodzinsky, Y. (2015). An fMRI study dissociating distance measures computed by Broca's area in movement processing: clause boundary vs identity. *Front. Psychol.* 6:654. doi: 10.3389/fpsyg.2015.00654

Schumacher, P. B., Piñango, M. M., Ruigendijk, E., and Avrutin, S. (2010). Reference assignment in Dutch: evidence for the syntax-discourse divide. *Lingua* 120, 1738–1763. doi: 10.1016/j.lingua.2009.12.003

Schwartze, M., Rothermich, K., and Kotz, S. (2012). Functional dissociation of pre-SMA and SMA-proper in temporal processing. *NeuroImage* 60, 290–298. doi: 10.1016/j.neuroimage.2011.11.089

Shapiro, L. P., and Levine, B. A. (1990). Verb processing during sentence comprehension in aphasia. *Brain Lang.* 38, 21–47. doi: 10.1016/0093-934X(90)90100-U

Shetreet, E., Friedmann, N., and Hadar, U. (2009). An fMRI study of syntactic layers: sentential and lexical aspects of embedding. *Neuroimage* 48, 707–716. doi: 10.1016/j.neuroimage.2009.07.001

Stowe, L. (1986). Evidence for on-line gap location. *Lang. Cogn. Process.* 1, 227–245. doi: 10.1080/01690968608407062

Stromswold, K., Caplan, D., Alpert N., and Rauch, S. (1996). Localization of syntactic comprehension by positron emission tomography. *Brain Lang.* 52, 452–473. doi: 10.1006/brln.1996.0024

Swinney, D., Ford, M., Frauenfelder, U., and Bresnan, J. (1988). "On the temporal course of gap-filling and antecedent assignment during sentence comprehension," in *Language Structure and Processing*, eds B. Grosz, R. Kaplan, M. Macken, and I. Sag (Standord, CA: Center for the Study of Language and Information).

Swinney, D., Zurif, E. B., and Nicol, J. (1989). The effects of focal brain damage on sentence processing: an examination of the neurological organization of a mental module. *J. Cogn. Neurosci.* 1, 25–37. doi: 10.1162/jocn.1989.1.1.25

Swinney, D., and Zurif, E. (1995). Syntactic processing in aphasia. *Brain Lang.* 50, 225–239. doi: 10.1006/brln.1995.1046

Swinney, D., Zurif, E., Prather, P., and Love, T. (1996). Neurological distribution of processing resources underlying language comprehension. *J. Cogn. Neurosci.* 8, 174–184. doi: 10.1162/jocn.1996.8.2.174

Tonkonogy, J. (1986). *Vascular Aphasia*. Cambridge, MA: MIT Press.

Van Valin, R. D., and LaPolla, R. J. (1997). *Syntax: Structure, Meaning, and Function*. Cambridge: Cambridge University Press. doi: 10.1017/cbo9781139166799

Vignolo, L. (1988). "The anatomical and pathological basis of aphasia," in *Aphasia*, eds E. C. Rose, R. Whurr, and M. A. Wyke (London: Whurr), 227–249.

Wise, R., Scott, S., Blank, S., Mummery, C., Murphy, K., and Warburton, E. (2001). Separate neural systems within Wernicke's area. *Brain* 142, 83–95. doi: 10.1093/brain/124.1.83

Zénon, A., Sidibé, M., and Olivier, E. (2014). Pupil size variations correlate with physical effort perception. *Front. Behav. Neurosci.* 8:286. doi: 10.3389/fnbeh.2014.00286

Zurif, E. B., Swinney, D., Prather, P., Solomon, J., and Bushell, C. (1993). An on-line analysis of syntactic processing in Broca's and Wernicke's aphasia. *Brain Lang.* 45, 448–463. doi: 10.1006/brln.1993.1054

Zurif, E., Swinney, D., Prather, P., and Love, T. (1994). Functional localization in the brain with respect to syntactic processing. *J. Psycholinguist. Res.* 23, 487–498. doi: 10.1007/BF02146687

Zurif, E., Swinney, D., Prather, P, Wingfield, A., and Brownell, H. (1995). The allocation of memory resources during sentence comprehension: evidence from the elderly. *J. Psycholinguist. Res.* 24, 165–182. doi: 10.1007/BF02145354

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer WM and handling Editor declared their shared affiliation, and the handling Editor states that the process nevertheless met the standards of a fair and objective review.

Copyright © 2016 Piñango, Finn, Lacadie and Constable. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 04 September 2015
doi: 10.3389/fpsyg.2015.01301

Cross-linguistic evidence for memory storage costs in filler-gap dependencies with wh-adjuncts

Arthur Stepanov and Penka Stateva*

Center for Cognitive Science of Language, University of Nova Gorica, Nova Gorica, Slovenia

This study investigates processing of interrogative filler-gap dependencies in which the filler integration site or gap is not directly subcategorized by the verb. This is the case when the wh-filler is a structural adjunct such as *how* or *when* rather than subject or object. Two self-paced reading experiments in English and Slovenian provide converging cross-linguistic evidence that wh-adjuncts elicit a kind of memory storage cost similar to that previously shown in the literature for wh-arguments. Experiment 1 investigates the storage costs elicited by the adjunct *when* in Slovenian, and Experiment 2 the storage costs elicited by *how quickly* and *why* in English. The results support the class of theories of storage costs based on the metric in terms of incomplete phrase structure rules or incomplete syntactic head predictions. We also demonstrate that the endpoint of the storage cost for a wh-adjunct filler provides valuable processing evidence for its base structural position, the identification of which remains a rather murky issue in current grammatical research.

Keywords: parsing, filler-gap dependency, thematic role, wh-adjunct, Active Filler Strategy, Slovenian

OPEN ACCESS

Edited by:
Matthew Wagers, University of California, Santa Cruz, USA

Reviewed by:
Dave Kush, Haskins Laboratories, USA
Pavel Logacev, University of Potsdam, Germany

****Correspondence:***
Arthur Stepanov, Center for Cognitive Science of Language, University of Nova Gorica, Vipavska 13, 5000 Nova Gorica, Slovenia
arthur.stepanov@ung.si

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: *27 March 2015*
Accepted: *16 August 2015*
Published: *04 September 2015*

Citation:
Stepanov A and Stateva P (2015) Cross-linguistic evidence for memory storage costs in filler-gap dependencies with wh-adjuncts. Front. Psychol. 6:1301. doi: 10.3389/fpsyg.2015.01301

Introduction

It has long been known that processing syntactic dependencies, in which two elements are syntactically related and linearly separated by intervening material, may be difficult for sentence comprehenders. An early study in Wanner and Maratsos (1978) showed that such difficulties arise, in particular, in processing incomplete filler-gap dependencies, in which a wh-phrase is syntactically related to the gap in the subcategorized position of the verb:

(1) Which book do you think that Colin recommended _ to the librarian?

Such long-distance dependencies are a source of syntactic complexity that the parser has to deal with over and above what is required for processing phrase structure and specific lexical items. One line of explanation for these difficulties faced by the human parser is that syntactic dependencies of this kind incur a tax on the working memory needed to temporarily store the antecedent or predictor, until a suitable element with which it can be associated is encountered in the partially processed input (Chomsky and Miller, 1963; Abney and Johnson, 1991; Gibson, 1991, 1998; Stabler, 1994; Lewis, 1996). Thus, in (1), the filler (*which* book) must somehow be temporarily stored in working memory until a suitable integration point is found. The working memory tax associated with storage cost leads to particular behavioral effects such as increased response times, or specific brain activity patterns at the neural level. Chen et al. (2005), in one of their self-paced reading experiments, manipulated the type of structure between a relative clause, where a wh-dependency is established, and a sentential complement, where it is not, as in the following examples:

(2) a. The announcement [that the baker from a small bakery in New York City received the award] helped the business of the owner.
b. The announcement [which the baker from a small bakery in New York City received ___] helped the business of the owner.

The critical region was the subject NP "the baker from a small bakery in New York City." By hypothesis, temporarily storing the wh-filler "which" initiates an incomplete syntactic dependency and a prediction of a subcategorizing or thematic-role assigning verb to complete the dependency. Thus, assuming that both conditions otherwise involve the same amount of lexical integrations, it is predicted that the critical region in (2b) should elicit greater reading times, showing a distributed slowdown effect, as opposed to (2a) where no wh-dependency is initiated, because of the storage effect[1]. Indeed, Chen et al. (2005) observed that reading times in (2b) were greater than in (2a) (see also Gordon et al., 2002; Grodner et al., 2002 for related studies).

In the Event Related Potentials paradigm, storage costs for wh-fillers appear as a modulation of left anterior negativity (a negative voltage deflection in the frontal, often left-lateralized, regions of the scalp) spread over the region between the filler and the gap (Kluender and Kutas, 1993; King and Kutas, 1995), followed also by a P600, a positive deflection effect at the gap or pre-gap position (Kaan et al., 2000). Phillips et al. (2005) also observed that the sustained negativity persisting throughout the wh-dependency until the point of its completion is independent of the length of a filler-gap dependency, appearing both in short-distance (single clause) and long-distance (multi-clausal) wh-dependencies. The authors interpret this sustained negativity as a reflection of the cost of holding the wh-phrase in working memory. A similar pattern of Event Related Potentials was also observed in German and Japanese (Fiebach et al., 2002; Ueno and Garnsey, 2007).

In the present study, we investigate whether wh-adjuncts like *how quickly, when* and *why* elicit a similar kind of storage cost as wh-arguments do. Wh-adjuncts are notably different from wh-arguments in ways that directly affect processing. Semantically, adjuncts can never have a basic semantic type: canonically, they may function as predicates of events in the sense of event semantics (Davidson, 1980), or proposition or event modifiers in the sense of compositional semantics (Heim and Kratzer, 1998). In that capacity, wh-adjuncts are special in that their base (that is, semantically and syntactically determined) position is not predicted by subcategorization and/or thematic role assigning properties of the verb. The absence of direct association with the verb raises an a priori possibility that wh-adjuncts do not instantiate a filler-gap dependency at all: rather, they could be simply processed in their surface position. This lines up with certain grammatical theories that do not postulate syntactic displacement or dependency in the case of wh-adjuncts, as opposed to wh-arguments, and assume that wh-adjuncts are base-generated in their surface position (see, e.g., Hukari and Levine, 1995 for an overview, and the discussion below). We can then ask the following:

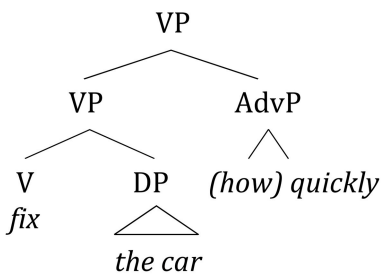
(3) a. Do wh-adjuncts instantiate filler-gap dependencies similar to wh-arguments, in the absence of thematic and/or subcategorization association with the verb?
b. Do all adjuncts incur similar storage costs?

As will become clear from the following discussion, we believe that the answer to (3a) is "yes," but the answer to (3b) is most likely "no." We do expect storage costs for (most) wh-adjuncts because their surface position must be syntactically linked to a base position linearly separated from it by intervening material. At the same time, recent advances in syntactic theory inform us that adjuncts differ with respect to their base positions. Consequently, we might expect different adjuncts to display different storage costs. The following section provides a basic overview of the major syntactic peculiarities that enter into processing considerations regarding filler-gap dependencies involving wh-adjuncts, and outlines the challenges presented by wh-adjuncts in light of the existing theories of storage costs and filler-gap dependencies in general.

Base/Integration Points of Wh-adjuncts

Syntactically, an adjunct is realized as a sister to an abstract syntactic node denoting the predicate that the adjunct modifies. An example of modifying an event predicate is shown in (4):

(4)



The actual attachment site of an adjunct may vary depending on the type of predicate that it modifies. Current theoretical research recognizes a multitude of positions in the syntactic structure of the sentence, where adjuncts of various semantic types may appear (e.g., Cinque, 1999). For the present purposes, we may adopt a simplified version of that typology and pinpoint at least four classes of adjuncts as in **Table 1** (cf. also Ernst, 2001; Rizzi, 2001).

The order of listing the adjunct types in **Table 1** roughly corresponds to their base structural position in the syntactic tree, or closeness to the root node. The attachment site of each adjunct type is determined by the corresponding phrase structure rule (e.g., S → AdvP S, or VP → VP AdvP) and corresponds to speakers' semantic intuitions. Structurally the "highest" are speaker oriented adverbs which we will not consider in this study. The next highest position is occupied by reason adjuncts and their corresponding *wh*-counterpart *why* (see also Experiment 2 for further details), placed above the sentential S node, in the domain of COMP (or Complementizer

[1] A reviewer points out that (2a) may be compatible with a filler-gap parse up to the item *the award*, which might affect comparison of storage costs between the two sentences. This potential confound is avoided in our design of Experiment 2 with the English materials (see below).

TABLE 1 | Grammatical typology of adjuncts and their wh-counterparts.

Adjunct types	Examples, non-wh-	Simplex wh-adjuncts	Complex wh-adjuncts
Speaker-oriented adjuncts, or evidentials	*Clearly, obviously, fortunately, in my opinion*	–	*How obviously, how fortunately*
Reason adjuncts	*Because*-adjuncts	*Why*	*For what reason*
Subject-oriented adjuncts	*Probably*, temporal, and spatial adjuncts	*When, where*	*In which room, after which holidays*
VP-oriented adjuncts	*Quickly, often, with a hammer* etc.	*How*	*How quickly, by what means*

Phrase in current syntactic terminology). This is followed by subject-oriented adjuncts that are adjoined to the S node (or Infl Phrase in many contemporary syntactic theories), not considered in this study either. Finally, VP-oriented adjuncts are adjoined to VP, thus are relatively "low" in the syntactic structure.

A major consequence of this diversity of the syntactic and semantic properties of wh-adjuncts is that, in contrast to arguments, the linear position of the adjunct in the sentence cannot be reliably predicted from the linear position of the corresponding verb and the information about the canonical word order in a language. Rather, the association of an adjunct to the verb in this case can only be *loose and indirect* [note that the lower VP constituent in (4) can syntactically be arbitrarily complex; other material can also be inserted by iterating the VP node]. In parsing, this translates into a state of affairs whereby a wh-filler cannot be reliably associated with a specific lexical stimulus, such as the verb. Rather, it may appear at an arbitrarily long distance from it[2]. With regard to VP-modifying wh-adjuncts, one can envision two possibilities as to where their integration point might lie.

The first possibility is that, despite the irrelevance of the thematic/subcategorization information, the parser follows some lexically-driven strategy to integrate the adjunct at or near the verb, similarly to wh-argument dependencies. Indeed, some grammatical models postulate a close syntactic relationship between (wh-)adjuncts and the verb outside the realm of the thematic relations. Such postulated relationship usually has a featural character: some morphosyntactic feature on the adjunct and a feature on its licensing head such as V or Infl must match or agree. In such approaches, different features may correspond to different adjuncts (Travis, 1988; Laenzlinger, 1996; Ernst, 2001). It is thus possible that the featural association of a wh-adjunct and the verb is reflected in the processing pattern, resembling or approaching the pattern of association of wh-argument fillers with corresponding verbs.

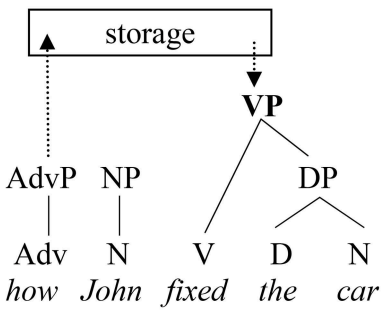
[2]Furthermore, there is an issue of *directionality*: some adjuncts may be left-adjoined or right-adjoined, following restrictions that may have syntactic, semantic or prosodic nature (for instance, prosodically heavy adjuncts tend to be right-adjoined, see Ernst, 2001). Even within the same class of VP-adjoined adjuncts, adverbs like *quickly* allow at least two grammatically acceptable positions compatible with the sisterhood to VP (with no easily discernible difference in meaning), while other manner adverbs like *poorly* do not allow such dual positioning (Bowers, 1993):

(i) a. John fixed the car quickly
b. John quickly fixed the car
(ii) a. John fixed the car poorly
b. *John poorly fixed the car

The second possibility is that the base position of the wh-adjunct (and, correspondingly, its integration point in a filler-gap dependency) requires online computation of an abstract syntactic node [cf. the lower VP in (4)] as a way of identification of the event-denoting predicate hosting the wh-adjunct, as well as some predictive information about its linear position with respect to that node. To illustrate schematically, let us assume a parsing algorithm with a storage component, using both bottom-up and top-down strategies and consulting the phrase structure module of the grammar. If such an algorithm is used to process the embedded part of a sentence like (5), the adverbial phrase *how* would be stored (possibly along with a pointer associating it with the relevant phrase structure rule; see also the discussion of the SLASH feature below) until the VP node is constructed online. Completion of the VP would trigger a subsequent application of the rule VP → VP AdvP, retrieving the stored phrase and integrating it at a grammatically permissible site [cf. (4)].

(5) I didn't know how John fixed the car

(6)



Formally, the difference between the two possibilities lies in the syntactic category of an element combining with an adjunct: a lexical head such as V^0 vs. a phrasal category such as VP, in the correspondent grammatical rule(s) guiding integration of the adjunct during online processing. In the absence of thematic and/or subcategorization criteria for integration, syntax seems to be a major relevant cue for predicting the integration site in this case (possibly supported by other cues such as plausibility). Consequently, no lexically-based strategy would be at issue; rather, the relevant integration algorithm would have to make reference to the syntactic category information in determining the integration point (see also Gibson, 1998, 2000 concerning processing costs of structural integrations). Similar considerations apply with respect to the processing costs of temporary storage of an wh-adjunct, as discussed below.

In sum, the grammatical distribution of (wh-)adjuncts in general is much more complex than that of (wh-)arguments.

Correspondingly, general processing predictions with respect to the base site of the filler-gap dependency headed by a wh-adjunct can hardly be formulated. Rather, gap predictions must be formulated item-specifically. In the absence of thematic/subcategorization information, such predictions have to take into account, at the very least, the semantic type of an adjunct, on the one hand, and the phrase structure rules generating it, on the other.

Storing Wh-adjuncts: Theoretical Predictions

From the storage perspective, investigating wh-adjunct dependencies is theoretically illuminating in at least two different respects. The first one concerns the role of the thematic factor in current theories of storage costs, and more generally, the role of lexically-based strategies of computation of these costs. The second regards the interaction and potential convergence of the processing and grammatical predictions concerning the endpoint of the storage costs, also in the context of the Active Filler Strategy. Below we consider these aspects in turn.

The Lexically-based vs. Syntactically-based Views on Storage Costs

Current processing theories of incomplete filler-gap dependencies focus, explicitly or implicitly, on the issue of the temporary storage of the wh-filler, its integration into the syntactic structure of the input, or both. For instance, the memory-based accounts of filler-gap dependencies (Gibson, 1998, 2000) assume that integration and storage incur separate memory costs, and the overall processing cost of a filler gap dependency is a function of both of these measures. These accounts are based on the general idea that integrations and storage share the same pool of memory resources and that this pool of resources is limited; consequently, exceeding the set limit at some point slows down performance (Baddeley, 1990; Just and Carpenter, 1992; Lewis, 1996)[3]. From the evidence accumulating from the previous psycho- and neurolinguistic studies (see Introduction), it can be conjectured that integration costs are associated with behavioral or neurophysiological markers showing up at certain discrete points of parsing, usually at or around the predicted gap site, whereas storage costs reveal themselves as extended intervals of specific behavioral or neural response over a range of input that coincides, or is very close to, the area between the filler and the gap, in the form of a reading slowdown or increased sustained voltage deflection in the ERP signal.

The theories of storage proposed to date differ with respect to the question as to what processing units may incur a memory storage cost (see Chen et al., 2005 for review). It has been proposed that storage be measured in units such as incomplete clauses (Kimball, 1973), incomplete phrase structure rules (Yngve, 1960; Chomsky and Miller, 1963), incomplete thematic role assignments (Hakuta, 1981; Gibson, 1991), incomplete Case dependencies (Stabler, 1994), and predicted syntactic heads (Gibson, 1998, 2000). For instance, under the theory taking incomplete phrase structure rules to be relevant storage units, a center-embedded structure as in *This is the malt that the rat that the cat that the dog worried killed ate* elicits storage costs quantifiable in terms of the number of the phrase structure rules such as S → NP VP that have to be kept in memory as more embedded material is processed. Similarly, under the predicted syntactic head theory, storage is quantified in terms of the number of syntactic heads expected to complete a dependency.

We believe that investigating wh-adjunct dependencies may reliably distinguish between these theories. In particular, the theories that take storage units to be incomplete thematic role assignments (Hakuta, 1981; Gibson, 1991) predict that wh-adjuncts should not elicit storage costs, simply because there are no thematic roles associated with them that need to be stored. A similar prediction is made by the theories that take the relevant storage units to be incomplete Case dependencies (Stabler, 1994): wh-adjuncts are usually adverbials or prepositional phrases; as such, they are not subject to the Case requirement, hence no Case information needs to be stored. Thematic roles, lexical head predictions and/or Case predictions are all part of the class of theories that take lexical factors as the cornerstone of relevance when it comes to computing storage costs.

On the other hand, theories that assume that incomplete phrase structure rules are stored during processing (see above) do predict storage costs for wh-adjuncts similarly to wh-arguments. As Chen et al. (2005) point out, these theories can be adapted to handle storage costs in filler-gap wh-dependencies utilizing the analytical tools of, e.g., head-driven phrase structure grammar (Pollard and Sag, 1994) and/or generalized phrase structure grammar (Gazdar et al., 1985). In these models, the mediation between the wh-filler and the verb is achieved via the SLASH feature which may propagate across syntactic nodes or rules down to the integration site thematically associated with the verb, and thus marks the path of the wh-dependency (Pollard and Sag, 1994; Sag and Fodor, 1994). The crucial point here is that the SLASH feature is insensitive to the syntactic category of the missing constituent. Thus, all else equal, it predicts the storage costs for wh-arguments, as well as for wh-adjuncts.

Similarly, theories that compute syntactic predictions in terms of expected syntactic heads (Gibson, 1998, 2000) predict storage costs for wh-adjuncts as well as wh-arguments. In contrast to the incomplete phrase structure rule theories, the expected syntactic head model does not make direct reference to hierarchical constituent structure, but only to its lowest level of representation, the level of syntactic heads. Let us assume for the moment that the gap expectation for *how* is at the point linearly following the lower VP constituent, as in (4). For concreteness, let us also assume the algorithm for quantitative estimation of storage costs based on predicted syntactic heads, as in Gibson (2000). Consider the relevant storage costs of the sentence in (7)

[3]Distinguishing the integration and storage costs empirically is not a trivial task. For instance, a classic explanation of the contrast in the parsing difficulty between center-embedded structures and the corresponding right-branching structures is that the former require a greater amount of storage space as opposed to the latter, and since the amount of memory resources available for sentence processing is limited (Miller, 1956), the difficulty arises at the point when the memory capacity is exceeded (Chomsky and Miller, 1963; Abney and Johnson, 1991; Lewis, 1996). But, as Gibson (1998) notes, there exists an alternative explanation that nested structures, by their nature, always require longer distance integrations between the respective syntactic heads, hence higher processing costs, than the right-branching structures. This caveat is obviously relevant to structures manifesting filler-gap dependencies as well.

(concentrating on the embedded clause) that can potentially be assigned in this model, with and without the wh-adjunct, all else being equal:

(7) I didn't know that/how you fixed the car yesterday

		Input word					
		…how	you	fixed	the	car	yesterday
(8)	Storage cost	3	2	2	2	1	0
		…that	you	fixed	the	car	yesterday
		2	1	1	1	0	0

Given the nature of adjuncts as event modifiers, the storage cost at the point when *how* is processed will be 3. This reflects, in addition to the associated gap position, two more heads to describe an event (e.g., *it happened* or *John arrived*). Upon encountering the subject *you*, the storage cost value is reduced to 2, expecting a predicate and (still) a gap. At the point when *the* comes, the parser expects the noun and a gap position. Finally, at *car*, only the gap position is expected. In the non-wh version, storage costs are correspondingly reduced.

To sum up, wh-adjuncts may provide important evidence to distinguish between theories that place crucial weight on the lexical properties of fillers and those that do not. If wh-adjuncts incur storage costs, that would argue in favor of the latter type of theories. The present study seeks to provide such evidence.

Endpoint of the Storage Costs

The second interesting aspect of storage costs has to do with understanding the way storage costs for wh-adjuncts are related to the two potential grammatical possibilities for the integration site considered above. The existing theories that take storage and integration both to be active components of the working memory, largely take it for granted that the integration site marks the retrieval of the wh-filler, hence the endpoint of the storage costs. For wh-arguments, the endpoint of the storage costs at the grammatically expected point (e.g., verb for the object filler) would not be particularly surprising. For wh-adjuncts, things are not that trivial. Current grammatical theories do not always offer reliable clues as to the end-/integration point of the wh-adjunct in the syntactic structure, due to their loose and mobile syntactic character that follows from the lack of thematic and/or subcategorizational anchors. The situation gets even more complicated considering that the integration site is different for different wh-adjuncts in the same language (see footnote 2 and Experiment 2). Naturally, wh-adjunct dependencies appear somewhat more elusive for tracking with current experimental methods than their wh-argument counterparts. The endpoint of the storage costs in this situation could then provide important processing evidence for grammatical theory, to the extent that it demarcates a likely integration site.

In this respect, it is also interesting to investigate the role of the Active Filler Strategy, a parsing strategy which assigns high priority to integrating the filler at the earliest point allowed by the grammar (see Fodor, 1978; Frazier and Clifton, 1989; de Vincenzi, 1991). The Active Filler Strategy bears on the "filled gap" effect of integrating a wh-argument like subject or object with its corresponding syntactic position in the input, as in the following sentences from the self-paced reading study in Stowe (1986):

(9) a. My brother wanted to know if Ruth will bring us home to Mom at Christmas.
b. My brother wanted to know $[\text{who}]_i$ G_i will bring us home to Mom at Christmas.
c. My brother wanted to know $[\text{who}]_i$ Ruth will bring $^*G_i$ us home to G_i at Christmas.

Longer reading times were reported at *us* in (9c) compared to (9a) and (9b). This is expected if the position of *bring* is the earliest potential position where the object wh-filler *who*, temporarily kept in the memory, can be integrated. That state of affairs causes reanalysis. In contrast, (9a) and (9b) involve no such reanalysis[4]. This and other studies investigating the Active Filler Strategy are usually based on processing verbal arguments. The interest in investigating the role of this strategy in processing wh-adjuncts consists primarily in determining (a) whether it is operative at all; and (b) if it is, what sort of evidence the parser uses in order to determine the earliest position, in the absence of thematic or lexically-oriented cues.

In the present study we report two self-paced reading experiments targeting wh-adjunct dependencies in Slovenian and English. Specifically, we focus on two examples of structurally low, VP-modifying adjuncts, as well as an example of a structurally high (reason) adjunct. Low or VP-modifying adjuncts offer a good source of evidence pertaining to research question (3a) above. Since the canonical word order in SVO languages presupposes some non-trivial distance between the occurrence of the filler and the VP in the linear representation of the interrogative sentence, a filler-gap dependency in this case can potentially be identified in parsing by a storage effect which extends across some part or all of the corresponding range in the input, much along the lines of the previous studies of storage costs incurred by wh-arguments. This is not the case with structurally high wh-adjuncts whose integration sites are likely to be close to their surface position or even identical to it (see also Section Storage Cost Predictions for *why*). Utilizing this idea, Experiment 1 aims at detecting a filler-gap dependency with the VP-modifying adjunct *kdaj* "when" in Slovenian, as well as investigating the endpoint of such dependency. Experiment 2, using English materials, addresses research question (3b) as well as (3a). It compares the storage cost patterns of the structurally low *wh*-adjunct *how quickly* and the structurally high *wh*-adjunct *why*, asking whether these processing patterns differ in a way that correlates with the syntactic and semantic properties of these two w*h*-items. This experiment also targets the endpoint of a filler-gap dependency in greater detail.

[4]Note that if *who* is ambiguous between subject and object, then the Active Filler Strategy also predicts increased reading times over the subject position (*Ruth*) in cases like (9b). In her study, Stowe found no increased reading times over the subject. However, Lee (2004) argues that a filled-gap effect appears when more material is added in between the filler and the (subject) gap, and thus sufficient time is available to the parser.

Inclusion of Slovenian in our study was justified on several grounds. Aside from the obvious benefit of expanding the empirical database of processing storage cost effects cross-linguistically, and the fact that Slovenian usually receives little attention in behavioral psycholinguistics, working with certain kinds of wh-adjunct dependencies in Slovenian turns out to be preferable as some wh-adjuncts in Slovenian are free of inherent lexical ambiguities typical of their counterparts in other languages, including English (as is the case of *when*, see below). This allows for a cleaner experimental design, avoiding potential confounds in the construction of stimuli. The present study is also the first, to our knowledge, comparing storage costs in filler-gap dependencies in two languages within the same experimental setup. Because of that, we show that storage costs elicited by wh-adjuncts are a language-independent phenomenon, a naturally expected result in the context of the general inquiry into the nature of the human parsing system.

Experiment 1: Slovenian *kdaj* ("when")

Experiment 1 is a self-paced reading study in which we investigate potential storage costs elicited by the structurally low, VP-modifying *wh*-adjunct *kdaj* "when" in Slovenian. If a wh-adjunct like *when* in the beginning of the sentence instantiates a filler-gap dependency as wh-arguments do, thus functioning as a filler, we may expect a storage cost effect extending across the range in the input which is commensurable with the structural distance between the filler and its corresponding gap in the VP area. The experiment tests this scenario.

In addition, Experiment 1 aims to shed light on the issue regarding the endpoint of the storage cost for *when*. As noted above, there are two main theoretical possibilities to consider with respect to this endpoint. One is that the dependency terminates at the verb, as is the case for wh-arguments. The other is that the dependency terminates at some point predicted by phrase structure rules for VP. Regarding the latter possibility, in a sentence with a transitive verb it makes sense to expect a gap at or after the relevant part of the argument structure, viz. verb plus object, is processed [cf. (4) above]. The working assumption, trivial for wh-arguments, but non-trivial for wh-adjuncts, is that the end of the storage costs (that is, a point where reading times are equalized compared to the input not involving a wh-dependency) signals the gap site. Based on the results in Chen et al. (2005) for wh-arguments, we thus expect to see a region of increased reading times to last until either the first or the second suspected gap site:

(10) a. I didn't know when John bought (G1?) the newspaper (G2?) in the kiosk
b. I didn't know that John bought the newspaper in the kiosk.

If the endpoint of the storage costs is at the verb, this would support the approach to storage costs based on the featural association of the verb and the adjunct (see above). On the other hand, if the endpoint of the storage costs is at or after the direct object, this would be consistent with the phrase structural theories, as well as with the predicted syntactic head theories of storage costs.

Note that *when* in English is ambiguous. It can be used in its truly interrogative sense (cf. *Peter asked when the parcel would arrive*) or in another, related, but non-interrogative, guise (cf. *Peter left when the parcel arrived*). When used in embedded contexts, the truly interrogative version of *when* is selected by a particular class of verbs such as *ask, wonder,* or *know*. When used in its non-interrogative sense, *when* does not need to be selected at all. It is often difficult to distinguish these two usages in English and other languages which use a single lexical item for both. In Slovenian, on the other hand, the two usages of *when* are lexically disambiguated: *kdaj* is used in the respective interrogative contexts, and *ko* in non-interrogative ones. Because of that, Slovenian is an excellent choice to study the online behavior of the interrogative *when* and rule out potential confounds caused by its non-interrogative usage (which may not trigger a wh-dependency at all).

We thus concentrated on *kdaj* in Slovenian, and compared performance over the region corresponding to the argument structure (in bold) in simple embedded wh-questions such as the following (further description of the items involved is discussed in the Section Materials below):

(11) a. Kritik je potrdil, da **je umetnik izdelal tisti**
Critic is confirmed that is artist created this
koš v svoji delavnici.
basket in his workshop
"The critic has confirmed that the artist created this basket in his workshop"
b. Kritik je potrdil, kdaj **je umetnik izdelal tisti**
Critic is confirmed when is artist created this
koš v svoji delavnici.
basket in his workshop
"The critic has confirmed (the date) when the artist created this basket in his workshop"

Since in (11a) the argument structure ends at the point ***koš***, this is the point where we expect the storage costs to equalize with those observed at the same point in (11b). Following the critical region was either a locative PP (e.g., *v svoji delavnici* "in his workshop") or a further optional specification of the object noun [e.g., (*koš*) *božičnih daril* "(basket) of X-mas presents"], which contained two to three words.

Methods

Participants[5]

Seventy-four monolingual speakers of Slovenian from the academic communities of the University of Nova Gorica and University of Ljubljana volunteered to participate in the experiment for no material compensation. All participants were naïve to the purposes of the study.

[5]The experiments in this study were carried out in accordance with the Declaration of Helsinki and the existing European and international regulations concerning ethics in research. All participants gave an informed consent prior to the beginning of testing.

Materials and Methodology

Twenty-four sets of sentences, each with the two conditions described above, were carefully constructed. Since we were interested in evaluating the actual "boundaries" of the filler-gap dependencies reflected in online storage costs, the sentences in the two conditions were exactly identical except for the value of the embedded clausal head, or Complementizer: this value was either a declarative *da* "that" or interrogative *kdaj*. To control for length of a wh-dependency, all sentences were made exactly 12 words long. Each sentence began with an introductory part involving a one-word subject, a past tense auxiliary and a main verb [cf. (11)]. The main verbs were carefully chosen so that they may embed either a wh-interrogative clause, or a declarative *that*-clause. In English, typical representatives of this class of verbs are *know* and *figure out* (e.g., *I know that the guests came* vs. *I know when the guests came*; see also Experiment 2). In general, at least for *when,* the set of such ambiguously embedding verbs is much larger in Slovenian than in English, so that there was no repetition of verb between the items. The fourth word is the embedded complementizer appearing in one of the two versions outlined above. Words five through nine represent the region of interest as they minimally describe an argument structure that can be modified by *when*. The sixth word is the embedded subject, the seventh is the embedded verb and the eighth and ninth words represent the direct object, where the eighth word was always a demonstrative determiner. This was done in order to make the object structurally "heavier," but not to the point when the complexity of its structure would potentially intervene with determination of the right boundary point. In choosing nouns used for embedded subjects and objects, as well as embedded verbs, we controlled for their plausibility and corpus frequency, for the latter using the FidaPLUS-JOS1M corpus (Erjavec et al., 2010).

The remaining three words always describe a location of the event in the form of a prepositional phrase compatible with the locative specification. The locational content of the prepositional phrase was chosen so that it would have a clear bias toward modification of the event, not of the last phrase (object). It should be also noted that Slovenian is a language where verbal clitics must always appear in the second position in the clause. The test sentences are all in past tense, whose grammatical manifestation in Slovenian requires a particular verbal clitic. That is why the second and fifth words in the test sentences are always verbal clitics, either singular or plural, depending on the grammatical number of the subject.

The target sentences were split into individualized lists balancing all factors in a Latin Square design, so that a different such list is activated for each participant. Each list was combined with 50 filler sentences of various syntactic types and of comparable length. The experimental items and fillers were thoroughly checked by a native speaker of Slovenian who is also a linguist. A complete list of target items along with their English glosses and translations is provided in Supplementary Material.

Subjects performed a self-paced reading task implemented by using the Ibex software (by Alex Drummond, http://spellout.net/ibexfarm/). We used a word-by-word centered-window presentation of stimuli. In this design, a subject initially sees two dashes in the center of the screen. By pressing the space bar, the first word in the sentence appears in place of the dashes. With each subsequent press of the space bar, the current word is replaced with the next word in the sentence, until the end of the sentence is reached. The reason we did not use the currently more popular moving window version of the self-paced reading task (see Just et al., 1982) was to rule out potential topological cues helping one to identify the left and right boundaries of a filler-gap dependency based on the positions of the relevant words or their placeholders (viz. dashes) in the linear representation of the sentence. Ruling out this possibility reinforces the scenario whereby processing filler-gap dependencies is based on the resources of working memory only, which is of primary interest from the point of view of evaluating storage costs. The order of stimulus presentation was pseudo-randomized for each participant by the experimental software and it was ensured that at least one filler intervenes between any two target items.

At the beginning of the experiment, participants were instructed to read the sentences at a natural pace and to be sure they understand what they read. To ensure that participants paid attention to the content of the reading task, half of the target items and one third of the fillers were followed by a yes-no comprehension question. Subjects were instructed to answer the question as quickly and accurately as possible. Feedback was provided when an incorrect response to a comprehension question was given, and subjects were told to take it into account as an indication to read more carefully. No feedback was given in cases of correct answer. Failure to respond within 4 s counted as an incorrect response. Before the start of the experiment, subjects read a short list of practice sentences and comprehension questions in order to familiarize themselves with the task. Each session lasted between 20 and 25 min per participant.

Statistical Procedures

We used the same statistical procedures for all experiments in this study. To control for differences in word length across conditions as well as overall differences in participants' reading speed, a regression equation predicting reading time from word length was constructed for each participant, on the basis of all filler and experimental items (see Ferreira and Clifton, 1986). At each word position, the reading time predicted by the participant's regression equation was subtracted from the actual measured reading time to obtain a *residual reading time*. The resulting residual reading times are the dependent variable used in all analyses (**Tables 2–4** also include raw reading times, to provide a more interpretable scale for the effects).

For all analyses of reading time data, we used linear mixed-effects models (Baayen et al., 2008), and for question-answering data we used a logistic mixed effects model for binary data (Jaeger, 2008). The only fixed effect in our analyses was COMP(lementizer), taking values corresponding to the respective [+interrogative] or [−interrogative] complementizers, with subjects and items entered as random

TABLE 2 | Mean (standard error) comprehension question performance in percent correct as a function of condition, by subject.

da	*kdaj*
82 (1.9)	85 (1.7)

effects. Our constructed models utilized the maximal random effect structure with random intercepts for subjects and items and random slopes for the fixed effect term in subjects and items (Barr et al., 2013). We report *p*-values based on the likelihood-ratio test, whereby a model containing the fixed effect of interest is compared to a model that is identical in all respects except the fixed effect in question. The *p*-values are computed by treating the *t* statistic resulting from linear mixed effects analysis as approximately normally distributed (justified for datasets of our size; see Baayen et al., 2008), as also supported by visual inspection of residual plots which did not reveal any obvious deviations from homoscedasticity or normality. Analyses were performed using the lme4 package (Bates et al., 2014) in R (R Development Core Team, 2011).

Results

Data from five participants were omitted from all analyses because of overall poor comprehension question performance (<67% accuracy overall). No subjects were removed on the basis of slow overall reading time (>4 standard deviations from the mean across subjects). Consequently, data from 69 subjects were used in subsequent analyses. For these subjects, reading time data from items with incorrectly answered comprehension questions were excluded from the analysis. In addition, residual reading time data points that were greater than three standard deviations from the subject mean were also excluded. This affected around 1.0% of the data overall for this experiment.

Comprehension questions

Overall, comprehension questions following the experimental items were answered correctly in 84% of the trials. The percentages of correct answers for each condition are presented in **Table 2**. A paired *t*-test revealed no significant effects [$t_{(357)} = -0.37$, $p = 0.71$]. To control for item (and subject) variability, we also fit a logistic mixed effects model and obtained similar results of COMP not being a significant predictor for the question response accuracy [$\chi^2_{(1)} = 0.07, p = 0.7935$].

Reading times

For the primary analyses, we treated each of the 12 words within each item as its own region, according to the following schema:

(12)

Kritik	je	potrdil	da/kdaj	je	umetnik	izdelal	tisti	koš	v	svoji	delavnici
critic	is	confirmed	that/when	is	artist	created	this	basket	in	his	workshop
MSubj	Cl1	MV	COMP	**Cl2**	**Subj**	**V**	**Det**	**Obj**	FU1	FU2	FU3

Figure 1 and **Table 3** show average residual reading times for each of the 12 primary regions per condition.

There was no significant effect of complementizer type in the COMP region [$\chi^2_{(1)} = 0.71$, $p = 0.3987$]. Since COMP is selected and/or subcategorized by the matrix verb, the absence of variation suggests that the parser is equally likely to expect a [−interrogative] and [+interrogative] complementizer after the selecting verbs. This is in line with the special properties of the verbs we used in our materials: they support both types of subcategorization. This result persists across each of the four verbs used in the stimuli validating the design in terms of balancing different types of subcategorization for the same verb.

As **Figure 1** and **Table 3** illustrate, the interrogative *kdaj* sentences were read slower than the declarative *da* sentences in the post-COMP area until region 9 (Obj). We have defined two aggregation regions in accord with the two different kinds of theoretical predictions considered above. Recall that the first class of theories predicts that the storage costs for wh-adjuncts are distributed more or less in conformity with those for wh-arguments, that is, they are bound to the verb. Thus the first region, indicated by the smaller circle on **Figure 1**, spans the range of stimuli between Cl2, the first post-COMP element, and V. The second type of theory predicts that the endpoint of storage costs extends beyond the verb, namely, across the VP domain generally. Thus the second aggregation domain, indicated by the larger circle on **Figure 1**, includes the first and extends further to the direct object phrase, until the first follow up word FU1.

In the first aggregated region spanning the area from Cl2 until V, linear mixed models revealed a main effect of COMP, which however shows up with a marginal significance [$\chi^2_{(1)} = 3.365$, $p = 0.0666$]. In the second, larger, aggregated region, there is a significant main effect of COMP [$\chi^2_{(1)} = 5.3896$, $p = 0.0203$], with the relevant portions of *kdaj* clauses being read about 10 ms/word ± 3.9 ms/word (standard errors) slower than *da* clauses. We also asked whether there is a main effect of COMP specifically in the direct object area (Det + Obj) differentiating our two aggregated regions. This turned out to be the case [$\chi^2_{(1)} = 4.427$, $p = 0.0353$], indicating that the slowdown in *kdaj* clauses persists across this particular area. Finally, regions FU1-FU3 following direct object showed no main effect of COMP, suggesting that there is no significant difference in reading times between the two conditions [$\chi^2_{(1)} = 0.08$, $p = 0.7795$].

Discussion

There were three main results of this experiment. The first result is that storage costs obtain for the wh-adjunct *kdaj* "when," similarly to wh-arguments. This result holds under the assumption that the resource-consuming memory processes relevant for sentence processing involve both storage and

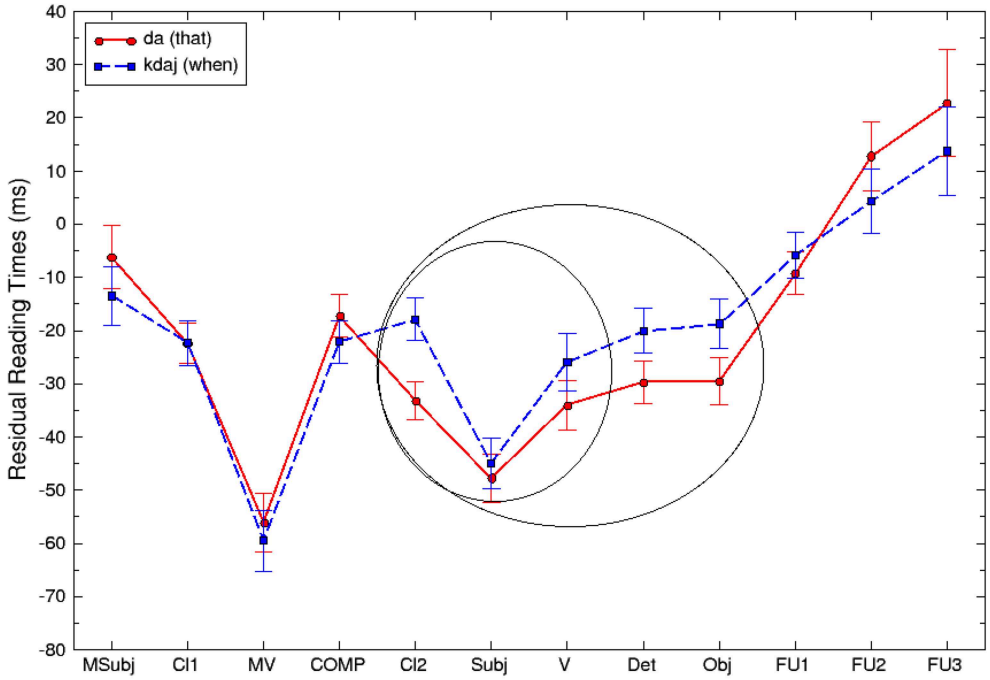


FIGURE 1 | Plot of mean (standard error) residual RTs per word by region in Experiment 1.

TABLE 3 | Mean residual RTs as ms/word by participants as a function of condition, for the post-COMP regions in Experiment 1, rounded to units (raw RTs in parentheses).

Region	*Da*	*Kdaj*
Cl2	−33 (421)	−18 (439)
Subj	−47 (443)	−45 (456)
V	−34 (459)	−26 (474)
Det	−30 (446)	−20 (459)
Obj	−29 (455)	−19 (473)
FU1	−9 (463)	−6 (469)
FU2	13 (502)	4 (499)
FU3	23 (520)	14 (504)

integrations into the partially processed structure, shared in some form by most theories of storage costs up to date. Since our compared conditions only differ in the COMP value, they involve the same number of integrations. Therefore, the increased reading times in the *kdaj*-condition is likely to be attributed to a storage effect. Furthermore, the temporal span of this effect suggests that it is related specifically to processing of *kdaj* which requires additional memory resources reflected in the reading slowdown.

The second result concerns the observed time-course pattern for the storage cost effect with respect to the predictions of the two classes of theories. If the first class of theories (the filler is associated with the verb) is correct, we should expect a significant difference across the first aggregated region, but not across the second. If the second class of theories (the gap is grammatically defined as following the lowest VP constituent) is correct, then we expect the storage cost effect across the second aggregated region, including the first region as well as the area differentiating the two regions. The results indicate that the latter is the case. We have seen that the difference in reading times persists until the end of the direct object area. Under the direct (feature-) association theories predicting association of the wh-adjunct with the verb, the continuing storage effect after the verb region would remain unexplained. At the same time the observed time-course pattern of the storage cost effect is consistent with the phrase structure theories predicting a gap, or the endpoint of the storage effect, after the direct object.

An alternative interpretation of this result, suggested by a reviewer, might be that the slowdown effect over the direct object area Det+Obj is due to (spill-over) integration costs, rather than storage costs *per se*. This interpretation would then be consistent with the theories which directly associate the gap with the verb, similarly to wh-arguments, and it would confine storage costs to the pre-V region only. While this possibility cannot a priori be ruled out given the design of Experiment 1, we believe it is unlikely to be the case. Such a scenario would imply that integration of a wh-adjunct filler is just too costly: it persists through a sequence of three items (V+Det+Obj) which takes considerable time (ca. 1.5 s; see **Table 3**). It is true that wh-adjunct fillers are semantically more complex than wh-arguments (see Section Base/Integration Points of Wh-adjuncts). However, the alleged difficulty appears incommensurable with the relatively simple semantics of *when* as well as with the general pattern of processing filler-gap dependencies generally. In particular, no spill-over effects have been reported in the previous studies

of filler-gap dependencies with wh-arguments (e.g., the ERP study of (Phillips et al., 2005) mentioned in Section Introduction observed a sustained P600 effect ending at the verb, interpreted by these authors as temporary storage cost). In addition, the fact that the post-V slowdown is restricted exactly to the entire direct object area (and not to some point before or after it) would be a suspicious coincidence under the spill-over scenario, whereas it is expected as a storage cost that conforms to the phrase structure-based theories (see above). Given these considerations, we continue to treat the direct object area as part of the relevant storage region.

The third result of Experiment 1, which stems from the second, suggests that the endpoint of the storage costs for a wh-adjunct can be a predictor of its potential gap position, or integration point. This result is important again in light of the grammatical theories that, due to excessive mobility of wh-adjuncts in the syntactic structure, often do not provide reliable diagnostics for their base position. The processing pattern is revealing in cases when the grammatical theory predicts more than one potential base position (see above), as well as in cases when it predicts a gap in a position different from the one found in a processing paradigm. An online processing study thus provides one with an efficient tool to carefully probe for the gap position in such non-trivial examples.

The results of this experiment are suggestive, but cannot be fully generalized because they are based on the processing of a single wh-item. It may be argued that the observed increased storage costs arise because of some specific lexical property of *kdaj* or, alternatively, because of some effect of interaction between this item and the syntactic structure independent of the storage cost effect. This experiment also raised an issue about the specific pattern of filler storage over the clausal subject regions. In particular, we would like to know whether the drop in the reading times is an idiosyncratic effect that occurs with specific wh-adjuncts, or representative of a more systematic pattern. Slovenian is a language whose grammar allows null/unexpressed subjects, so one could imagine a scenario where the storage cost effect expected over the subject would actually already be encoded over the preceding clitic (Cl2) region, given that this is a verbal clitic morphologically specified with the morphosemantic features of the subject (person, number, gender). Consequently, for instance, under the distance-based theory of storage costs reading the actual subject would not count toward calculating the overall storage costs for the wh-adjunct. Finally, not all wh-adjuncts are created equal. Unlike wh-arguments that are usually NPs with predicted syntactic behavior dictated by thematic considerations, wh-adjuncts may differ dramatically from each other from a syntactic point of view. There is thus an important question as to whether other wh-adjuncts elicit a similar kind of a storage cost effect, possibly correlating with their lexical, syntactic and semantic properties. Experiment 2 addresses these issues.

Experiment 2: English *how quickly* and *why*

Experiment 1 was concerned with the wh-adjunct *when* in Slovenian, which falls under the category of VP-modifying adverbs attached relatively low in the syntactic tree (see **Table 1**). Experiment 2 uses English materials in a self-paced reading task and takes the investigation of storage cost effects in wh-adjunct dependencies further. The goal of this experiment was threefold. First, we wanted to replicate the Slovenian pattern of storage costs in a language in which storage costs for wh-arguments have been previously investigated in reasonable detail and at present are better understood (see Section The Lexically-based vs. Syntactically-based Views on Storage Costs), with the aim to strengthen the cross-linguistic dimension of our inquiry. English is a natural choice in this regard. Second, now that there are reasons to believe that wh-adjuncts elicit storage costs as much as wh-arguments do, the main question we ask is whether these storage costs correlate with the syntactic base position of a particular wh-adjunct, along the lines outlined in Section Endpoint of the Storage Costs. Thus in Experiment 2 we focused on the comparison between the VP modifier *how quickly* and the wh-adjunct *why*. As shown in the syntactic literature, the syntactic behavior of *why* is quite different from that of VP-modifying adjuncts, and the most robust grammatical evidence for that again comes from English (see Section Storage Cost Predictions for *why*). Since we wanted to compare the patterns of storage costs for these two modifiers, our corresponding processing predictions can therefore be better grounded in this language.

Yet another goal of Experiment 2 was to more closely investigate the integration point of the wh-adjunct in light of the relevant storage costs. Experiment 1 showed that the parser may have to wait until the direct object is parsed in order to integrate the wh-adjunct. In this respect, we were interested in the role of the Active Filler Strategy as the parser's tendency to fill the adjunct gap as soon as possible. In particular, in cases of complex direct object phrases such as *a glass of water,* does the parser wait for bottom-up evidence that the end of the direct object constituent has been reached in order to discharge the wh-adjunct, or does it do it as soon as this becomes grammatically permissible - in our example, upon encountering *a glass* (and not waiting till the end of the direct object to determine whether the phrase is complete)? This question gains particular importance in light of the proposals in the literature that derive the Active Filler Strategy from a requirement to saturate a thematic role of the wh-filler as soon as possible (Pritchett, 1992; Gibson et al., 1994; Aoshima et al., 2004). If the Active Filler Strategy is indeed a thematic-oriented strategy, it should not be relevant in the case of wh-adjunct processing. On the other hand, if the Active Filler Strategy is, in principle, independent of the thematic factor (and may or may not interact with it), then it, or some version of it, should apply in the case of wh-adjunct dependencies also, and the gap should be filled on the first grammatically permissible occasion.

Storage Cost Predictions for *how quickly*

Syntactically, *how quickly* is a low, VP-modifying adjunct, and in this property it is similar to *kdaj* used in Experiment 1. Both items also have a comparable semantic status of event modifiers. Processing-wise, *how quickly* may be slightly more complex than

kdaj because it contains an additional word[6]. For the purposes of this experiment, we will, however, treat *how quickly* as a single unit (both words were presented simultaneously to participants). Given the results of Experiment 1 using the VP adjunct *kdaj*, we expect a storage cost effect for *how quickly* across a range of input extending between the filler and the syntactically determined gap site or endpoint, which would comport with the filler's VP-modifying syntactic status. The presence of a storage effect of *how quickly* would provide further evidence strengthening the empirical validity of storage costs incurred by VP-modifying adjuncts, both on a cross-item as well as on a cross-linguistic basis.

Storage Cost Predictions for *why*

Our particular interest in *why* in the present study is dictated by the growing consensus in grammatical research that the syntactic and semantic status of *why* is principally different from that of VP-modifying adjuncts like *how quickly* and *when*. Specifically, *why* has different scopal properties, different restrictions on co-occurrence with other wh-phrases, different behavior under syntactic ellipsis, sentential negation and other root phenomena (e.g., Subject-Aux inversion), compared to the other wh-adjuncts. *Why* also has semantic properties that make it different from the other *wh*-adjuncts. Whereas the latter are either event or predicate modifiers, *why* is a functor over an entire proposition (thus a question *Why did John leave the room?* has some sort of a proposition as the answer, e.g., *Because he was hurrying,* rather than a predicate modifier such as *quickly*). As an explanation for this differing behavior, it has been proposed in the syntactic literature that *why* is base-generated in its surface syntactic position in COMP at the left periphery of the sentence, or in a position very close to COMP (Bromberger, 1992; Rizzi, 2001; Stepanov and Tsai, 2008; Shlonsky and Soare, 2011). This amounts to the claim that *why* does not instantiate a filler-gap dependency in the usual sense of a long-distance dependency requiring encoding, storage and subsequent retrieval of the wh-filler. Under the standard compositional semantics approach, if *why* is a functor of propositions, *why* would then be interpreted as a sister of a syntactic node denoting a proposition, which is consistent with its base-generation at the clausal left periphery.

These syntactic and semantic accounts make a very clear prediction for a psycholinguistic study: if *why* does not initiate a filler-gap dependency, then there should be no storage effect in the case of *why*, as opposed to *how quickly*. All else equal, processing-wise, *why* is predicted to behave similarly to the complementizer *that* in a pair of sentences like (13): both expect a proposition afterwards.

(13) a. Peter knows that John fixed the car
b. Peter knows why John fixed the car

We thus expect that *how quickly* and *why* will show a contrast in terms of expected storage costs. While *how quickly* should elicit storage costs similarly to Slovenian *kdaj*, *why* is not expected to elicit any additional storage costs compared to the *that* control. Experiment 2 tests this prediction for English.

Methods

Participants

Eighty seven adult volunteers from the Glasgow community in the UK participated in this experiment voluntarily for no material compensation. All participants were recruited via email and social networking forums. All declared themselves as monolingual native speakers of English and were naïve to the purposes of the study.

Materials

Twenty-four sets of sentences with embedded clauses were carefully constructed[7]. Similarly to Experiment 1, the sentences in each set were exactly identical except for the value of the embedded COMP(lementizer). This time COMP takes one of the three possible values, each defining the respective condition: (1) *that*; (2) *why;* and (3) *how_quickly*. To control for the length of the wh-dependency, all sentences were made exactly 15 words long and matched by syllable structure to the best extent possible. An example is given in (14)[8]:

(14) The reporter didn't know *that/why/how_quickly* the soldier shot the panel of doctors in the hospital

Similarly to Experiment 1, each sentence begins with a four-word main clause including a two-word subject, a verbal modifier (e.g., negative *didn't* or an adverb) and a main verb in past tense. The main verbs were chosen so that they may embed either a wh-interrogative clause, or a declarative *that* clause. We chose four such ambiguously subcategorizing verbs: *know, forget*, *explain,* and *find_out*, which were equally represented among the set of experimental items (six instances each). The fifth word is the embedded Complementizer appearing in one of the three versions outlined above. Words six through twelve represent the main area of interest as they correspond to the verbal argument structure. The sixth and seventh words are always an embedded subject of the definite description type [the N], and the eighth is the embedded verb.

Words nine through twelve represent the direct object, which was always of the form [the N of N]. The first N is always

[6] As pointed out at the beginning of Section Experiment 1: Slovenian kdaj ("when"), using *when* in English can potentially be confounded by its lexically ambiguous status. Other simplex wh-adjuncts in English like *where* and *how* might be subject to similar concerns if used in embedded clauses, as in our study.

[7] Preparation of the English materials and subject recruitment for Experiment 2 was implemented by Calum Riach (see Riach, 2014, supervised by the first author).

[8] A reviewer raises a concern about potentially greater pragmatic oddity/implausibility of the *how quickly* sentences as compared to the *why* counterparts, which could then lead to an increase of reading times for the former in the post-verbal (direct object) region. Even though we controlled for general plausibility upon constructing the items, we conducted an additional norming-like evaluation with the aim to see if our items are biased in this direction. We asked two native speakers of English who were not involved in creating the items, to indicate, for each item, which version out of the two sounded more natural (plausible) to them, compared to the other. Both speakers preferred the *why* reading slightly more often (by 4 and 5 items out of 24, respectively) than the *how quickly* reading. We then fit mixed effects models entering preference as a fixed factor along with COMP, asking whether preference affects the reading times, in the direct object area. We found no main effect of preference ($p = 0.2246$ and $p = 0.6839$, respectively) and no interaction with the factor COMP. This suggests that the post-verbal reading times are not affected by pragmatics/plausibility, at least in obvious ways.

lexically ambiguous between a standard noun and a classifier, e.g., *glass*. The second N is either a mass noun (*water*) or bare plural (*doctors*). The reason why the direct object was intentionally made structurally more complex has to do with investigating the Active Filler Strategy. If this, or similar strategy requiring gap filling as soon as possible is in place also for wh-adjuncts, we would expect the end of storage costs to occur after the first N, given a possible gap site at this location in the context of the partially processed input. Conversely, if the Active Filler strategy is not operative in the case of wh-adjuncts, then, under the assumption that phrase structure (still) guides the integration point, the gap would be expected at or after the second N. In other words, in a sentence like *I don't know how quickly John finished the drink of beer* integration of *how quickly* may occur either after *drink,* or after *beer*. The remaining three words in the sentence (words 13–15) always describe a location of the event in the form of a prepositional phrase compatible with the specification expressed by *where*.

The target sentences were split into individualized lists balancing all factors in a Latin Square design, so that a different such list is activated for each participant by the experimental software. Each such list was combined with 50 filler sentences of various syntactic types and of comparable length. The order of stimulus presentation was also pseudo-randomized separately for each participant and it was ensured that the presentation begins with a filler and that at least one filler intervenes between any two target items. A complete list of target items is provided in Supplementary Material.

The procedure was identical to the procedure in Experiment 1, except that half of the filler sentences were accompanied with a yes-no comprehension question. Computation of residual reading times and the statistical analysis procedures all followed those used in Experiment 1. In addition, Tukey's pairwise comparisons were performed on our fitted models using the *glht()* function in R's "multcomp" package (e.g., Hothorn et al., 2008).

Results

Ten subjects were excluded because of coding errors that led to distorting stimulus presentation in several trials. In addition, six subjects were excluded due to low comprehension question accuracy (<67%) and/or low overall reading time (>4 standard deviations from the mean across subjects). This left the data from 71 subjects to be used in the analyses. Overall, comprehension questions were answered correctly in 88% of the trials. Residual RT data points (pooled across all regions and conditions) that were greater than three standard deviations from the mean were excluded from all analyses, affecting around 1.4% of the data overall for this experiment.

For the primary analyses, we treated each of the words as its own region (omitting the main clause area). **Figure 2** and **Table 4** show average residual reading times for each of the 11 primary regions per condition.

We have then defined four aggregated regions of interest, as shown in (15):

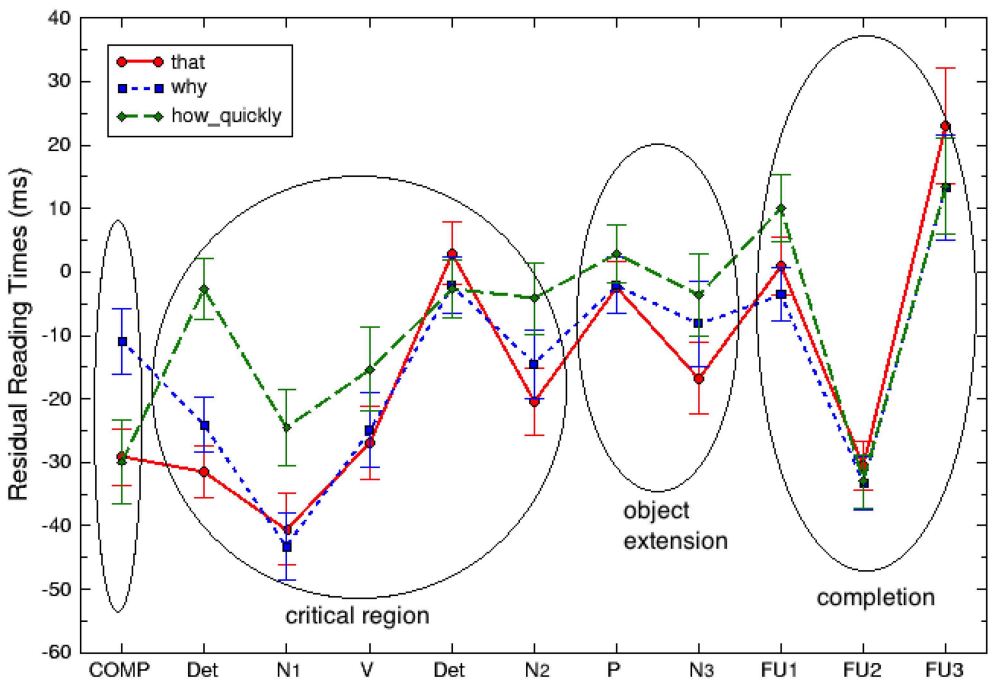


FIGURE 2 | Plot of mean (standard error) residual RTs per word by region in Experiment 2.

(15)

that / how_quickly / why	the	soldier	shot	the	panel	of	doctors	in	the	hospital
COMP	**Det**	**N1**	**V**	**Det**	**N2**	P	N3	FU1	FU2	FU3
	CRITICAL REGION					OBJECT EXTENSION		COMPLETION		

The first region includes embedded COMP which takes one of the three condition-defining values, *that, why* or *how quickly*. Following that is the critical region which represents components of the argument structure of the embedded verb. The object extension is always an *of*-phrase (see above). The final region is a locative PP including three follow-up words. **Table 5** includes estimated mean residual as well as raw reading times, illustrating a per region comparison among the three conditions.

At the leftmost COMP region, there is a significant variation in reading times [$\chi^2_{(2)} = 6.2785$, $p = 0.04331$]. *Post-hoc* Tukey estimations among the pairs of conditions indicate that *why* is read slower than both *that* and *how quickly* (pair *why/that*: $z = 2.157$, $p = 0.0783$; pair *why/how_quickly*: $z = -2.889$, $p = 0.0105$; pair *that/how_quickly*: $z = -1.146$, $p = 0.4844$). The tendency to read *why* slower than *that* and *how quickly* persists across each of the four subgroups of items defined by the respective embedding verbs (*know, forget, explain, find out*) calculated separately, though does not quite reach significance within any of the subgroups ($p > 0.05$). With respect to the *why*/*that* and *how_quickly*/*that* pairs, this result appears to contrast with Experiment 1 where there were no notable differences in the rate of reading *da* and *kdaj*.

Moving on to the critical region, linear mixed models revealed that COMP significantly affects reading times across the range until the first N of the direct object, the first suspected integration site of the wh-adjunct dependency [$\chi^2_{(2)} = 8.5873$, $p = 0.01365$], with the *how quickly* clauses being read about 12 ms/word ± 4.3 ms/word (standard errors) slower than the corresponding *that* clauses, and the *why* clauses virtually not affected at all with a difference of 0.3 ± 4.3 ms (standard errors) from the *that* clauses. Post-hoc pairwise Tukey comparisons confirm that the critical region of *how quickly* clauses was read significantly slower compared to *that* clauses ($z = 2.830$, $p = 0.01284$). *How quickly* clauses were also read slower than *why* clauses ($z = 3.335$, $p = 0.00241$). Finally, *why* clauses in the critical region were read with a rate similar to that of *that* clauses ($z = -0.459$, $p = 0.89048$).

In the context of estimating an endpoint of the storage cost effect and a possible impact of the Active Filler Strategy, we also asked whether the slowdown in the reading times for the *how quickly* clauses persists specifically over the part of the direct object area (Det+N2), similarly to Experiment 1. We found that, in that sub-region, *how quickly* clauses are read about 8 ms/word slower than the *that* clauses and about 5 ms/word slower than the *why* clauses. However, the effect does not reach significance [$\chi^2_{(2)} = 2.60$, $p = 0.27$].

As **Figure 2** demonstrates, *how quickly* clauses also tend to be read slower in the object extension region (the *of*-phrase), all the way up to the first follow-up word FU1. However, no main effect of COMP was estimated at the object extension region overall [$\chi^2_{(2)} = 2.1047$, $p = 0.3491$], or at each of the two word regions comprising it [region P: $\chi^2_{(2)} = 2.6271$, $p = 0.2687$; region N_3: $\chi^2_{(2)} = 2.2795$, $p = 0.3199$]. Finally, at the completion region (locative PP) no significant difference in reading times across the three conditions is observed either [$\chi^2_{(2)} = 0.6078$, $p = 0.7379$].

TABLE 4 | Mean residual RTs as ms/word by participants as a function of condition in Experiment 2, rounded to units (raw RTs in parentheses).

Region	*that*	*how quickly*	*why*
COMP	−29 (380)	−30 (429)	−11 (391)
Det	−32 (372)	−3 (400)	−24 (379)
N_1	−40 (396)	−24 (412)	−43 (394)
V	−27 (412)	−15 (420)	−25 (409)
Det	3 (405)	−3 (400)	−2 (400)
N_2	−20 (402)	−4 (410)	−15 (407)
P	−2 (394)	3 (399)	−2 (394)
N_3	−17 (404)	−4 (417)	−8 (413)
FU1	1 (397)	10 (411)	−4 (393)
FU2	−30 (372)	−33 (370)	−33 (369)
FU3	23 (450)	14 (439)	13 (441)

TABLE 5 | Mean residual RTs as ms/word by participants as a function of condition, for the four aggregated regions in Experiment 2, rounded to units (raw RTs in parentheses).

Aggregated region	*that*	*how quickly*	*why*
COMP	−29 (380)	−30 (429)	−11 (397)
Critical region	−23 (397)	−9 (408)	−21 (419)
Object extension	−7 (399)	−1 (408)	−5 (403)
Completion	−2 (406)	−6 (406)	−3 (401)

Discussion

There were three main results of this experiment. The first notable result was the replication of the pattern of reading times observed in Experiment 1. In particular, *how quickly* clauses were read slower than *that* clauses in the critical region. Since, as in Experiment 1, the number of structural integrations is the same, the slowdown is likely to be due to a storage effect.

Furthermore, since both *how quickly* and Slovenian *kdaj* ("when") share key syntactic characteristics typical for VP-modifying adjuncts, it can be concluded that such adjuncts elicit a storage effect similar to the one reported previously for wh-arguments, namely subjects and objects, and, furthermore, that this effect may be language-independent.

The second result of Experiment 2 was divergence of the patterns of the storage costs for *why* and *how quickly*. These diverging patterns would seem puzzling at face value, but they receive a natural explanation if grammatical considerations are taken into account. Since, *how quickly* needs to be kept in memory long enough to reach its integration point in the VP domain, storing it incurs a tax, much along the lines of the previous research on temporary storage of wh-arguments. At the same time, *why* does not need to be kept in memory (or it does for a very short time) because its integration site is more or less at the point where it is encountered. In this respect *why* behaves like a declarative complementizer. This result suggests that grammatical rules concerning base-generation of wh-adjuncts may serve as a reliable predictor of their storage costs, and, conversely, that observed storage cost effects provide processing evidence for the grammatical statements regarding the base position of specific (wh-)adjuncts in the syntactic structure of the sentence.

The third result, related to the second, concerns the endpoint of the storage costs for *how quickly*. Recall that inclusion of the *of*-phrase into the direct object region was motivated by our interest in the role of the Active Filler Strategy in wh-adjunct dependencies. In particular, if this or similar strategy is active, the endpoint should be observed at or around the right boundary of the critical region. If it is not active, then, under the phrase structural restrictions, we would expect the storage effect also over the object extension, the integration point then being at or right after that region.

We observed no significant difference in reading times in the direct object region, even though there is a tendency to read *how quickly* sentences slower than both *that* and *why* sentences, in that region. Thus, we did not fully replicate the result of Experiment 1 which revealed a reliable difference in the reading times between *kdaj* and *da* sentences in the direct object area. Based on the results of Experiment 1, we concluded that the parser consults the relevant phrase structural information while attempting to integrate the wh-adjunct. Given that, the reason why the English participants did not show a difference in the reading times in the direct object area following the verb could be because the grammatically permissible integration point of the adjunct *how quickly* is not the same as that for the wh-adjunct *kdaj* ("when") namely, following the direct object. As we saw in Section Base/Integration Points of Wh-adjuncts, the grammatically licensed base position of adjuncts may be either pre-verbal or postverbal. It might be, then, that the base position of *how quickly* is actually preverbal [cf. example (ib) in fn. 2], and if so, the parser would not have to wait until the direct object in order to integrate this wh-adjunct.

The materials in Experiment 2 contained complex object noun phrases such as *the panel of doctors*. We wanted to see if the storage cost effect ends after the first, or second noun, in order to determine whether the Active Filler Strategy is operative in the case of wh-adjunct dependencies. The results of Experiment 2, namely, the absence of a reliable effect both at the first noun (in the Det+N2 region) as well as across the entire direct object area, do not permit us at this point to make a definitive conclusion in one or the other direction. Thus the possibility that the Active Filler strategy applies also in the case of wh-adjunct dependencies, cannot be ruled out.

A somewhat surprising accompanying result of Experiment 2 was a slowdown in reading the embedded *why* item itself, compared to reading times for embedded *that* and *how quickly*. The relevance of this result lies in the domain of processing subcategorized information, in particular, subcategorized complementizers. With respect to the *why/that* pair, this contrasts with Experiment 1 where there were no notable differences in the rate of reading *kdaj* vs. *da* in Slovenian. It is not clear whether verbal subcategorization for a specific question word should lead to an increased processing effort reflected in reading times, or the observed difference in reading time is simply a baseline effect. Grammatically, the verbs selected for this experiment are equally likely to select for any wh-item, or for a declarative complementizer. Previous studies on filler-gap dependencies in embedded interrogatives (notably fewer than those that investigate filler-gap dependencies in relative clauses with an invariant relativizer such as *which*) did not report any difference in reading times at the embedded COMP, e.g., between wh-arguments vs. complementizer *if* (Stowe, 1986; Lee, 2004). A number of processing factors may in principle modulate expectations for a particular subcategorization frame. For instance, studies of garden path effects suggest that verb subcategorization frequencies have an immediate effect on sentence processing (Trueswell et al., 1993; Garnsey et al., 1997; Hare et al., 2003; Snedeker and Trueswell, 2004, see also Mitchell, 1987). One may also estimate predictability of a specific verbal subcategorization by calculating its conditional probability in the context of a subcategorizer, based on corpus data (cf. Levy, 2008). **Table 6** lists predictability of each of the three COMP items for each of the four embedding verbs. As **Table 6** indicates, for every verb with the exception of the *explain-why* bigram the predictability drops along the continuum *that-how-why* (we take the predictability of *how* to be representative for estimating the reading time for *how_quickly*). If predictability (negatively) correlates with reading times logarithmically (Hale, 2001; Levy, 2008), then it is in principle possible that the reading times increase past some critical threshold in predictability, thus making *why* read slower. It is also possible that *why* is read slower because it is different from other wh-adjuncts, as well as from the complementizer *that*: as pointed out above, it is a functor over propositions, as opposed to VP adjuncts that are

TABLE 6 | Co-occurrence of the target verbs with respective COMPs, calculated as conditional probability P(COMP |VERB) = P (COMP ∩ VERB)/P(VERB), where P (COMP ∩ VERB) is a probability of the respective bigram, based on the British National Corpus (Mark Davis/Brigham Young University, http://corpus.byu.edu/bnc/).

Context/COMP	*that*	*how*	*why*
know	0.083	0.033	0.011
forget	0.097	0.017	0.002
explain	0.082	0.047	0.078
find out	0.042	0.068	0.023
MEAN	0.076	0.04125	0.0285

predicate modifiers, and to *that* which is just a clause-introducer. Or, again, this could be just a baseline artifact. To further clarify this issue, the follow up Experiment 3 was conducted.

Experiment 3: *why* vs. *that*

Experiment 3 had the same design as Experiment 2. This time we concentrated only on the subcategorization aspects of COMP, asking whether reading the subcategorized *why* takes additional processing effort compared to reading the embedded *that*.

Methods

Participants

The procedure of subject recruitment was similar to Experiment 2. 26 English-speaking monolingual subjects volunteered to participate in this study for no material compensation.

Materials and Procedure

Experiment 3 used a subset of the English materials used in Experiment 2. We used the same 24 target items but this time COMP only had values *that* and *why*. The rationale for choosing these items was to control for the (absence of) possible filler-gap effects at COMP, given that neither of these items instantiate a filler-gap dependency proper, as Experiment 2 has demonstrated.

Subjects saw 24 items in a pseudo-randomized order, interspersed with 52 fillers. Similarly to Experiment 2, half of the filler items were accompanied by a comprehension question.

Results and Discussion

Overall, comprehension questions were answered correctly in 87% of the trials. No subject was excluded on the basis of comprehension accuracy or slow overall reading times (>4 standard deviations from the mean across subjects). Overall, comprehension questions were answered correctly in 87% of the trials. Residual reading time data points that were greater than three standard deviations from the mean were excluded from all analyses, affecting around 0.8% of the data for this experiment.

There was no main effect of COMP at the embedded complementizer [$\chi^2_{(1)} = 0.768$, $p = 0.3808$]. This suggests that subcategorization does not affect the reading times of complementizers *that* and *why*. Although with only 26 participants this experiment had less statistical power than Experiment 2, this result largely corroborated that of Experiment 1. However, there is still a tendency to read *why* slower than *that*, by about 5–20 ms depending on the matrix verb [mean overall RRT (*that*) = −11 ms; mean overall RRT (*why*) = −1 ms]. Thus, if a predictability effect of the kind outlined above exists, it is very weak and requires a substantially larger statistical sample than the population size in this study to reliably reveal itself.

General Discussion

In the beginning of this article, we viewed wh-adjunct interrogatives as an important and previously under-investigated empirical ground for testing theoretical predictions pertaining to the following aspects of storage costs in filler-gap dependencies: (1) the thematic factor and the role of lexically-based strategies of computation of online storage costs; and (2) the processing and grammatical predictions concerning the endpoint of the storage costs, also in the context of the Active Filler strategy. Below we evaluate the main results of this study in light of these aspects, and point to some further issues.

The Thematic Factor and the Lexically-based Strategies Revisited

Both Experiment 1 and Experiment 2 showed a reliable storage effect related to wh-adjuncts modifying a verbal phrase (VP), that is, Slovenian *kdaj* "when" and English *how quickly*. This effect is not predicted by the class of the theories that calculate temporary storage costs in terms of the number of unassigned/incomplete thematic roles (Hakuta, 1981; Gibson, 1991), as well as in terms of the number of unassigned/incomplete Case features (Stabler, 1994). The reason is that, being non-referential syntactic entities, wh-adjuncts do not receive a thematic role from the verb, and they generally do not need Case from the verb, their Case feature being satisfied either adjunct-internally (as in the case of wh-adjunct PPs such as *on which table*), or absent at all, as in the present study. On the other hand, our results support the class of storage cost theories that do not make reference to the thematic, Case or referential status of the filler. These include theories that estimate storage costs in terms of temporarily stored incomplete phrase structure rules or their close counterpart such as the SLASH feature of HPSG (see Section The Lexically-based vs. Syntactically-based Views on Storage Costs), as well as in terms of the number of incomplete syntactic heads (Gibson, 1998, 2000). These latter theories can thus be extended to wh-argument as well as wh-adjunct dependencies.

Note that the relevant principal distinction between these two classes of theories of storage cost metric lies in the amount of theoretical weight they place on a (lexicon-oriented) internal featural specification of the filler as opposed to its (syntax-oriented) structural environment. The Case/thematic role metrics of storage costs capitalize on the thematic argument and/or the NP status of the filler. Even though theta-roles, as well as Case, have always been commonly understood as part of the syntactic computation in the grammar, it was also clear that they have a strong lexico-semantic component. In contrast, the incomplete phrase structure and incomplete syntactic head metrics of storage costs capitalize on the syntactic status of the filler, that is, its structural relation with respect to other syntactic constituents specified at the level of syntax, as in the former case, or syntactic-head driven expectations, as in the latter. Our results thus support a more syntax-oriented and less lexicon-oriented view of temporary storage costs in filler-gap dependencies[9].

This view harmonizes with the grammatical status of (wh-)adjuncts. Since wh-adjuncts, unlike wh-arguments, are not grammatically associated with the verb directly, the integration point of a wh-adjunct in a filler-gap dependency, or its gap

[9]It should be noted that in phrase structure theories such as HPSG (see Section The lexically-based vs. Syntactically-based Views on Storage Costs) the distinction between the lexical and syntactic modules is not as clear cut as in other phrase structure theories (for instance, the transformational generative grammar). But even in that framework, the SLASH feature assigned to a lexical head (e.g., verb) is basically part of the syntactic computation establishing a relation between that head and other syntactic elements in the structure, thus the relevant processing predictions for a filler-gap dependency could arguably be made, again, on a syntactic basis.

site, is not signaled by the relevant stimulus encountered in the input (viz. the verb). Rather, it is determined on the basis of computing an abstract syntactic node [cf. (4)] with which the wh-adjunct can be associated, in the partially processed input. It is thus reasonable to suppose that the temporarily stored information associated with the syntactically constructed host, is itself of a syntactic nature, so that this kind of computation can be performed at the same, syntactic, level. This is consistent with the modular theory of parsing (Fodor, 1978), where storing and integration can potentially be performed during the first, syntactic, pass, as well as with the interactive theories, with a qualification that no access to non-syntactic (e.g., thematic) sources of information would be needed in the case of wh-adjuncts.

The Active Filler Strategy Revisited

The Active Filler Strategy (see Section Endpoint of the Storage Costs) was originally formulated independently from the subcategorization or theta-role assignment properties. A number of later works (e.g., Pritchett, 1992; Gibson et al., 1994; Aoshima et al., 2004) argued that the Active Filler Strategy in filler-gap dependencies reduces to the parser's need to satisfy thematic requirements of the fronted wh-phrase as soon as possible. The results of our study did not rule out the possibility that the Active Filler Strategy is operative also in wh-adjunct dependencies that are not thematically-based. If this possibility ultimately turns out to be true, that line of argument would be questioned. In this regard, we would like to briefly revisit some of the empirical evidence offered in the literature in support of recasting the Active Filler Strategy in thematic terms and consider an alternative, non-thematic interpretation of that evidence.

One empirical argument in favor of reinterpreting the Active Filler Strategy in terms of thematically-based statements comes from Aoshima et al. (2004) and is based on their experimental investigation of the Active Filler effect in Japanese, an SOV language where objects precede verbs. Aoshima et al. (2004) considered sentences with a left-scrambled wh-word that was an object of the verb in the embedded clause, as in (17) [their (7b)] which is interpreted as an embedded wh-question. Note the question word *–ka* marking the scope of that embedded question and appearing as a verbal suffix: this marker is obligatory in that context and is taken to be an interrogative complementizer:

(16) a. Dare-ni John-wa [Mary-ga sono hon-o
whom-dat John-top Mary-nom that book-acc
ageta-ka] itta.
gave-Q said
"John said to whom Mary gave that book."

The authors provide experimental evidence that the Japanese readers associate the scrambled wh-word with the most embedded clause of a multi-clause sentence (given the presence of the matrix subject). They argue that the wh-phrase *dare-ni* is already associated with the (bracketed) embedded clause even before the embedded verb is encountered, on the basis of a Japanese counterpart of the "filled gap" effect (Stowe, 1986, see also Section Endpoint of the Storage Costs). In particular, the readers show a surprise effect if instead of the marker–*ka* they encounter a different marker–*no* in the same context. The authors argue that if the parser's goal were simply to create a gap as soon as possible, then there would be no motivation to interpret the fronted wh-phrase inside the (most) embedded clause. Rather, the parser would posit a gap in the main clause (after the subject *John-wa*), and that gap would then be unaffected by further (embedded) structure. On the other hand, the embedded clause interpretation is expected, if the parser's objective is to satisfy thematic requirements of the verb or of the wh-phrase: the most embedded clause in an SOV language provides the first opportunity to accomplish that. In that case, the authors argue, the parser "repositions" the main clause gap as an embedded clause gap by reanalysis. On these grounds, they conclude that the Active Filler Strategy is a thematically-driven strategy (the authors also argue that the active search initiated by the parser in order to integrate the wh-phrase cannot be driven solely by the requirement to associate with the question marker; see this work for details).

The argument thus builds on the observed parser's tendency to search for the first available verb to associate with the wh-filler (see also Pritchett, 1992; Gibson et al., 1994 for similar arguments). A thematic association is indeed a natural explanation of this tendency, but, we believe, not the only one. Indeed, an association of the argument wh-filler with the verb can also be accomplished by a phrase structure rule such as V → NP V, whereby the verb is a right sister of the relevant phrase. From the perspective of the parser, a lexical strategy such as "this wh-phrase must be a thematic argument of some verb, let's go and find that verb as soon as possible" is equally plausible as a syntactic strategy such as "this wh-phrase must be a structural sister of some verb, let's go and find that verb as soon as possible." In the scenario of incremental structure building considered above, this amounts to storing the relevant phrase structure rule with an open slot (a verb in this case) in the working memory until a suitable candidate for filling in the slot is found, fully consistent with the theories of incomplete phrase structure rules. For the case of wh-arguments associated with verbs, the two strategies are virtually indistinguishable. They have the same empirical consequences, since the grammatical theory tells us that theta roles are assigned in a very local structural configuration, easily expressible with the usual machinery of phrase structure rules (e.g., Haegeman, 1994). Wh-adjuncts, however, provide a useful empirical ground for distinguishing the two strategies. The thematic/lexical strategy is not easy to restate in this case, precisely because thematic considerations are irrelevant here, whereas in the syntactic strategy, all that is needed is just to replace the relevant phrase structure rule (e.g., VP → VP Adj). The syntactic strategy additionally implies that the parser is sensitive to abstract syntactic nodes as well as to lexical items, but this is a common assumption made in the parsing literature which is simply reinforced here. A different version of a syntactically-oriented strategy is de Vincenzi's (1991) re-interpretation of the Active Filler Strategy in terms of his Minimal Chain Principle: "Avoid postulating unnecessary

chain members at S-structure, but do not delay required chain members" (p. 13).

Processing Evidence for the Base Position of Wh-adjuncts

The results from Experiment 1 and Experiment 2 are also relevant for grammatical theories regarding the base location of particular adjuncts. As mentioned in Section Base/Integration Points of Wh-adjuncts, the flexible phrase structural status of syntactic adjuncts makes it often difficult to pinpoint their base position for the purposes of explanatory syntactic analyses. This is in contrast with wh-arguments, whose base positions are usually trivially (modulo linear directionality of arguments as a parameter distinguishing, for instance, SVO from SOV languages) deduced on the basis of the linear positions of the respective predicates. With regard to wh-adjuncts, the endpoint of storage costs may provide a valuable, though admittedly indirect, processing evidence regarding these base positions. For instance, in Experiment 1 the object is a noun phrase. The end of the storage costs appears to be marked at or around the end of that noun phrase. Thus, by adjusting for the incremental character of online sentence processing, one may make an informed guess about the narrow structural area where the gap postulated by the mental grammar must lie, for each particular wh-adjunct under consideration. At the very least, the processing pattern provides us with a reasonable idea regarding directionality of the wh-adjunct gap relative to the verb. Furthermore, the absence of a continuing storage costs pattern for *why* observed in Experiment 2 is compatible with predictions of the grammatical theory regarding the non-postulation of the gap for this item. We thus have a reason to believe that the endpoint of storage costs may provide useful processing evidence for the grammatical theory of wh-adjuncts.

Overall, our results in Experiment 1 and Experiment 2 suggest that the role of the thematic factor in parsing should not be overestimated. While there is good evidence that the parser is generally sensitive to the argument structure of verbs (e.g., Rayner et al., 1983; Clifton et al., 1991; Friederici and Frisch, 2000), as far as filler-gap dependencies are concerned, the argument structure cannot be the (only) type of information that the parser makes use of during the temporary storage of the filler. Storage costs as a measure of complexity in parsing wh-adjunct dependencies suggest that phrase structure must play a role as well. In this respect, our results are consistent with the theories of integration costs employing complexity metrics of integration that do not take into account thematic information, but are based on different units of comparison, such as the number of intervening discourse referents. Our results are also consistent with the recent proposal that information from preverbal NPs may be sufficient to trigger active gap creation without having access to the verbal information including argument structure, in a kind of "hyper-active" manner (Omaki et al., 2015). In other words, the verb may not play an instrumental role in filler-gap dependencies even in the case of wh-arguments after all. In conjunction with our results on storage effects, this raises an interesting question as to whether the thematic factor can be dispensed with altogether in the processing theories of filler-gap dependencies, and replaced with the corresponding phrase structural statements. Note that in the case of wh-arguments, the thematic information is largely mirrored with the phrase structural information (this state of affairs is formalized in grammatical theory in various forms, such as "the Projection Principle," cf. e.g., Chomsky, 1986). Thus a direct object is usually a sister of the transitive verb, and a subject is a sister of the VP. Further relevant tests for probing the role of the thematic factor independently of phrase structure may potentially include thematic and non-thematic uses of *where,* as in *where did John V the book?* in conjunction with verbs like *put* (thematic) and *see* (non-thematic), and similar constructions for other wh-adjuncts.

Concluding Remarks

The present study provided converging cross-linguistic evidence from Slovenian and English, two languages belonging to different language families, that processing filler-gap dependencies with wh-adjuncts as fillers elicit storage costs across the range of the filler-gap dependency ending approximately at the points predicted by the grammatical theories for particular adjuncts. Our findings provide evidence to the class of storage cost models that are based on computation of the number of incomplete phrase structure rules or, alternatively, the number of incomplete syntactic heads. Our results underscore the non-thematic character of storage costs, and at the same time support the principle-based approach to parsing that draws on grammatical knowledge, specifically phrase structure, as a primary source of parsing decisions (Berwick and Weinberg, 1984; Pritchett, 1988, 1991, 1992; Gibson, 1991; Gibson et al., 1994; Weinberg, 1999).

Acknowledgments

This research was supported in part by the Heisenberg Fellowship of the German Research Foundation to the first author (STE-1851/1) and by the Slovenian Research Agency (program No. P6-0382). We are grateful to the two reviewers and Topic Editor Matthew Wagers for helpful comments and suggestions, and to Calum Riach, Petra Mišmaš, Jeremy Yeaton, Beth Phillips and Viviane Déprez for valuable assistance. Portions of this material were presented at AMLaP-2014 at the University of Edinburgh.

Supplementary Material

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpsyg.2015.01301

References

Abney, S. P., and Johnson, M. (1991). Memory requirements and local ambiguities of parsing strategies. *J. Psycholinguist. Res.* 20, 233–250. doi: 10.1007/BF01067217

Aoshima, S., Phillips, C., and Weinberg, A. (2004). Processing filler-gap dependencies in a head-final language. *J. Mem. Lang.* 51, 23–54. doi: 10.1016/j.jml.2004.03.001

Baayen, R. H., Davidson, D. J., and Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *J. Mem. Lang.* 59, 390–412. doi: 10.1016/j.jml.2007.12.005

Baddeley, A. (1990). *Human Memory*. Needham Heights, MA: Allyn and Bacon.

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: keep it maximal. *J. Mem. Lang.* 68, 255–278. doi: 10.1016/j.jml.2012.11.001

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2014). *lme4: Linear Mixed-effects Models using Eigen and S4_. R Package Version 1*. Available online at: http://lme4.r-forge.r-project.org/

Berwick, R., and Weinberg, A. (1984). *The Grammatical Basis of Linguistic Performance*. Cambridge, MA: MIT Press.

Bowers, J. (1993). The syntax of predication. *Linguist. Inq.* 24, 591–656.

Bromberger, S. (1992). "What we don't known when we don't know why," in *On What We Know We Don't Know*, ed S. Bromberger (Chicago, IL: The University of Chicago press), 145–169.

Chen, E., Gibson, E., and Wolf, F. (2005). Online syntactic storage costs in sentence comprehension. *J. Mem. Lang.* 52, 144–169. doi: 10.1016/j.jml.2004.10.001

Chomsky, N., and Miller, G. A. (1963). "Introduction to the formal analysis of natural languages," in *Handbook of Mathematical Psychology*, vol. 2, eds R. D.Luce, R. R. Bush and E. Galanter (New York, NY: Wiley), 269–321.

Chomsky, N. (1986). *Barriers*. Cambridge, MA: MIT Press.

Cinque, G. (1999). *Adverbs and Functional Heads: A Cross-linguistic Perspective*. New York, NY: Oxford University Press.

Clifton, C., Speer, S., and Abney, S. P. (1991). Parsing arguments: phrase structure and argument structure as determinants of initial parsing decisions. *J. Mem. Lang.* 30, 251–271. doi: 10.1016/0749-596X(91)90006-6

Davidson, D. (1980). *Essays on Actions and Events*. Oxford: Clarendon Press.

de Vincenzi, M. (1991). *Syntactic Parsing Strategies in Italian*. Dordrecht: Kluwer Academic Publishers.

Erjavec, T., Fiser, D., Krek, S., and Ledinek, N. (2010). "The JOS linguistically tagged corpus of Slovene," in *Proceedings of 7th International Conference on Language Resources and Evaluations* (Valletta: European Language Resources Association), 1806–1809.

Ernst, T. (2001). *The Syntax of Adjuncts*. Cambridge, MA: Cambridge University Press.

Ferreira, F., and Clifton, C. (1986). The independence of syntactic processing. *J. Mem. Lang.* 25, 348–368. doi: 10.1016/0749-596X(86)90006-9

Fiebach, C., Schlesewsky, M., and Friederici, A. (2002). Separating syntactic memory costs and syntactic integration costs during parsing: the processing of German wh-questions. *J. Mem. Lang.* 47, 250–272. doi: 10.1016/S0749-596X(02)00004-9

Fodor, J. D. (1978). Parsing strategies and constraints on transformations. *Linguist. Inq.* 9, 427–473.

Frazier, L., and Clifton, C. Jr. (1989). Successive cyclicity in the grammar and the parser. *Lang. Cogn. Process.* 4, 93–126. doi: 10.1080/01690968908406359

Friederici, A. D., and Frisch, S. (2000). Verb argument structure processing: the role of verb-specific and argument-specific information. *J. Mem. Lang.* 43, 476–507. doi: 10.1006/jmla.2000.2709

Garnsey, S. M., Pearlmutter, N. J., Myers, E., and Lotocky, M. A. (1997). The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *J. Mem. Lang.* 37, 58–93. doi: 10.1006/jmla.1997.2512

Gazdar, G., Klein, E., Pullum, G., and Sag, I. (1985). *Generalized Phrase Structure Grammar*. Oxford: Blackwell.

Gibson, E., Hickok, G., and Schütze, C. T. (1994). Processing empty categories: a parallel approach. *J. Psycholinguist. Res.* 23, 381–405. doi: 10.1007/BF02143946

Gibson, E. (1991). *A Computational Theory of Human Linguistic Processing: Memory Limitations and Processing Breakdown*. Pittsburgh, PA: Carnegie Mellon University Unpublished doctoral dissertation.

Gibson, E. (1998). Linguistic complexity: locality of syntactic dependencies. *Cognition* 68, 1–76. doi: 10.1016/S0010-0277(98)00034-1

Gibson, E. (2000). "The dependency locality theory: a distancebased theory of linguistic complexity," in *Image, Language, Brain*, eds Y. Miyashita, A. Marantz and W. O Neil, (Cambridge, MA: MIT Press), 95–126.

Gordon, P., Hendrick, R., and Levine, W. (2002). Memory-load interference in syntactic processing. *Psychol. Sci.* 13, 425–430.

Grodner, D., Gibson, E., and Tunstall, S. (2002). Syntactic complexity in ambiguity resolution. *J. Mem. Lang.* 46, 267–295. doi: 10.1006/jmla.2001.2808

Haegeman, L. (1994). *Introduction to Government and Binding Theory, 2nd Edn.* Oxford: Blackwell.

Hakuta, K. (1981). Grammatical description versus configurational arrangement in language acquisition: the case of relative clauses in Japanese. *Cognition* 9, 197–236. doi: 10.1016/0010-0277(81)90001-9

Hale, J. (2001). "A probabilistic Earley parser as a psycholinguistic model," in *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies* (Stroudsburg, PA: Association for Computational Linguistics), 1–8. doi: 10.3115/1073336.1073357

Hare, M., McRae, K., and Elman, J. L. (2003). Sense and structure: meaning as a determinant of verb subcategorization preferences. *J. Mem. Lang.* 48, 281–303. doi: 10.1016/S0749-596X(02)00516-8

Heim, I., and Kratzer, A. (1998). *Semantics in Generative Grammar*. Malden, MA: Blackwell Publishers.

Hothorn, T., Bretz, F., and Westfall, P. (2008). Simultaneous inference in general parametric models. *Biom. J.* 50, 346–363. doi: 10.1002/bimj.200810425

Hukari, T. E., and Levine, R. D. (1995). Adjunct extraction. *J. Linguist.* 31, 195–226. doi: 10.1017/S0022226700015590

Jaeger, T. F. (2008). Categorical data analysis: away from ANOVAs (transformation or not) and towards logit mixed models. *J. Mem. Lang.* 59, 434–446. doi: 10.1016/j.jml.2007.11.007

Just, M. A., and Carpenter, P. A. (1992). A capacity theory of comprehension: individual differences in working memory. *Psychol. Rev.* 99, 122–149. doi: 10.1037/0033-295X.99.1.122

Just, M. A., Carpenter, P. A., and Woolley, J. D. (1982). Paradigms and processes and in reading comprehension. *J. Exp. Psychol.* 3, 228–238. doi: 10.1037/0096-3445.111.2.228

Kaan, E., Harris, A., Gibson, E., and Holcomb, P. (2000). The P600 as an index of syntactic integration difficulty. *Lang. Cogn. Process.* 15, 159–201. doi: 10.1080/016909600386084

Kimball, J. (1973). Seven principles of surface structure parsing in natural language. *Cognition* 2, 15–47. doi: 10.1016/0010-0277(72)90028-5

King, J. W., and Kutas, M. (1995). Who did what to when? Using word- and clauselevel ERPs to monitor working memory usage in reading. *J. Cogn. Neurosci.* 7, 376–395. doi: 10.1162/jocn.1995.7.3.376

Kluender, R., and Kutas, M. (1993). Bridging the gap: evidence from ERPs on the processing of unbounded dependencies. *J. Cogn. Neurosci.* 5, 196–214. doi: 10.1162/jocn.1993.5.2.196

Laenzlinger, C. (1996). "Adverb syntax and phrase structure," in *Configurations: Essays on Structure and Interpretation*, ed A. M. Di Sciullo (Somerville, MA: Cascadilla Press), 99–127.

Lee, M. W. (2004). Another look at the role of empty categories in sentence processing (and grammar). *J. Psycholinguist. Res.* 33, 51–73. doi: 10.1023/B:JOPR.0000010514.50468.30

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition* 106, 1126–1177. doi: 10.1016/j.cognition.2007.05.006

Lewis, R. L. (1996). Interference in short-term memory: the magical number two (or three) in sentence processing. *J. Psycholinguist. Res.* 25, 93–115. doi: 10.1007/BF01708421

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychol. Rev.* 63, 81. doi: 10.1037/h0043158

Mitchell, D. C. (1987). "Lexical guidance in human parsing: locus and processing characteristics," in *Attention and Performance XII: The Psychology of Language*, ed M. Coltheart (Hillsdale, NJ: Lawrence Erlbaum Associates), 601–618.

Omaki, A., Lau, E. F., Davidson White, I., Dakan, M. L., Apple, A., and Phillips, C. (2015). Hyper-active gap filling. *Front. Psychol.* 6:384. doi: 10.3389/fpsyg.2015.00384

Phillips, C., Kazanina, N., and Abada, S. H. (2005). ERP effects of the processing of syntactic long-distance dependencies. *Cogn. Brain Res.* 22, 407–428. doi: 10.1016/j.cogbrainres.2004.09.012

Pollard, C., and Sag, I. (1994). *Head-Driven Phrase Structure Grammar*. Chicago; Stanford: University of Chicago Press and CSLI Publications.

Pritchett, B. L. (1988). Garden path phenomena and the grammatical basis of language processing. *Language* 64, 539–576. doi: 10.2307/414532

Pritchett, B. L. (1991). "Subjacency in a principle-based parser," in *Principle-based Parsing: Computation and Psycholinguistics*, eds R. C. Berwick, S. P. Abney and C. Tenny (Dordrecht: Kluwer), 301–345.

Pritchett, B. L. (1992). *Grammatical Competence and Parsing Performance*. Chicago, IL: University of Chicago Press.

Rayner, K., Carlson, M., and Frazier, L. (1983). The interaction of syntax and semantics during sentence processing: eye movements in the analysis of semantically biased sentences. *J. Verbal Learn. Verbal Behav.* 22, 358–374. doi: 10.1016/S0022-5371(83)90236-0

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. Austria: R Foundation for Statistical Computing Vienna.

Riach, C. (2014). *Processing wh-adjuncts under the Active Filler Strategy.* Master's thesis, University of Nova Gorica, Nova Gorica.

Rizzi, L. (2001). "On the position Int(errogative) in the left periphery of the clause," in *Current Studies in Italian syntax. Essays offered to Lorenzo Renzi*, ed G. Cinque and G. Salvi (Oxford: Elsevier), 287–296.

Sag, I., and Fodor, J. D. (1994). "Extraction without traces," in *Proceedings of the Thirteenth Annual Meeting of the West Coast Conference on Formal Linguistics*, (Stanford, CA: CSLI Publications), 365–384.

Shlonsky, U., and Soare, G. (2011). Where's "why"? *Linguist. Inq.* 42, 651–669. doi: 10.1162/LING_a_00064

Snedeker, J., and Trueswell, J. C. (2004). The developing constraints on parsing decisions: the role of lexical-biases and referential scenes in child and adult sentence processing. *Cogn. Psychol.* 49, 238–299. doi: 10.1016/j.cogpsych.2004.03.001

Stabler, E. (1994). "The finite connectivity of linguistic structures," in *Perspectives on Sentence Processing*, eds C Clifton, L. Frazier and K. Rayner (Hillsdale, NJ: Erlbaum), 303–336.

Stepanov, A., and Tsai, W. T. D. (2008). Cartography and licensing of wh-adjuncts: a cross-linguistic perspective. *Nat. Lang. Linguist. Theory* 26, 589–638. doi: 10.1007/s11049-008-9047-z

Stowe, L. A. (1986). Parsing wh-constructions: evidence for on-line gap location. *Lang. Cogn. Process.* 1, 227–245. doi: 10.1080/01690968608407062

Travis, L. (1988). *The syntax of adverbs. McGill Working Papers in Linguistics, Special Issue on Comparative German Syntax*. Montreal, QC: McGill University, 280–310.

Trueswell, J. C., Tanenhaus, M. K., and Kello, C. (1993). Verb-specific constraints in sentence processing: separating effects of lexical preference from garden-paths. *J. Exp. Psychol.* 19:528. doi: 10.1037/0278-7393.19.3.528

Ueno, M., and Garnsey, S. M. (2007). "Gap-filling vs. filling gaps: event-related brain indices of subject and object relative clauses in Japanese," in *Proceedings of the 15th Japanese/Korean Linguistics Conference,* eds N. H. McGloin and J. Mori (Stanford, CA: CSLI Publications), 288–301.

Wanner, E., and Maratsos, M. (1978). "An ATN approach in comprehension," in *Linguistic Theory and Psychological Reality*, eds M. Halle, J. Bresnan and G. Miller (Cambridge, MA: MIT Press), 119–161.

Weinberg, A. (1999). "A minimalist theory of human sentence processing," in *Working Minimalism*, eds S. Epstein and N. Hornstein (Cambridge, MA: MIT Press), 282–315.

Yngve, V. H. (1960). A model and an hypothesis for language structure. *Proc. Am. Philos. Soc.* 104, 444–466.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2015 Stepanov and Stateva. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 20 May 2015
doi: 10.3389/fpsyg.2015.00654

An fMRI study dissociating distance measures computed by Broca's area in movement processing: clause boundary vs. identity

Andrea Santi[1]*, *Angela D. Friederici*[2], *Michiru Makuuchi*[2] and *Yosef Grodzinsky*[3,4]

[1] Department of Linguistics, University College London, London, UK, [2] Department of Neuropsychology, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany, [3] Edmond and Lily Safra Center for Brain Research and Language, Logic and Cognition Center, The Hebrew University of Jerusalem, Jerusalem, Israel, [4] Institute of Neuroscience and Medicine (INM-1), Forschungszentrum Jülich, Jülich, Germany

Behavioral studies of sentence comprehension suggest that processing long-distance dependencies is subject to interference effects when Noun Phrases (NP) similar to the dependency head intervene in the dependency. Neuroimaging studies converge in localizing such effects to Broca's area, showing that activity in Broca's area increases with the number of NP interveners crossed by a moved NP of the same type. To test if NP interference effects are modulated by adding an intervening clause boundary, which should by hypothesis increase the number of successive-cyclic movements, we conducted an fMRI study contrasting NP interveners with clausal (CP) interveners. Our design thus had two components: (I) the number of NP interveners crossed by movement was parametrically modulated; (II) CP-intervention was contrasted with NP-intervention. The number of NP interveners parametrically modulated a cluster straddling left BA44/45 of Broca's area, replicating earlier studies. Adding an intervening clause boundary did not significantly modulate the size of the NP interference effect in Broca's area. Yet, such an interaction effect was observed in the Superior Frontal Gyrus (SFG). Therefore, the involvement of Broca's area in processing syntactic movement is best captured by memory mechanisms affected by a grammatically instantiated type-identity (i.e., NP) intervention.

Keywords: fMRI, working memory, syntactic processing, movement, Broca's area

OPEN ACCESS

Edited by:
Matthew Wagers, University of California, Santa Cruz, USA

Reviewed by:
Ellen F. Lau, University of Maryland, USA
Michael Walsh Dickey, University of Pittsburgh, USA

***Correspondence:**
Andrea Santi, Department of Linguistics, University College London, Chandler House, 2 Wakefield Street, London WC1N IPF, UK
a.santi@ucl.ac.uk

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: *19 December 2014*
Accepted: *04 May 2015*
Published: *20 May 2015*

Citation:
Santi A, Friederici AD, Makuuchi M and Grodzinsky Y (2015) An fMRI study dissociating distance measures computed by Broca's area in movement processing: clause boundary vs. identity. Front. Psychol. 6:654. doi: 10.3389/fpsyg.2015.00654

Introduction

There is extensive evidence that Broca's area is taxed by sentences with *movement* both from neuropsychological studies of patients and neuroimaging studies of healthy adults (Just et al., 1996; Stromswold et al., 1996; Caplan et al., 1999; Ben-Shachar et al., 2003, 2004; Fiebach et al., 2005; Grewe et al., 2005). Less complex relations, such as simple phrasal composition and local agreement have also been shown to activate/depend on this region (Pallier et al., 2011; Carreiras et al., 2012), however, they have not done so as consistently across methods and populations, as movement (for lack of evidence for simple composition in imaging see Humphries et al., 2005; Brennan et al., 2012). Our goal in this paper is to push our understanding of this special relation between movement and Broca's area even further.

Recent work suggests that activation of Broca's area with syntactic movement may be specifically tied to memory interference, as activity appears to increase with each additional NP intervener within the movement dependency (Santi and Grodzinsky, 2007b; Makuuchi et al., 2013). In the current fMRI study we ask whether this interference effect is modulated by the number of intervening clause boundaries (0 vs. 1). As a clausal boundary increases the number of movements, this manipulation is particularly relevant to theories that place a special role for Broca's area in computing movement dependencies (Grodzinsky, 2000; Grodzinsky and Santi, 2008). Below we elaborate on the structural properties of movement that can be cashed-in as costly for processing mechanisms potentially located within this brain region. While many theoretical positions have been put forth in accounting for this effect, we will argue for the strength of an interference-based account, where interveners are of the same syntactic/semantic type as the moved phrase (type-identical interference henceforth), as opposed to others, for example the number of iterations of a local movement operation.

In sentences with *Movement* (2), a single Noun has (at least) two dependent positions that provide distinct interpretations (e.g., in (2) interrogative and thematic). Only one of these positions is pronounced (2,3), the other(s) copy is <bracketed>, silent and is where the noun is interpreted (thematically) as an argument of a predicate. In contrast, sentences without movement (1) have no silent copy and only one interpretive position for each noun.

1. The boy likes the girl.
2. **Who** does the boy like <***who***>?
3. **Who** <***who***> likes the boy?

Many investigations into movement processing have been based on the object vs. subject movement asymmetry. Object movement (2) unlike subject movement (3) has lexical material intervening between the pronounced position of the noun and where it gets thematically interpreted. Furthermore, the ordering of arguments is non-canonical in the case of object movement (Object-Subject-Verb, above).

The difficulty associated with processing object compared to subject movement has been largely attributed to the degree of referential similarity between the intervening argument(s) and the moved one (Gordon et al., 2001). In a behavioral study, Gordon et al. (2001) studied subject and object extracted relative clauses whereby the head of the relative clause was an NP that was a definite description (e.g., "the barber" in 4 and 5) and the NP within the relative clause was either also descriptive (e.g., "the lawyer") or a proper name (e.g., "Joe"). Reading times at the two critical words (those underlined in the example sentence in 4 and 5) demonstrated an interaction. Reading times were longer for object-extracted relative clauses compared to subject-extracted ones, when the NP within the relative clause was of the same type as the filler (i.e., descriptive). When a proper name was used there was little if any difference between object and subject extracted relative clauses.

4. The barber that **the lawyer/Joe** admired <*the barber*> climbed the mountain.
5. *The barber* that <*the barber*> admired the *lawyer/Joe climbed* the mountain.

This result demonstrates that the parser is sensitive to the syntactic and/or semantic similarity of features (e.g., +sing, +animate, +definite) between referential items.

Additional behavioral studies have reinforced the idea that long distance dependencies, more generally, are difficult to process when there is a similar intervener. These studies do not focus on referential features, but the syntactic position of the intervening material (Van Dyke, 2007). For example, a subject of a complement clause creates more interference within a subject-verb dependency than does the same NP within an object PP. Thus, a broad range of features (+nom, +animate, +singular, +definite, etc.) may contribute to similarity-based interference during dependency resolution, but their degree of contribution may depend on the particular dependency under investigation.

The finding that Broca's area is sensitive to object movement (Just et al., 1996; Stromswold et al., 1996; Caplan et al., 1999; Fiebach et al., 2005; Grewe et al., 2005) is reinforced by more sophisticated parametric fMRI studies. These studies quantified how taxing movement is by the amount of intervening units between the dependent elements, where units (i.e., "interveners") have most often been defined as animate, singular, descriptive NPs. Animate, singular descriptive NPs were selected, as they share syntactic and semantic features with the moved phrase, thereby introducing semantic/syntactic identity based interference in memory processes (Gordon et al., 2001). For an example of a parametric manipulation of number of similar interveners, see 6(a–d) from Makuuchi et al. (2013).

6. *a. Ich glaube, der Mann zeigte dem Kind den Onkel gestern Abend.*
   b. I think, the$_{NOM}$ man showed the$_{DAT}$ boy the$_{ACC}$ uncle last evening.
   c. I think, *the$_{DAT}$ boy* **the**$_{NOM}$ **man** showed <*the$_{DAT}$ boy*> the$_{ACC}$ uncle last evening.
   d. I think, *the$_{ACC}$ uncle* **the**$_{NOM}$ **man** showed **the**$_{DAT}$ **boy** <*the$_{ACC}$ uncle*> last evening.

The baseline sentence is presented in 6a in German and 6b presents the English gloss. In this baseline sentence all arguments are in their base position. In 6c, the direct object has moved in front of the subject (crossing 1 NP) whereas in 6d the indirect object has moved in front of the subject (crossing 2 NPs). Previous parametric studies investigated the neural reflections of the number of NPs crossed (i.e., interveners) by a *single* moved NP (Santi and Grodzinsky, 2007b; Makuuchi et al., 2013) *or* of the number of NPs displaced by syntactic movement (Friederici et al., 2006) across different languages (English, German) and movement constructions (Scrambling, Topicalization, Relative Clauses). Their results provide a neurocognitive generalization: Broca's area is sensitive to movement distance measured by the number of similar interveners (in this case type-identical NPs) that moved NPs cross. In conjunction with the results from additional fMRI studies, this interference appears to be occurring proactively rather than retroactively, given that dependencies,

which are not predictable until the tail of the dependency (e.g., reflexive binding and parasitic gaps), do not engage Broca's area (Santi and Grodzinsky, 2007a,b). Thus, it would seem that object movement is taxing due to maintenance of a prediction (i.e., gap for an NP) that crosses type-identical interveners (i.e., NP).

A recent fMRI study (Glaser et al., 2013) showed similarity of an NP intervener to the head of the dependency is critical in driving activation in BA44 and 45 (i.e., Broca's area). This particular study did not assess interference within a movement dependency, but a subject-verb (agreement) dependency. The high interference condition had an intervening subject NP (visitor)[1] within a complement clause (8), whereas the low interference condition had an intervening NP (that was not subject) within a PP (7). The greater activation within Broca's area for (8) than (7) was interpreted to reflect the main verb (i.e., was complaining) cueing for the retrieval of a subject NP whereby an intervening subject NP resulted in greater interference. Thus, unlike our conclusions above, they assume that similarity-based interference effects in Broca's area occur during a cue-based retrieval.

7. *The client* who had arrived after the important **visitor** that day *was complaining* about the investigation.
8. *The client* who implied that the **visitor** was important that day *was complaining* about the investigation.

Whether interference is occurring proactively or retroactively, conflict resolution can apply in recovering the correct representation (Thompson-Schill et al., 1997; Novick et al., 2005; Thothathiri et al., 2012). Thothathiri et al. (2012) specifically suggest that non-canonical structures activate Broca's area due to syntactic competition between an agent-first hypothesis and the actual syntactic representation, which is patient-first in the cases of object-relatives and passives. Conflict resolution is relied-on to distinguish the correct from the incorrect representation. Thus, conflict resolution may apply following interference and be the basis of the observed activation in Broca's area.

Although there is indication that Broca's area is engaged by interference generated by the number of NP interveners (whether affecting proactive, retroactive, or both aspects of processing) crossed by a movement dependency, movement may engage additional processing mechanisms within this region. However, the nature of the tests conducted thus far cannot address this. Multiple distinct computations within Broca's area is not unreasonable, given that it contains multiple anatomical subregions with presumably distinct functions (Amunts et al., 2010). Our goal in this study is to determine whether movement has effects in Broca's area above and beyond those imposed by the semantic/syntactic identity of intervening NPs within a movement dependency. Specifically, does the number of movements affect activation in any subregions of Broca's area or surrounding regions, as another neurolinguistic account of Broca's area has proposed it is involved in computing syntactic movement (Grodzinsky, 2000). We investigated this with sentences involving iterations of a movement operation (i.e., successive cyclic movement) as compared to sentences with a single movement (within a clause) but with an equal number of NPs crossed by that movement. The following provides a brief description of how movement proceeds successive-cyclically after which we will further elaborate on the complexity dimensions tested.

As discussed above, movement involves an interpretation of a phrase in a position that is not pronounced (i.e., silent copy). In those examples we were concerned with a single clause. By comparison, in sentences with multiple clauses, the wh-phrase (i.e., who) moves from a thematic (i.e., *doer* or *doee*), silent position which it "vacates" (gap) to a "filled" position (filler), in which it is pronounced, by stopping off at the left edge of *each* intervening clause and leaving behind a silent copy in each of them (10). Evidence that the wh-phrase moves through intermediate CPs (i.e., CP3 in 9) on the way to its final destination (i.e., CP2 in 10) comes from grammaticality contrasts, as in (9) vs (10). Both (9) and (10) are composed of 3 clauses (CPs). Note that in (9) the *wh*-phrase (i.e., who) crosses more words than in (10) along the path from thematic interpretation to its pronounced position, but (10) is ungrammatical and (9) is not. This grammaticality contrast can be explained by considering that in (9) the *wh*-phrase has an intermediate landing position available (left edge of CP3) that is not available in (10) because the intermediate position is already filled by another wh-phrase (*which boy*). It has thus been proposed that wh-phrases *must* move successively through each CP on the way to their final landing position, leaving traces or silent copies (identified by phrases in angled brackets) in these intermediate positions, because failure to do so results in ungrammaticality (10). This captures the successive-cyclic nature of *movement*.

9. [$_{CP1}$ I know [$_{CP2}$ *who* the teacher from Norway thinks [$_{CP3}$ <*who*> the boy likes <*who*>]]].
10. *[$_{CP1}$ I know [$_{CP2}$ *which girl* the teacher thinks [$_{CP3}$ which boy likes <*which girl*>]]].

Further, evidence for intermediate landing positions is provided by language acquisition studies, which show children produce wh-words in these intermediate positions (Thornton, 1995). Likewise, Psycholinguistic studies have provided support for intermediate positions (Gibson and Warren, 2004), through demonstrating that intermediate positions ease processing of a sentence-final silent copy relative to comparable length dependencies not involving embedded CPs, achieved through nominalization.

The current study had two design features: (1) we manipulated the number of NP interveners crossed by a

[1]Although note that in (7) the NP is modified by the adjective, "dangerous," whereas in (8) it is not. Thus type similarity between the intervening NP and the head of the dependency also differs across this contrast.

moved NP (Baseline:NP/CPS0, 1NP intervener:NP/CP/O1, 2NP intervener:NP/CPO2 in **Table 1**) and (2) compared successive cyclic movement to a single movement while controlling for number of intervening NPs (see **Table 1**). The first part of the design allowed us to relate the novel design/results to previous results that investigated a parametric manipulation in the number of intervening NPs. The second part allows us to test whether number of movements has an effect above the number of similar NPs crossed.

The baseline condition (CP/NPS0) involves a local subject (S) movement, hence crossing 0 similar NPs. This was compared to movement that crossed 1 similar NP (CP/NP O1); in order to accomplish this the object of the most embedded clause was moved across the subject of that same clause. Furthermore, this was compared to movement that crossed 2 similar NPs (CP/NP O2), which was accomplished again via object movement, either across the two subjects of the two most embedded clauses (CP condition) or across the direct object and subject in a single clause, containing a bi transitive verb (NP condition). This contrast of size to similarity addresses what form of information increases complexity of memory mechanisms in Broca's area. Thus, by comparing condition CPO2 to CPO1 we have a contrast in number of CPs crossed (2 vs. 1) and in contrasting NPO2 to NPO1 we have a contrast in number of NPs crossed (2 vs. 1). Furthermore, collapsing across the two types of interveners we can re-assess the parametric effect of number of similar interveners in comparing the current work to past results.

Although our primary interest was in investigating the effect of multiple movements, it is important to note that multiple movements have a couple of consequences that in and of themselves may increase processing complexity. The multiple movements coincide with a larger syntactic size of the "interveners" (i.e., CP) or put otherwise refers to movement that crosses a clausal boundary. In successive-cyclic movement we are crossing multiple clauses rather than a single one containing some multiple of NPs. CPs contain many more functional projections (i.e., CPs, and tense and agreement checking nodes) and as such are syntactically more complex[2]. Wagers and Phillips (2014) show that movement within a clause involves active maintenance of both coarse (e.g., category) and fine-grained (lexical semantic) information about the antecedent, but across clauses there is active maintenance of just the coarse-grained information, whereby fine-grained lexical information needs to be retrieved at the gap. Thus, a clause boundary manipulation should engage retrieval processes more than one without.

We can test whether crossing a clausal boundary of a wh-movement dependency has an effect on the fMRI signal above that of similarity of the intervener (i.e., NP) to the moved constituent by comparing crossing of 2CPs to 1CP with crossing 2NPs to 1NP (in a single clause). Note, in **Table 1**, the type contrast (CP, NP) does not differ in terms of dependency distance when there is either 1 or 0 intervener. Thus, one would only expect a difference between the intervener types when comparing 2 vs. 1 intervener (i.e., hence CPS0 is grayed out in **Table 1** to highlight the conditions contributing to the expected interaction effect). In summary, an enhancement of activation for crossing a clause could indicate 1 of 2 related processes: (1) number of movements or (2) taxing retrieval mechanisms more due to crossing a clausal boundary.

Any results from the current study that demonstrate an effect of number of CPs over NPs cannot distinguish between number of movements, syntactic size of the intervening material and crossing a clausal boundary. Nonetheless, the data will critically show whether or not Broca's area is sensitive to movement (and syntactic size or crossing clausal boundary) beyond similarity of the interveners to the head of the dependency. The potential complexity factors induced by successive-cyclic movement above type-identical based interference are interrelated (perhaps, reflections of different levels of analyses) and as such not easily disentangled, these include: (1) number of movements (and silent copies) (2) syntactic size of intervening material (between the pronounced and thematically interpreted NP), which involves the crossing of a clausal boundary.

TABLE 1 | Example Stimuli.

CP INTERVENER SENTENCES	
CPS0	I said the neurosurgeon knew *which resident* liked the porter
CPO1	I said the neurosurgeon knew [*which porter* **the resident liked** <which porter>]$_{CP1}$
CPO2	I knew [*which porter* **the neurosurgeon said**]$_{CP2}$ [<which porter> **the resident liked** <which porter>]$_{CP1}$
NP INTERVENER SENTENCES	
NPS0	I knew *which neurosurgeon* showed the resident to the porter
NPO1	I knew *which resident* [**the neurosurgeon**]$_{NP1}$ showed <which resident> to the porter
NPO2	I knew *which porter* [**the neurosurgeon**]$_{NP1}$ showed [**the resident**]$_{NP2}$ to <which porter>

The CP conditions include CPS0, CPO1, CPO2, and the NP conditions include NPS0, NPO1, NPO2. Subject movement conditions that cross 0 NPs (S0) conditions have an embedded wh-subject phrase that does not leave its clause. The object movement condition have an embedded wh-object phrase that crosses one (O1) or two (O2) interveners (defined as either CPs or NPs), respectively. CPS0 and NPS0 along with CPO1 and NPO1 are no different in terms of intervention across a movement dependency.

Methods

Subjects

Twenty one subjects participated in the study (after exclusion of two participants from the analysis due to low behavioral performance in the fMRI study (<65%)[3]. The average age of

[2]This is relevant given that Glaser et al. (2013) compared an intervening CP to an intervening PP in testing effects of (subject) NP interference, where the intervening CP condition was also the condition with "high-syntactic interference" and resulted in greater activation in Broca's area. The question remains whether this greater activation is due to greater syntactic structure intervening the dependency or the subject status of the NP within this structure.

[3]This level of accuracy is based on the fact that the sentences are quite complicated and additionally the offline comprehension questions were difficult, as they involved a thematic role reversal. Further, it is not necessarily the case that incorrect answers to an off-line comprehension question correspond to an incorrect parse online. Rather it may simply be the product of an incorrect memory of that parse. As will be discussed later on, in the lab prior to fMRI scanning, participants performed at 75% or higher in each condition and as is normal, performance became a bit worse in the peculiar environment of an fMRI machine.

participants was 19.90 years, and 12 were female. All subjects were right-handed according to the Edinburgh Handedness Inventory (Oldfield, 1971), had normal or corrected-to-normal vision, a score above 3 on the Daneman and Carpenter Reading Span Test (Daneman and Carpenter, 1980), and gave informed consent in accordance with the ethics committee of the Montreal Neurological Institute (MNI).

Stimuli

The design of the stimuli crossed NUMBER of intervener (0, 1, 2) with TYPE of intervener (CP, NP). Although as pointed out in the Introduction, the distinction across "Type" for our purposes only arises when the Number of interveners is 2. Each condition was made up of 40 sentences and every sentence was between 17 and 19 syllables in length. In the Intervening CP condition there were two embedded clauses allowing for two successive-cyclic movements from baseline. In the intervening NP condition, to allow for movement over multiple NPs, but not CPs, there was one embedded clause that contained a double-object verb. The following sections contain further detailed descriptions of the parameterization of distance for each intervener type (see **Table 1** for example stimuli and Supplementary Materials for full list of Stimuli).

Intervening CPs

All sentences started with a pronoun (I, we, he, she) followed by a verb that takes a sentential complement (thought, claimed, hoped, said). The sentential complement was composed of an NP and another verb (knew, learned, announced) that takes a sentential complement. This second embedded clause was composed of an NP, a verb and direct object NP. In CPS0 there is no movement over a clause but there is (or may be) movement into a CP (unless one does not assume string vacuous movement)[4]. From baseline there is movement of the second embedded object to the front of the second embedded clause (CPO1 in **Table 1**), or movement of the second embedded object to the front of the first embedded clause (CPO2).

Intervening NPs

Likewise in the NP condition the sentences began with a pronoun (I, we, he, she) followed by a verb that takes a sentential complement (knew, announced, learn), the sentential complement was made up of a subject NP, a double object verb (introduced, described, showed, recommend) and its direct object NP and indirect object NP. In the baseline condition, NPS0, there is no movement over an NP, but there is (or may be) movement of the subject into a CP (unless one does not assume string vacuous movement). From baseline there is movement of the direct object in front of the embedded subject (NPO1). The second parameterization moved the indirect object over both the direct object and embedded subject (NPO2).

Only two subjects had an individual condition with an average accuracy of 65%, but over all conditions each subject performed above 70% and typically above 80%.

[4]An alternative baseline with no movement, but rather a complementizer was considered. It was ruled out as problematic, given it would be the only condition without a wh-phrase. This confound would make interpretations of the data difficult.

Procedure

To assure the participants were processing and understanding the sentences a yes/no question about stimulus content followed 50% of the sentences. Half of these required a "yes" response and half a "no" response. Questions requiring a "no" response involved a thematic role reversal (see 7–8 below). Given the difficulty of the task, we wanted to be assured that there would be a low exclusion rate in the fMRI study. Thus we screened subjects before fMRI scanning for behavioral performance days to weeks before the actual fMRI session. During screening, participants performed the task on 50% of the stimuli and were included for the fMRI study if they performed at 75% or greater in every condition. We screened 52 people whereby 28 satisfied all requirements (including handedness, language and behavioral performance). Of the remaining 24 that did not satisfy the screening requirements, 14 of them did not satisfy the requirements for behavioral performance alone. Of those 14, 10 still scored above 75% on average across all conditions. Thus, many simply performed below the conservative threshold on 1 or 2 of the conditions. Half of the subjects were screened on one-half of the sentences and the other on the complementary set. Both groups of participants saw the complementary set of comprehension questions from their screening session in the actual fMRI study, and both saw the entire set of sentences (the full set of items across all conditions). Thus, the half they saw in practice they saw again during the fMRI study that was run days or weeks later. Therefore, comprehension sentences only appeared on 50% of the trials in the fMRI study.

7. I said the neurosurgeon knew which porter the resident liked.
8. Did I say the neurosurgeon knew which resident the porter liked?

The stimuli were programmed with Presentation software (Neurobehavioral Systems, Inc., Albany, California, USA) on a Windows PC. The stimuli were projected onto a screen at the back of the MRI and then reflected into a mirror attached to the head coil. The sentences appeared word/phrase by word/phrase (see **Figure 1**). Each word/phrase appeared for 700 ms with 100 ms between. The comprehension question was presented for 4000, 100 ms after the sentence. On trials without comprehension questions there were 3 scans (4.8 s) of blank screen inter-trial interval (ITI), whereas on trials followed by comprehension questions there were 2.5 scans (4 s) of ITI. Half of the stimuli were presented in each of two runs. See **Figure 1** for a depiction of the trial dynamics. Trial order and additional interspersed silence (10*12.8 + 10*9.6 s) for jittering stimulus onset was optimized by optseq (http://surfer.nmr.mgh.harvard.edu/optseq/) with the presentation of the trial being jittered by 0 or 800 ms from the onset of the scan. Run order was counterbalanced across participants. An MRI compatible response box for comprehension question responses was placed in the participants' left hand to avoid potential motor activation overlapping with typically left frontal language activation.

Image Acquisition

Functional and structural data were acquired on a 3T Siemens magnetom Triotim. Twenty-six slices, 4 mm thick oriented

CP	**700 ms**	**100 ms**	**700ms**	**100 ms**	**700ms**	**100 ms**	**700ms**	**100 ms**	**700ms**	**100 ms**	**700ms**	**100 ms**	**4000ms**
	I said		the neurosurgeon		knew		**which resident**		liked		the porter		Question
NP	**700 ms**	**100 ms**	**700ms**	**100 ms**	**700ms**	**100 ms**	**700ms**	**100 ms**	**700ms**	**100 ms**	**700ms**	**100 ms**	**4000ms**
	I knew		**which neurosurgeon**		showed		the resident		to		the porter		Question

FIGURE 1 | Trial Dynamics of sentence presentation phrase/word by phrase/word.

AC-PC, with full coverage of the frontal, temporal, and occipital lobes and partial coverage of the parietal lobes were acquired (TR = 1.6 s, TE = 30 ms, Flip angle = 90°, FOV = 25.6 × 25.6 cm^2, 64 × 64 matrix). Superior aspects of the parietal lobe could not be included to maintain the desired functional and anatomical resolution. Voxels were 4 × 4 ×4 mm in volume. There were 176, 1 mm thick structural scans acquired with an MPRage sequence (TR = 2300 ms, TE = 2.98 ms, FOV = 256 ×240 mm, 256 × 240 matrix). During scanning, an air vacuum pillow and sponges were used to stabilize the head.

Analysis

Behavioral Data

Mean reaction times (RT) and accuracy for each subject and condition was entered into a 2 TYPE (NP, CP) by 3 Distance (0, 1, 2) Repeated Measures ANOVA (both by subjects and by items).

fMRI Data

The first 4 volumes of each fMRI run were removed from the analysis, in order to exclude magnetic saturation effects. The data were analyzed in SPM8 (available at http://www.fil.ion.ucl.ac.uk/spm/). Functional images were aligned to the first image and resliced in order to correct for motion. Then coregistration between functional and anatomical images was performed. Anatomical images were segmented and normalized to MNI space. The resultant transformation matrix was applied to the functional images that were subsequently spatially smoothed with an 8 mm FWHM Gaussian kernel. The data were modeled with regressors for each sentence condition and 1 regressor for all comprehension questions and convolved with a canonical model with a time derivative. The time derivative was applied to handle slice timing differences (Henson et al., 1999). A high pass filter with a cut-off of 128 s was applied to the data. The contrast images for each condition of each subject were submitted to a second-level (group) analysis: (1) 2TYPE(CP, NP) × 3Number(S0, O1, O2) within-subject ANOVA. F-test of the interaction was FWE corrected for multiple comparisons. T-tests were used to test for a linear effect of Number [-1 0 1 -1 0 1] (CP/NPO2>CP/NPO1>CP/NOS0) to replicate previous studies that have demonstrated an effect of number of NPs intervening a movement dependency. Main effects of Type (CP>NP) and (NP>CP) were coded as t-tests as well [1 1 1 -1 -1 -1] and [-1 -1 -1 1 1 1], respectively. These effects compare multiple syntactic factors (e.g., verb argument structure, number of clauses) so the interpretation of any such results need to be made with caution, but nonetheless provide further data considering syntactic differences in processing. Additionally, the interaction to test for an effect of syntactic size (CPO2-CPO1>NPO2-NPO1) was coded as a t-tests [0 -1 1 0 1 -1]. Again, this particular interaction test was to address whether an intervening clause modulates the effect of an intervening NP. The effect of an additional clause vs. NP is only provided by the 2 intervener condition (CPO2 vs. NPO2 condition). The t-test maps were thresholded at voxel-wise $p < 0.005$ for signal intensity and by a cluster size where only significant clusters ($p < 0.05$) were reported.

The anatomy toolbox (www.fz-juelich.de/ime/spm_anatomy_toolbox; Eickhoff et al., 2005) was used for the identification of cytoarchitectonic probability of cluster localization. The Marsbar toolbox (Available at http://marsbar.sourceforge.net) was used for extracting Percent Signal Change from clusters.

Results

Behavioral Results

The accuracy results demonstrated very high (>85%) accuracy rates (see **Figure 2**). A main effect of DISTANCE was nevertheless observed over subjects [$F_{1(2, 40)} = 15$, $p < 0.001$] and items [$F_{2(1.476, 57.55)} = 8.95$, $p < 0.001$]. Pairwise comparisons with a Sidak correction for multiple corrections showed that S0 was significantly more accurate than O1 ($p = 0.007$) and O2 ($p < 0.001$), but that O1 and O2 did not significantly differ from one another ($p = 0.251$) both in the subjects and items analysis. Although the conditions did not directly differ there was a significant linear decrease in accuracy (i.e., Linear effect) with DISTANCE in the subjects [$F_{1(1, 20)} = 37.60$, $p < 0.001$] and items [$F_{2(1, 39)} = 17.19$, $p < 0.001$]. This indicates that accuracy demonstrated a decreasing trend with increasing distance even though direct contrasts did not turn out significant. Neither the main effect of TYPE or the interaction of TYPE*DISTANCE were significant. The RT results (see **Figure 3**) likewise demonstrated a main effect of DISTANCE in the subjects [$F_{1(2, 40)} = 6.94$, $p < 0.003$] and items [$F_{2(2, 78)} = 6.35$, $p = 0.003$] and Linear Effect of DISTANCE in the subjects [$F_{1(1, 20)} = 10.66$, $p < 0.004$] and items [$F_{2(1, 39)} = 12.62$, $p = 0.001$]. The main effect of DISTANCE was due to a faster reaction time for S0 than O2 ($p = 0.012$) in the subjects analysis and due to a faster reaction for S0 than both O1 ($p = 0.03$) and O2 ($p = 0.001$). There was also a main effect of TYPE in the subjects [$F_{1(1, 20)} =$

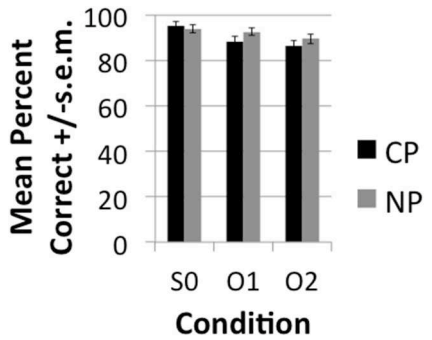


FIGURE 2 | Mean percent +/−s.e.m. correctly answered comprehension questions broken down by condition.

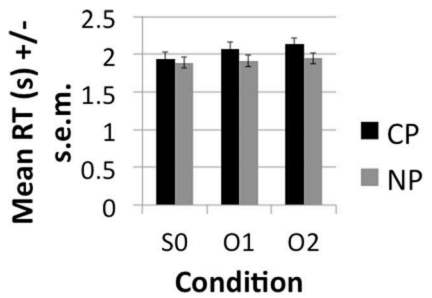


FIGURE 3 | Mean Reaction Time (RT) in seconds (s) ± s.e.m. to comprehension question broken down by condition.

19.23, $p < 0.001$] and items [$F_{2(1, 39)} = 28.9$] analyses and an interaction between TYPE and DISTANCE that was approaching significance in the subjects [$F_{1(2, 40)} = 3.22$, $p < 0.051$], but not the items [$F_{2(2, 78)} = 1.85$, $p = 0.164$] analysis. The trend of an interaction was due to CP interveners having a greater effect on slowing RT with increasing number of interveners than NP interveners. The main effect of TYPE was due to a slower RT for CP (mean = 2.05, SE = 0.084) than for NP (mean = 1.92, SE = 0.073).

fMRI Results

The current study tested whether an additional clause boundary within a wh-movement dependency has an effect on the fMRI signal above that of similarity of the intervener (i.e., NP) to the moved constituent. That is, it tested whether an additional clause boundary would have a greater effect on the fMRI signal than that of NP interveners.

Number of Intervener NPs

A significant linear effect of Number of interveners was observed bilaterally in the Inferior Frontal Gyrus (IFG) and in the Caudate Nucleus (see **Table 2** and **Figures 4**, **5**. for details). The anatomy toolbox, identified the peak LIFG activation (−40, 12 26) was within BA 44 with a probability of 30%. As can be seen in **Figure 5**, the LIFG activation is strongest in BA44 and spreads into the posterior portion of BA45. Across both the BA44 and 45

TABLE 2 | Regions activated by a linear effect of Distance.

Location	BA	Voxels	Clusterp	PeakZ	Coordinates
Left Inferior Frontal Gyrus	44/45	812	0.01	4.07	−40 12 26
				3.63	−38 4 34
				3.14	−42 22 16
Caudate Nucleus	–	488	0.038	3.70	−8 10 4
				3.51	8 8 0
Right Inferior Frontal Gyrus	44/45	436	0.048	3.38	52 12 26
				3.24	36 10 26
				2.91	44 4 34

Thresholded at a voxel-wise $p < 0.005$ and corrected for multiple comparisons by a cluster level $p < 0.05$.

probability maps, thresholded at 30% (as in **Figure 5**), 404 voxels of the linear activation cluster are within the maps (i.e., 50% of the cluster overlaps with the maps). When using unthresholded probability maps (i.e., 10–100%), 590 voxels of the cluster are contained within the probability maps of BA44/45 (i.e., 73%). The activation that is not overlapping with the probability maps is mostly due to medial and posterior extension of the activation. In addition to LIFG activation, both the caudate and right Broca's area demonstrated activation, but this activation occupied a much smaller cluster (about half the size) then that on the left. Further, in fMRI it is difficult to know the necessity of the area(s) activated and from additional studies, it would appear that right Broca's area is often activated (amongst patients and healthy participants), but unlike LIFG, is not causally involved in language processes (Thiel et al., 2006).

Additional Clause Boundary

There was no interaction effect in Broca's area, either defined by a linear increase (0 to 1 to 2) that is greater for CP interveners than NP interveners or in terms of the (2–1 intervener) subtraction having a greater effect for CP compared to NP interveners. In the *t*-tests, a significant interaction effect defined by a greater effect of number of interveners (2 vs. 1) in the CP condition than the NP condition was, however, localized bilaterally in the Superior Frontal Gyrus (SFG; see **Table 3** and Supplementary Figure).

Effect of Syntactic Type

There was a significant effect of TYPE in the Superior Temporal Sulcus (STS) and the Inferior Occipital Gyrus (IOG, see **Table 4**; **Figure 6**). This effect was due to the CP condition producing greater activation than the NP condition. There were no significant clusters that demonstrated greater activation for the NP condition over the CP condition.

Discussion

The novel result that this study presents is that while Broca's area is sensitive to the number of type-identical interveners in long distance wh-movement, this effect is not augmented by **a clausal boundary**. Other less prominent areas of activation that demonstrated this same effect were found in the right homolog of Broca's area and the caudate nucleus. On the other hand, more

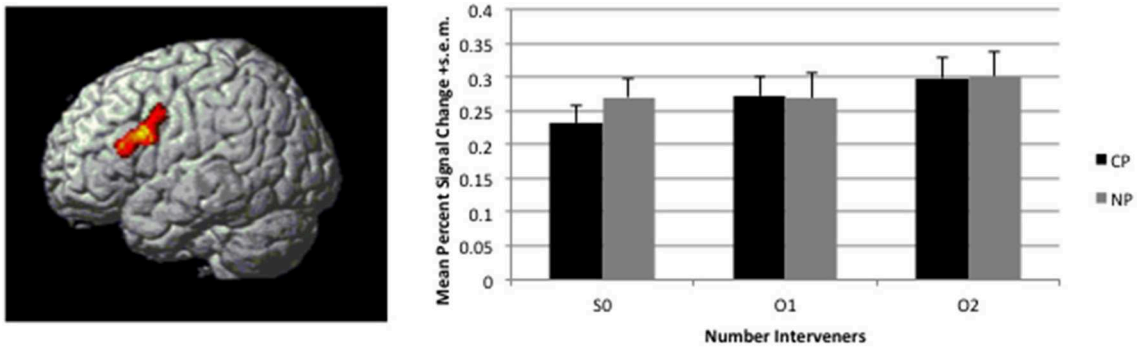


FIGURE 4 | Linear effect of distance (S0<O1<O2) observed in the 2Type(CP, NP) × 3Number(S0, O1, O2). Map is thresholded at voxel-wise $p < 0.005$ and corrected for multiple comparison through cluster size $p < 0.05$. Activation is overlaid on a rendered surface of the brain. The histogram to the right presents the mean percent signal change +s.e.m. over the entire cluster broken down by condition (percent signal change was calculated in Marsbar per subject).

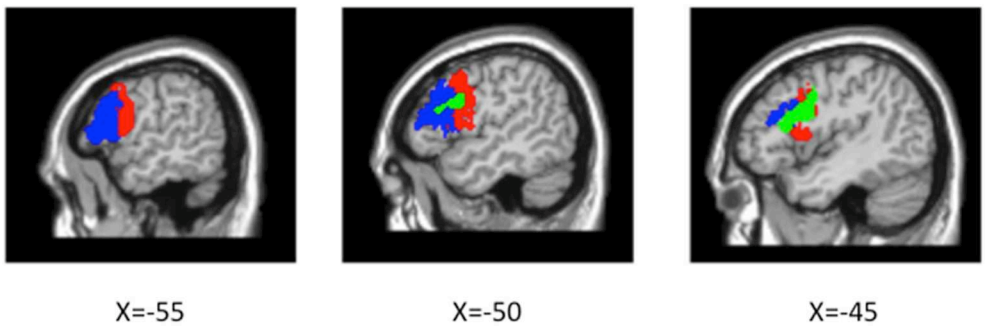


FIGURE 5 | Cytoarchitectonic probability maps of BA 44 (red) and BA 45 (blue) thresholded at 30% overlap overlaid on canonical average brain and linear main effect (voxel-wise $p < 0.005$, cluster-level $p < 0.05$; green) overlaid on top.

TABLE 3 | Regions activated by a 2Type(CP, NP) × 2Distance(O1, O2) interaction where (CPO2-CPO1)>(NPO2-NPO1).

Location	BA	Voxels	Clusterp	PeakZ	Coordinates
Superior Frontal Sulcus/Gyrus (SFS/SFG)		1696	0.001	4.95	44 38 32
				4.20	−6 42 50
				4.18	40 36 40

Thresholded at a voxel-wise $p < 0.005$ and corrected for multiple comparisons by a cluster level $p < 0.05$. This is one cluster that contains the bilateral SFS/G.

TABLE 4 | Regions activated by a main effect of Type (CP vs. NP).

Location	BA	Voxels	Clusterp	PeakZ	Coordinates
Left Inferior Occipital Gyrus		611	0.022	5.25	−22 −92 −8
Left Superior Temporal Sulcus		597	0.023	3.91	−56 −8 −12
				3.54	−52 −24 −4
				3.32	−50 −32 0

Thresholded at a voxel-wise $p < 0.005$ and corrected for multiple comparisons by a cluster level $p < 0.05$.

superior areas (i.e., SFG) were augmented by a clausal boundary (or the syntactic size of the intervener).

The result in Broca's area is consistent with psycholinguistic data that has demonstrated that the similarity of the interveners to the head of a movement dependency increases processing difficulty (Gordon et al., 2001). Our results further expand on these results in two ways: (1) by demonstrating that a syntactically similar intervener, but not an intervening clausal boundary, increases activation in Broca's area, its right homolog, and the basal ganglia, (2) a clausal boundary further increases complexity of a movement dependency, as evidenced by a marginally significant behavioral effect on offline RTs to verification questions and increased activation within the SFG and to some degree the left superior temporal cortex.

Broca's Area and Syntactic/Semantic Similarity Based Interference

Broca's area has been repeatedly reported to be engaged by object movement dependencies (Just et al., 1996; Stromswold et al., 1996; Caplan et al., 1999; Ben-Shachar et al., 2003, 2004; Fiebach et al., 2005; Grewe et al., 2005). Here we have explicitly framed movement distance in terms of number of

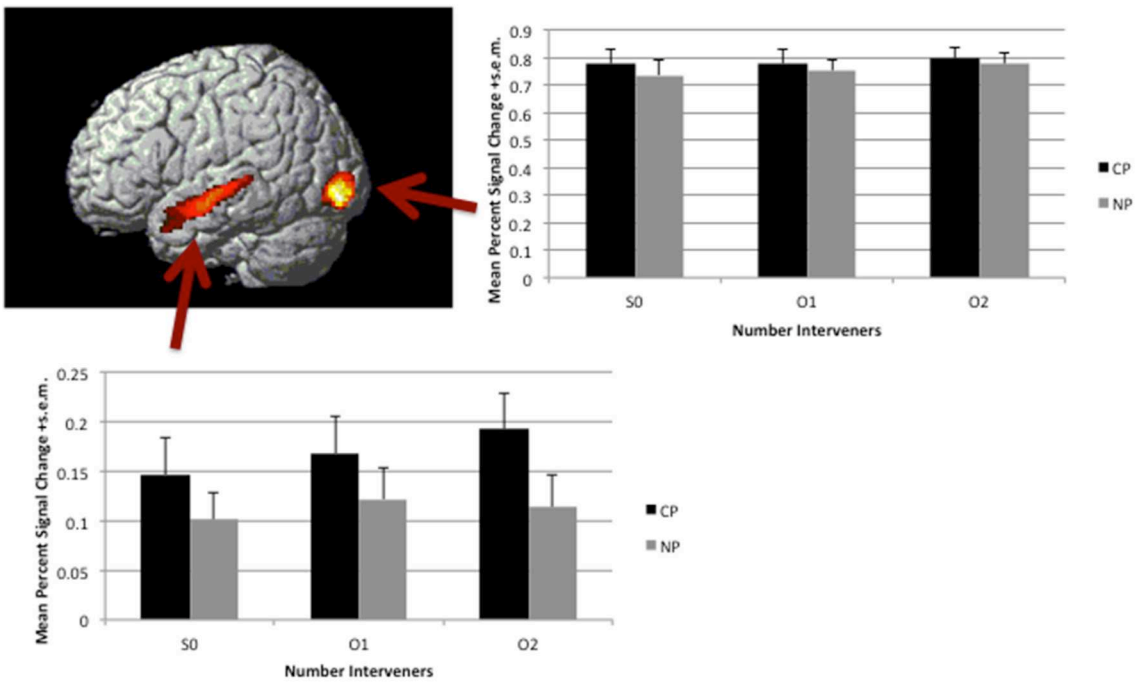


FIGURE 6 | Main effect of Type (CP>NP). Map is thresholded at voxel-wise $p < 0.005$ and corrected for multiple comparison through cluster size $p < 0.05$. Activation is overlaid on a rendered surface of the brain. The histogram to the right presents the mean percent signal change +s.e.m. over the entire Inferior Occipital Gyrus cluster broken down by condition (percent signal change was calculated in Marsbar per subject). The histogram below presents the mean percent signal change +s.e.m. over the entire Superior Temporal Sulcus cluster broken down by condition (percent signal change was calculated in Marsbar per subject).

type-identical interveners (NPs). This definition of distance not only holds of our results but also covers some other closely related studies (Santi and Grodzinsky, 2007b; Makuuchi et al., 2013). The cross-study consistency with Makuuchi et al. (2013) holds in terms of anatomical location but also in terms of experimental paradigm (reading word/phrase by word/phrase; comprehension questions assessing thematic role reversal) and methods for data analysis. The two studies differ in terms of the syntactic constructions tested. Makuuchi et al. (2013) investigated two types of movement dependencies in German, Scrambling and Topicalization. Here we studied embedded wh-movement. Similar to what is reported here, Makuuchi et al. (2013) found a linear effect of number of NP interveners on the fMRI signal with a peak in BA44, spreading into BA45.

A previous study by Santi and Grodzinsky (2007b) found the activation for increasing number of NPs within a movement dependency to be centered more anteriorly than the current study. This previous study differed from the current one in many ways. For one, intermittent scanning was used (thus the point of the scan may have been biased to BA45 processing, if BA45 has an earlier or later peak in processing relative to BA44), presentation modality was auditory, and the data analysis was slightly different (parametric effect was taken into consideration in the model). Nonetheless, these three studies are relatively similar and provide corroborating evidence for the role of Broca's area in being sensitive to movement distance defined over intervening constituents (i.e., NPs) that are similar to the head of the dependency.

Based on previous fMRI studies that find *unpredictable* syntactic dependencies (i.e., Reflexive Binding) do not activate Broca's area when NPs intervene, we suggest that the observed similarity-based interference effects are based on predictive processes. In particular, that there is storage of a prediction (NP gap) that is affected by similar, intervening NPs. The implication is that maintaining syntactic predictions increases activation in Broca's area and these predictions are affected by interveners that are identical in type. In the case of a movement relation, the parser is predicting a gap, which could involve storage of a category (i.e., NP) (Wagers and Phillips, 2014) or possibly an even more detailed feature profile (+sing, +animate, +nominal). When a *potential* gap site is reached this may cause reactivation of the entire lexical content of the filler or not (if maintained), but in either case the presence of another (type-identical) NP at a potential gap location will cause interference.

The behavioral data similarly shows that the number of intervening NPs affects both accuracy and RT. However, there is some indication that RT is primarily affected by CPs and not NPs (at least in the analysis by subjects but not the analysis by items). It is important to bear in mind that these are offline measures so how they directly relate to online measures of interference is more difficult to ascertain.

Our conclusions are consistent with Glaser et al. (2013) in showing that Broca's area is activated when there is interference by type-identical NP interveners in resolving syntactic dependencies. However, our conclusions differ in that Glaser et al. (2013) attribute this interference to occur during cue-based retrieval rather than along the prediction path. Remember that Glaser et al. (2013) do not investigate a movement dependency, but nonetheless one that is predictable, a subject-verb dependency (Van Dyke and McElree, 2006; Van Dyke, 2007). Trying to generalize across these two types of dependencies may not be the right approach. Further study is required to establish the degree of similarity in the memory mechanisms used to resolve these two dependencies and whether they are dependent on cue-based retrieval or maintenance of a prediction.

In general, the perspective that Broca's region engages in conflict resolution (Thompson-Schill et al., 1997; Novick et al., 2005; Thothathiri et al., 2012) is compatible with either interference during prediction or cue-based retrieval. In terms of prediction, it would attribute a conflict to wanting to release the filler as soon as possible and the actual representation, in which the "potential" gap site is already filled with a type identical NP. In resolving this conflict, the parser will need to maintain the gap prediction and do so for the correct NP. In terms of cue-based retrieval, this perspective would attribute conflict resolution to deciding between which of the type-identical NPs is the actual argument (head) of the verb or gap.

Lastly, it is worth noting that although about three-quarters of the activation lies within Broca's area (BA44/45), it also extends medially and posteriorly from Broca's area, including areas that connect Broca's area to other regions.

A Syntactic Working Memory

Given that the syntactic complexity effects observed in Broca's area depend on long distance dependencies, some have argued that its functional role is to provide a syntactic working memory (Caplan and Waters, 1999). The results from this study indicate that, if so, the size of the intervening structure is not as relevant as its similarity to the moved constituent. It appears that the critical dimension is the similarity of syntactic structure/features between the moved phrase and those intervening along the path of movement.

Phonological Working Memory

Some would argue that this linear effect of distance in Broca's are is related to a phonological working memory (Rogalsky et al., 2008) rather than syntactic/semantic similarity. However, contradictory evidence has been provided from various other studies (Caplan et al., 2000; Santi and Grodzinsky, 2007b). Even though Rogalsky and Hickok (2009) interpret their results to be due to phonological working memory, even their results demonstrate that there is activation in Broca's area during concurrent speech articulation. Thus, there is no clear evidence indicating that the observed syntactic complexity effect can be reduced to a phonological working memory.

Basal Ganglia, WM, and Syntactic Complexity

In addition to Broca's area (and its right homolog), a linear effect of interveners was observed in the basal ganglia. Makuuchi et al. (2013) also observed a linear effect of interveners within a movement dependency in the basal ganglia, however, there the activation was observed in the globus pallidus rather than the caudate nucleus. Further, a variety of related studies have found that the basal ganglia is sensitive to syntactic complexity (Prat and Just, 2011) and syntactic anomaly (Moro et al., 2001). Thus, this result is consistent with the region being engaged in the network that computes syntax.

Interaction Effect in MFG/SFG

The effect of syntactic size demonstrated an effect beyond type-identity interveners, bilaterally in the MFG/SFG, though most predominantly in the SFG. This was an unexpected finding, particularly with respect to the peak activation that lies anteriorly. The more posterior extent of the activation observed in the left hemisphere is similar to that seen in studies investigating the processing of Japanese scrambled sentences (Kinno et al., 2008) and one study that was interested in general distance effects within subject-verb agreement dependencies (Makuuchi et al., 2009). In fact, this posterior area is in very close proximity to the posterior end of the linear effect cluster (that is extending beyond Broca's area). Thus, this area seems to demonstrate some general engagement when Working Memory increases during sentence processing.

The bulk of the activation for the interaction, however, extends further anteriorly and is bilateral, thus demonstrating a distinction from these previous studies. Interpretations of the effect should, therefore, be made cautiously. Moreover, the plot of percent signal change by condition within this cluster provides a difficult picture to interpret. It seems as though there is activation for 1 intervening object in the double object construction, but no activation with 1 intervening object in the embedded clause structure. Then the pattern inverts for 2 intervening objects.

Auxiliary Brain Areas and Contrasts

Superior Temporal Cortex and CP>NP

In addition to examining the effect of distance (similarity and syntactic size) we looked at differences between the two Types of constructions. The results of this contrast need to be treated with care since many syntactic variables are concurrently manipulated (since this was not our primary interest). Although the CP and NP conditions contain the same number of NPs, the CP condition has an additional verb, whereas the NP condition has the preposition *to*. Furthermore, the argument structure of the verbs differ, the NP condition contains ditransitive verbs that the CP condition does not. Rather the CP condition contains more verbs that take sentential complements. Given these differences in verb argument structure, there is a consequent effect on the degree of syntactic structure building. The CP condition embeds clauses on each verb, thereby generating more syntactic structure. Having acknowledged the variety of differences across the TYPE contrast, the results of this contrast can be used to speak to current functional interpretations of the regions observed by this contrast that make reference to these syntactic variables. The anterior-to-posterior superior temporal gyrus activation is consistent with previous studies that have found the mid-to-superior posterior temporal gyrus sensitive to argument structure and syntax (Ben-Shachar et al., 2003, 2004; Friederici et al.,

2009; Santi and Grodzinsky, 2010) and the anterior temporal cortex to structure building (Humphries et al., 2005; Rogalsky and Hickok, 2009; Brennan et al., 2012). It is of interest to note, however, that the peak and focus of the activation is in the middle temporal cortex and not more anteriorly. In fact, in lowering the p-value ($p < 0.001$), the anterior temporal activation disappears and the mid-to-posterior activation remains. Thus, this contrast predominately depends on the mid-to-posterior superior temporal gyrus, rather than its anterior portion.

In looking at the percent signal change across conditions for this cluster, it is clear this effect is observed regardless of the number of intervening NPs. Additionally, it suggests a linear trend in activation with increasing number of NP interveners for the condition with multiply embedded CPs that is not observed for the condition with a ditransitive verb in a singly embedded CP. Similarly, the offline behavioral RT data demonstrate some evidence for increasing RTs with increasing number of interveners in the multiply embedded clause condition. This was observed in a marginally significant interaction effect (only in the by-subjects analysis, however). At appropriately thresholded levels, the fMRI data do not demonstrate such an interaction in the STS. However, when the voxel-wise p-value is dropped to $p < 0.05$ (voxel-wise and uncorrected for cluster size) the STS is observed in the interaction effect map. That is, a greater linear increase in activation with an increasing number of intervening NPs is observed in the multiply embedded clause condition over the single embedded clause one. Recall, in the two intervening object NP condition, the movement is occurring over a clausal boundary in the multiple clause condition, but is within a clause in the double object condition. Wagers and Phillips (2014) demonstrate that not all properties of the filler are maintained across a clause boundary, requiring their retrieval at the gap site. Thus the data, though not significant, show a trend for the STS to be sensitive to multiple movements or retrieval demands.

Inferior Occipital Gyrus and CP>NP

Not only was the superior temporal sulcus activated by the contrast in syntax type, but so was the inferior occipital gyrus. The location of activation is consistent with that observed by Makuuchi et al. (2013) in their contrast between Topicalization and Scrambling. The authors interpret this to be due to visual attention driven by the case-marked NP that appears sentence initial in wh-movement (topicalization), but in the embedded clause in scrambled structures. If the activation is due to increased attention, then in this study it must be for a different reason than the presence of an early case-marked NP that predicts the additional NPs. First there is no case-marking in the present study and, if anything, the filler appears earlier in the NP condition than the CP condition. A distinct potential syntactic factor that could be generating predictions and increasing visual attention in the CP compared to NP condition is that there are more open clausal phrases in the CP condition, leading to more predictions of verbs and arguments. Generally, these findings are consistent with other studies that demonstrated top-down effects in visual areas based on lexical predictions (Dikker and Pylkkänen, 2011).

Conclusions

The current parametric study manipulated the number of NP interveners in a movement dependency while also manipulating the presence of a clausal boundary across such a dependency. The results demonstrated a linear effect of number of interveners in Broca's area but no interaction between the number of interveners and the presence of a clausal boundary. More superiorly and bilaterally in the SFG there was an interaction due to the clausal boundary having a greater effect on the number of interveners than NPs. The STS demonstrated greater activation for the multiple clausal embedding condition than the single clausal embedding condition regardless of the number of interveners. As there were multiple distinctions across these conditions it is difficult to attribute the activation to a particular factor. Further, there was a trend within the STS in being sensitive to movement over a clause boundary, although not significant. In conclusion, Broca's area is sensitive to the number of interveners that are similar to the moved constituent and the activation is not augmented by an additional movement, or movement over a clausal boundary. Thus, type-identical interference rather than movement or crossing of a clausal boundary increases activation in Broca's area.

Acknowledgments

Partial support for this project was provided by an Alexander von Humboldt Foundation Research Award (YG), SSHRC (standard grand #410-2009-0431), Canada Research Chairs (YG), the Edmond and Lily Safra Center for Brain Research (ELSC, to YG), European Research Council (ERC-2010-360 AdG20100407 awarded to AF) and the German Ministry of Education and Research (BMBF; Grant Nr. 01GW0773 AF).

Supplementary Material

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpsyg.2015.00654/abstract

References

Amunts, K., Lenzen, M., Friederici, A. D., Schleicher, A., Morosan, P., Palomero-Gallagher, N., et al. (2010). Broca's region: novel organizational principles and multiple receptor mapping. *PLoS Biol.* 8:e1000489. doi: 10.1371/journal.pbio.1000489

Ben-Shachar, M., Hendler, T., Kahn, I., Ben-Bashat, D., and Grodzinsky, Y. (2003). The neural reality of syntactic transformations: evidence from functional magnetic resonance imaging. *Psychol. Sci.* 14, 433–440. doi: 10.1111/1467-9280.01459

Ben-Shachar, M., Palti, D., and Grodzinsky, Y. (2004). Neural correlates of syntactic movement: converging evidence from two fMRI experiments. *NeuroImage* 21, 1320–1336. doi: 10.1016/j.neuroimage.2003.11.027

Brennan, J., Nir, Y., Hasson, U., Malach, R., Heeger, D. J., and Pylkkanen, L. (2012). Syntactic complexity predicts anterior temporal activity during natural story listening. *Brain Lang.* 120, 163–173. doi: 10.1016/j.bandl.2010.04.002

Caplan, D., Alpert, N., and Waters, G. (1999). PET studies of syntactic processing with auditory sentence presentation. *NeuroImage* 9, 343–351. doi: 10.1006/nimg.1998.0412

Caplan, D., Alpert, N., Waters, G., and Olivieri, A. (2000). Activation of Broca's area by syntactic processing under conditions of concurrent articulation. *Hum. Brain Mapp.* 9, 65–71. doi: 10.1002/(SICI)1097-0193(200002)9:2<65::AID-HBM1>3.0.CO;2-4

Caplan, D., and Waters, G. S. (1999). Verbal working memory and sentence comprehension. *Behav. Brain Sci.* 22, 77–126. doi: 10.1017/S0140525X99001788

Carreiras, M., Pattamadilok, C., Mesequer, E., Barber, H., and Devlin, J. T. (2012). Broca's area plays a causal role in morphosyntactic processing. *Neuropsychologia* 50, 816–820. doi: 10.1016/j.neuropsychologia.2012.01.016

Daneman, M., and Carpenter, P. A. (1980). Individual differences in working memory and reading. *J. Verb. Learn. Verb. Behav.* 19, 450–466. doi: 10.1016/S0022-5371(80)90312-6

Dikker, S., and Pylkkänen, L. (2011). Before the N400: effects of lexical-semantic violations in visual cortex. *Brain Lang.* 118, 23–28. doi: 10.1016/j.bandl.2011.02.006

Eickhoff, S. B., Stephan, K. E., Molhberg, H., Grefkes, H., Fink, G. R., Amunts, K., et al. (2005). A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. *NeuroImage* 25, 1325–1335. doi: 10.1016/j.neuroimage.2004.12.034

Fiebach, C. J., Schlesewsky, M., Lohmann, G., von Cramon, D. Y., and Friederici, A. D. (2005). Revisiting the role of Broca's area in sentence processing: syntactic integration versus syntactic working memory. *Hum. Brain Mapp.* 24, 79–91. doi: 10.1002/hbm.20070

Friederici, A. D., Fiebach, C. J., Schlesewsky, M., Bornkessel, I. D., and Von Cramon, D. Y. (2006). Processing linguistic complexity and grammaticality in the left frontal cortex. *Cereb. Cortex* 16, 1709–1717. doi: 10.1093/cercor/bhj106

Friederici, A. D., Kotz, S. A., Scott, S. K., and Obleser, J. (2009). Disentangling syntax and intelligibility in auditory language comprehension. *Hum. Brain Mapp.* 31, 448–457. doi: 10.1002/hbm.20878

Gibson, E., and Warren, T. (2004). Reading-time evidence for intermediate linguistic structure in long-distance dependencies. *Syntax* 1, 55–78. doi: 10.1111/j.1368-0005.2004.00065.x

Glaser, Y. G., Martin, R. C., Van Dyke, J. A., Hamilton, A. C., and Tan, Y. (2013). Neural basis of semantic and syntactic interference in sentence comprehension. *Brain Lang.* 126, 314–326. doi: 10.1016/j.bandl.2013.06.006

Gordon, P. C., Hendrick, R., and Johnson, M. (2001). Memory interference during language processing. *J. Exp. Psychol.* 27, 1–13. doi: 10.1037/0278-7393.27.6.1411

Grewe, T., Bornkessel, I., Zysset, S., Wiese, R., Von Cramon, D. Y., and Schlesewsky, M. (2005). The emergence of the unmarked: a new perspective on the language-specific function of Broca's Area. *Hum. Brain Mapp.* 26, 178–190. doi: 10.1002/hbm.20154

Grodzinsky, Y. (2000). The neurology of syntax: language use without Broca's area. *Behav. Brain Sci.* 23, 1–21. doi: 10.1017/S0140525X00002399

Grodzinsky, Y., and Santi, A. (2008). The battle for Broca's region. *Trends Cogn. Sci.* 12, 474–480. doi: 10.1016/j.tics.2008.09.001

Henson, R. N. A., Buechel, C., Josephs, O., and Friston, K. J. (1999). The slice-timing problem in event-related fMRI. *NeuroImage* 9, 125.

Humphries, C., Love, T., Swinney, D., and Hickok, G. (2005). Response of anterior temporal cortex to syntactic and prosodic manipulations during sentence processing. *Hum. Brain Mapp.* 26, 128–138. doi: 10.1002/hbm.20148

Just, M. A., Carpenter, P. A., Keller, T. A., Eddy, W. F., and Thulborn, K. R. (1996). Brain activation modulated by sentence comprehension. *Science* 274, 114–116. doi: 10.1126/science.274.5284.114

Kinno, R., Kawamura, M., Shioda, S., and Sakai, K. (2008). Neural correlates of noncanonical syntactic processing revealed by a picture-sentence matching task. *Hum. Brain Mapp.* 29, 1015–1027. doi: 10.1002/hbm.20441

Makuuchi, M., Bahlmann, A., Anwander, A., and Friederici, A. D. (2009). Segregating the core computational faculty of human language from working memory. *Proc. Natl. Acad. Sci. U.S.A.* 106, 8362–8367. doi: 10.1073/pnas.0810928106

Makuuchi, M., Grodzinsky, Y., Amunts, K., Santi, A., and Friederici, A. D. (2013). Processing noncanonical sentences in Broca's region. *Cereb. Cortex* 23, 694–702. doi: 10.1093/cercor/bhs058

Moro, A., Tettamanti, M., Perani, D., Donati, C., Cappa, S. F., and Fazio, F. (2001). Syntax and the brain: disentangling grammar by selective anomalies. *Neuroimage* 13, 110–118. doi: 10.1006/nimg.2000.0668

Novick, J. M., Trueswell, J. C., and Thompson-Schill, S. L. (2005). Cognitive control and parsing: reexamining the role of Broca's area in sentence comprehension. *Cogn. Affect. Behav. Neurosci.* 5, 263–281. doi: 10.3758/CABN.5.3.263

Oldfield, R. C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia* 9, 97–113. doi: 10.1016/0028-3932(71)90067-4

Pallier, C., Devauchelle, A., and Dehaene, S. (2011). Cortical representation of the constituent structure of sentences. *Proc. Natl. Acad. Sci. U.S.A.* 108, 2522–2527. doi: 10.1073/pnas.1018711108

Prat, C. S., and Just, M. A. (2011). Exploring the neural dynamics underpinning individual differences in sentence comprehension. *Cereb. Cortex* 21, 1747–1760. doi: 10.1093/cercor/bhq241

Rogalsky, C., and Hickok, G. (2009). Selective attention to semantic and syntactic features modulates sentence processing networks in anterior temporal cortex. *Cereb. Cortex* 19, 786–796. doi: 10.1093/cercor/bhn126

Rogalsky, C., Matchin, W., and Hickock, G. (2008). Broca's area, sentence comprehension, and working memory: an fMRI study. *Front. Hum. Neurosci.* 2:14. doi: 10.3389/neuro.09.014.2008

Santi, A., and Grodzinsky, Y. (2007a). Taxing working memory with syntax: bihemispheric modulations. *Hum. Brain Mapp.* 28, 1089–1097. doi: 10.1002/hbm.20329

Santi, A., and Grodzinsky, Y. (2007b). Working memory and syntax interact in Broca's area. *Neuroimage* 37, 8–17. doi: 10.1016/j.neuroimage.2007.04.047

Santi, A., and Grodzinsky, Y. (2010). fMRI adaptation dissociates syntactic complexity dimensions. *NeuroImage* 51, 1285–1293. doi: 10.1016/j.neuroimage.2010.03.034

Stromswold, K., Caplan, D., Alpert, N., and Rauch, S. (1996). Localization of syntactic comprehension by positron emission tomography. *Brain Lang.* 52, 452–473. doi: 10.1006/brln.1996.0024

Thiel, A., Habedank, B., Herlolz, K., Kessler, J., Winhuisen, L., Haupt, W. F., et al. (2006). From the left to the right: how the brain compensates progressive loss of language function. *Brain Lang.* 98, 57–65. doi: 10.1016/j.bandl.2006.01.007

Thompson-Schill, S. L., D'Esposito, M., Aguirre, G. K., and Farah, M. J. (1997). Role of left inferior prefrontal cortex in retrieval of semantic knowledge: a reevaluation. *Proc. Natl. Acad. Sci. U.S.A.* 94, 1492–14797. doi: 10.1073/pnas.94.26.14792

Thornton, R. (1995). Referentiality and wh-movement in child English: Juvenile D-Linkuency. *Lang. Acquisition* 4, 139–175. doi: 10.1080/10489223.1995.9671662

Thothathiri, M., Kim, A., Trueswell, J. C., and Thompson-Schill, S. L. (2012). Parametric effects of syntactic-semantic conflict in Broca's area during sentence processing. *Brain Lang.* 120, 259–264. doi: 10.1016/j.bandl.2011.12.004

Van Dyke, J. A. (2007). Interference effects from grammatically unavailable constituents during sentence processing. *J. Exp. Psychol.* 33, 407–430. doi: 10.1037/0278-7393.33.2.407

Van Dyke, J. A., and McElree, B. (2006). Retrieval interference in sentence comprehension. *J. Mem. Lang.* 55, 157–166. doi: 10.1016/j.jml.2006.03.007

Wagers, M. W., and Phillips, C. (2014). Going the distance: memory and control processes in active dependency construction. *Q. J. Exp. Psychol.* 67, 1274–1304. doi: 10.1080/17470218.2013.858363

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2015 Santi, Friederici, Makuuchi and Grodzinsky. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 23 June 2015
doi: 10.3389/fpsyg.2015.00860

Discourse accessibility constraints in children's processing of object relative clauses

Yair Haendler[1], Reinhold Kliegl[2] and Flavia Adani[1]*

[1] Department of Linguistics, University of Potsdam, Potsdam, Germany, [2] Department of Psychology, University of Potsdam, Potsdam, Germany

Children's poor performance on object relative clauses has been explained in terms of intervention locality. This approach predicts that object relatives with a full DP head and an embedded pronominal subject are easier than object relatives in which both the head noun and the embedded subject are full DPs. This prediction is shared by other accounts formulated to explain processing mechanisms. We conducted a visual-world study designed to test the off-line comprehension and on-line processing of object relatives in German-speaking 5-year-olds. Children were tested on three types of object relatives, all having a full DP head noun and differing with respect to the type of nominal phrase that appeared in the embedded subject position: another full DP, a 1st- or a 3rd-person pronoun. Grammatical skills and memory capacity were also assessed in order to see whether and how they affect children's performance. Most accurately processed were object relatives with 1st-person pronoun, independently of children's language and memory skills. Performance on object relatives with two full DPs was overall more accurate than on object relatives with 3rd-person pronoun. In the former condition, children with stronger grammatical skills accurately processed the structure and their memory abilities determined how fast they were; in the latter condition, children only processed accurately the structure if they were strong both in their grammatical skills and in their memory capacity. The results are discussed in the light of accounts that predict different pronoun effects like the ones we find, which depend on the referential properties of the pronouns. We then discuss which role language and memory abilities might have in processing object relatives with various embedded nominal phrases.

Keywords: child language, relative clauses, discourse, pronouns, intervention locality, visual-world paradigm

OPEN ACCESS

Edited by:
Colin Phillips, University of Maryland, USA

Reviewed by:
Cynthia Thompson, Northwestern University, USA
Akira Omaki, Johns Hopkins University, USA

****Correspondence:***
Yair Haendler, Department of Linguistics, University of Potsdam, Karl-Liebknecht-Strasse 24-25, 14476 Potsdam, Germany
yair.haendler@uni-potsdam.de

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: *08 January 2015*
Accepted: *11 June 2015*
Published: *23 June 2015*

Citation:
Haendler Y, Kliegl R and Adani F (2015) Discourse accessibility constraints in children's processing of object relative clauses. Front. Psychol. 6:860. doi: 10.3389/fpsyg.2015.00860

Introduction

Relative Clause Processing in Children and Adults

The acquisition of relative clauses has been studied extensively and in a large variety of languages (Brandt et al., 2009; Arnon, 2010; Adani, 2011; Arosio et al., 2012; Belletti et al., 2012; Adani et al., 2014, among others). The existing research focuses mainly on the asymmetry between child performance on subject-extracted relatives (SRs) and object-extracted relatives (ORs), examples of which are provided in (1) and (2), respectively. In the examples, the head of the relative clause is the noun it modifies (*the bunny*). The underscore marks the position in the embedded clause from which the head noun is extracted: subject position in SRs and object position in ORs.

(1) The bunny that __ is chasing the horse
(2) The bunny that the horse is chasing __

In head-initial languages, it is a robustly attested finding that young children have difficulties comprehending and producing ORs, but not SRs (see Gutierrez-Mangado, 2011 for a reversed pattern in Basque). Children's errors with ORs are mainly expressed by the interpretation of these sentences as SRs. An account that aims to explain the SR–OR asymmetry in acquisition is proposed by Friedmann et al. (2009), following earlier work by Grillo (2005, 2009). This approach provides an explanation in terms of intervention locality, based on the syntactic principle of Relativized Minimality (RM; Rizzi, 1990 and subsequent work). We will refer to Friedmann et al.'s (2009) approach as the RM account.

Relativized Minimality is based on the configuration in (3), in which X is a constituent that moves from its original (gap) position Y crossing an intervening constituent Z.

(3) X . . . Z . . . Y

According to the RM Principle, a local relation between X and Y is impossible if Z is a potential candidate for that local relation. Such a case occurs when Z intervenes between X and Y and when Z is structurally similar to X. These two co-occurring conditions give rise to a locality intervention effect and, thus, to difficulties in parsing the structure. Friedmann et al. (2009) show how this configuration and the conditions that create intervention effects apply to the structure of SRs and ORs[1]. In the case of relative clauses, the authors identify the feature [+NP], or 'lexical restriction,' as the one that, when present on both X and Z, makes them structurally similar. In (1) and (2), repeated as (4) and (5), both X and Z are lexically restricted, or in other words: they are both full DPs. But only in the OR Z intervenes between X and Y. For this reason, according to Friedmann et al. (2009) ORs with two full DPs are difficult for children whereas SRs with two full DPs are not.

(4) [The bunny] that __ is chasing [the horse]
X Y Z

(5) [The bunny] that [the horse] is chasing __
X Z Y

The RM account predicts significant improvement in child comprehension of ORs when the head (X) is a full DP, whereas the embedded subject (Z) is not. Children are therefore predicted to perform more accurately on an OR with a full DP head and an embedded subject which is a personal pronoun, a DP that lacks the [+NP] feature. Friedmann et al. (2009, p. 75) tested this prediction examining child comprehension of Hebrew ORs with an embedded subject which is a null pronoun. The following example is taken from their paper.

(6) Tare li et ha-sus she- mesarkim oto
show to-me ACC the-horse that- *pro* brush-3rd-pl him
'Show me the horse that someone is brushing'
(literally, 'the horse that they are brushing')

The Hebrew *pro* subject in (6) is an impersonal subject that agrees with the 3rd-person plural form, as evidenced by the Person and Number agreement marking on the embedded verb *brush*. This impersonal, or arbitrary *pro* is used to describe the action of an unspecified agent. Friedmann et al. (2009) found that children understood ORs like (6) more accurately than ORs with a full DP head noun and a full DP embedded subject. They explained the improved comprehension as due to the attenuation of the intervention locality effect, caused by the fact that the head of the OR is a full DP but not its embedded pronominal subject. Crucially, the prediction is that any type of pronoun in the embedded subject position will improve comprehension, since what matters is the lack of lexical restriction, a property shared by all personal pronouns. This prediction receives further support from studies that find relatively accurate child performance on ORs whose embedded subject is an overt 3rd-person pronoun (Brandt et al., 2009), a 2nd-person pronoun (Kidd et al., 2007) or a 1st-person pronoun (Arnon, 2010).

Other accounts that explain OR processing based on adult performance make similar predictions. Warren and Gibson (2002, 2005) propose that sentence processing is determined by the number of new referents that intervene between a moved element (filler) and the gap site in which it is integrated into the structure. The greater the number of intervening referents (e.g., noun phrases, verbs) the harder it is to keep track of the filler until the gap site is encountered and the filler-gap dependency is resolved [a similar idea is advanced by O'Grady (2011)]. Under this view, an intervening pronoun reduces processing cost since it does not introduce a new discourse referent: it serves as a link to an already given one. Indeed, adults have less difficulty with doubly nested ORs and object clefts whose embedded-most DP is a pronoun, as compared to cases in which all the nominal phrases in the structure are full DPs (Warren and Gibson, 2002, 2005). Other accounts explain the difficulty with ORs in terms of similarity between the DP head and the embedded subject DP. It has been found that an OR becomes easier to parse when these two constituents are sufficiently dissimilar. For instance, ORs with two full DPs are more costly to process than ORs in which the head is a full DP and the embedded subject is a proper name (Gordon et al., 2004), or a 2nd-person pronoun (Gordon et al., 2001). Other studies define the difficulties with OR processing in terms of cue-based interference (Lewis and Vasishth, 2005; Lewis et al., 2006; van Dyke and McElree, 2006). Under this view, the similarity between the DP head and the embedded DP is defined by the cues that these two constituents bear. When a constituent (e.g., the DP head in an OR) is encountered it is encoded in memory. Later on, in the gap position, it has to

[1]The RM principle was first developed to explain intervention locality effects in extraction from weak islands (Rizzi, 1990). The approach was later extended to explain intervention effects in ORs, assuming a structural proximity between the latter and the original island phenomena (Grillo, 2005, 2009; Friedmann et al., 2009; Rizzi, 2013).

be retrieved from memory in order to be integrated into the structure. At this point, its (syntactic, semantic, or other) cues are analyzed in order to decide whether the filler-gap dependency can be resolved. If another constituent (e.g., the embedded subject DP in an OR) shares similar cues with those of the encoded constituent this second set of cues will interfere with the processing of the first one, increasing the overall processing cost of the structure. In an OR with an embedded pronoun, the cues of the intervening pronoun are sufficiently different from those of the encoded head noun, thus reducing the processing cost.

As can be seen, there is an affinity between the RM account and the accounts reviewed in the last paragraph, although the former is the only one whose predictions have been tested in experiments with children. All these accounts appear to share the prediction that an OR with an embedded pronominal subject is less costly for processing than an OR in which both the head noun and the embedded subject are full DPs. Moreover, at least some of these approaches (Gordon et al., 2001; Lewis et al., 2006), like the RM account, attribute the difficulties in OR processing to the (dis)similarity between the DP head and the embedded subject DP in terms of cues or features. Importantly, however, each of these studies tested the effect of only one pronoun type on OR processing. The only exception is Warren and Gibson's (2002) study with adults, to which we will return later. The present study is the first to assess the comprehension of ORs with different embedded pronominal subjects in children. That is, we will test the prediction that ORs with different pronouns in the embedded subject position should be equally easy for children, as compared to ORs with two full DPs. Comparing the effects of different pronoun types is particularly interesting given studies that show that pronouns with different referential properties affect sentence processing differently in adults (Warren and Gibson, 2002; Carminati, 2005).

We have recently shown (Haendler et al., 2015) that there is a relation between children's performance on ORs with different types of embedded referring expressions (full DP, different personal pronouns)[2] and their language skills, as measured by standardized tests for receptive grammatical abilities. These language or grammatical skills (we will use the two terms interchangeably) were defined as the average score on three subtests from Siegmüller et al. (2010). The tests assessed the comprehension of (a) canonical and non-canonical declarative sentences (SVO and OVS); (b) sentences containing reflexives and pronouns; (c) various types of relative clauses (right-branching and center-embedded; SRs and ORs). In the discussion, we will elaborate on what grammatical skills are assumed to underlie children's performance on these three language tests. Concerning the results, we found that children were most accurate on ORs with an embedded 1st-person pronoun (OR + 1pro; *The horse that I chase*), independently of their scores on the language tests. In ORs with an embedded 3rd-person pronoun (OR + 3pro; *The horse that it chases*) and ORs with a full DP head and an embedded full DP (OR + 2DP; *The horse that the bunny chases*), which were overall more difficult, children's performance interacted with their grammatical skills: children with higher scores on the language tests were more accurate on these conditions than children with lower scores.

In the present paper, we extend this picture by looking at memory skills and assessing whether they interact with language abilities in the modulation of children's performance on the three OR types. In other words, we want to see whether both language and memory have an impact on children's OR processing, and whether their effects are independent of one another or whether they interact. In the latter case, we want to see what kind of relation between language and memory skills emerges during OR processing. This kind of analysis will help distinguish between effects that are purely due to children's language skills, effects that are purely memory-dependent and effects that are caused by both types of cognitive abilities.

Memory and the Processing of Object Relative Clauses

The relevance of memory for the processing of relative clauses has been vastly investigated. To begin with, Friedmann et al. (2009) speculate that the difficulty with an OR containing two full DPs lies in children's limited memory capacity. During the processing of such a structure, one needs to hold in memory the featural specifications of the DP head and the embedded DP and compare them in order to determine their (dis)similarity (see also Adani et al., 2010). When the features of the DP head and of the embedded DP are similar, such as when they are both full DPs, the comparison of the features is more costly and memory capacity is overloaded. However, when the features on the DP head and on the embedded DP are sufficiently different, as in the case of an OR with an embedded pronominal subject, comparing the features becomes less demanding for memory resources and the comprehension of the OR is facilitated.

The reviewed accounts on adult processing similarly suggest that memory abilities constrain the processing of ORs (for a comprehensive review, see Wagers and Phillips, 2014). According to Gibson (1998, 2000) and Warren and Gibson (2002, 2005; see also O'Grady, 2011), the difficulty associated with keeping track of the filler while processing newly introduced discourse referents is related to available memory resources. The greater the number of new discourse referents that intervene between the filler and its gap site, the longer the filler has to be kept in memory until the filler-gap dependency is resolved. Therefore, people with strong memory capacity will be facilitated in maintaining the filler in memory while processing the sentence until the gap position is reached. Gordon et al.'s (2001, 2004) proposal that the processing cost of an OR is determined by the (dis)similarity between the DP head and the embedded DP is also related to memory capacity. The idea is that dissimilar DPs burden memory to a lesser extent, making the distinction of the two constituents during sentence processing easier. Finally, the processing mechanism assumed under the cue-based interference account (Lewis and Vasishth, 2005; Lewis et al., 2006; van Dyke and McElree, 2006) similarly draws on memory resources. If the set of cues of a previously encoded constituent (the DP head of an OR) and that of the

[2]We use the term 'referring expression' to mean any linguistic form that relates to some discourse referent. This term thus includes both definite noun phrases (full DPs) and pronouns (see Fukumura and van Gompel, 2012; Serratrice, 2013).

intervening DP are similar, memory capacity will be overloaded, resulting in an increased processing cost. If the two sets of cues are dissimilar, memory resources will be less burdened and the sentence will be easier to process.

The relation between children's memory abilities and their comprehension of syntactically complex sentences has been vastly studied. Different studies have used different kinds of tests to measure memory, yielding mixed results. Some studies found a relation between children's off-line response accuracy and their performance on listening span tasks (Montgomery et al., 2008; Montgomery and Evans, 2009; Weighall and Altmann, 2011), backward digit span tasks (Engel de Abreu et al., 2011; Boyle et al., 2013) and forward digit span tasks (Arosio et al., 2011, 2012; Engel de Abreu et al., 2011). An association has been found also between similar memory tasks and children's on-line sentence processing (Booth et al., 2000; Roberts et al., 2007). However, no systematic relation has been found between the score on any specific memory test and children's performance on any specific language task (Kidd, 2013). Particularly relevant for the present study is Arosio et al.'s (2012) work. Using a picture-selection task, they tested 7-years-old German-speaking children on the comprehension of SRs and ORs, disambiguated either by case-marking on the determiner of the embedded DP or by number-marking on the embedded verb. The authors found that children were more accurate on case-disambiguated than on number-disambiguated ORs. Also relevant is their finding that children's score on a forward span test was a reliable predictor of their comprehension of ORs (but not SRs).

In the present study, we administered to children both a forward and a backward digit span task. The memory measure was calculated as the average score on the two tests. As we have seen, both the forward and the backward span tests have been widely used in studies with children. Moreover, these tasks are typically assumed to reflect two kinds of memory components in Baddeley's classical model (Baddeley, 1986; Baddeley et al., 2009): the forward digit span task is believed to reflect the operation of the phonological loop, a short-term storage of phonological information; the backward digit span task is assumed to reflect the operation of the central executive, which is responsible for the coordination and elaboration of the stored information. The former is often referred to as *verbal short-term memory*; the latter as *verbal working memory* (Kidd, 2013). The fact that no systematic relation has been demonstrated between any of these two tests and a specific performance pattern on language comprehension led us to combine the scores on the two tasks into one, more general measure of memory capacity. The disadvantage in doing so is that we cannot look at separate effects caused by the two kinds of memory abilities (short-term memory and working memory). The advantage is that such a general memory measure is more robust and reliable for the analysis, since it combines data collected in two different tasks. The mixed findings in the literature regarding the relation between the two span tasks and certain language abilities leaves the qualitative analysis of the role of memory highly speculative. Hence, by using the composite score, we gain a stronger measure for the quantitative analysis of children's memory capacity.

Referential Properties and Discourse Accessibility

As we have seen, the prediction we are testing is that any type of embedded pronoun should facilitate children's performance on ORs to an equal extent. However, there is extensive literature focusing on differences between pronouns in terms of their referential properties. A case in point is the different way of establishing reference of 1st- and 2nd-person pronouns on the one hand, and 3rd-person pronouns on the other hand. When a participant in a linguistic act constructs a discourse model, 1st- and 2nd-person pronouns are directly integrated into that model since they refer, respectively, to the speaker and the interlocutor, two discourse referents which are always available and highly accessible (Recanati, 1993; Erteschik-Shir, 1997; Ariel, 2001). Moreover, the referents of these pronouns are derived from the lexical meaning of the pronouns themselves: 1st-person pronoun ('I,' 'we') = speaker; 2nd-person pronoun ('you') = interlocutor. This is similar to the way in which a regular noun phrase (e.g., 'the horse') establishes reference. The discourse referent of the noun phrase is derived from its lexical meaning, despite the fact that it is marked with 3rd-person (unlike 1st- and 2nd-person pronouns) and although it is not referring to a participant in the linguistic act (like 'speaker' or 'interlocutor'). By contrast, the referent of a 3rd-person pronoun ('it,' 'they,' and demonstratives such as 'this,' 'that') is derived from the discourse, in a process of pronoun resolution in which the pronoun relates to an antecedent in the linguistic or extra-linguistic context (Heim, 1991; Legendre and Smolensky, 2012).

There is experimental evidence that such differences in discourse accessibility of pronouns affect the processing of sentences in which they occur. Warren and Gibson (2002) found that adults perceive doubly nested ORs with an embedded 1st- or 2nd-person pronoun as less complex, as compared to such structures with an embedded 3rd-person pronoun. Moreover, adult on-line processing of pronoun resolution in infrequent circumstances (when the pronoun antecedent is a previously mentioned object, rather than subject) is facilitated when that pronoun is marked with 1st- or 2nd-person, rather than 3rd-person (Carminati, 2005). These effects, assumed to be caused by the referential properties of pronouns, have not been tested yet in children. But a number of studies suggest children are sensitive to discourse properties of pronouns as well. First, in line with the pronoun asymmetry described above, children acquire the ability to correctly interpret 1st- and 2nd-person pronouns before 3rd-person pronouns (Brener, 1983; Girouard et al., 1997; Legendre et al., 2011; Legendre and Smolensky, 2012). Moreover, there is substantial evidence indicating that children are sensitive to the discourse properties that determine pronoun usage and interpretation (Song and Fisher, 2005, 2007; Spenader et al., 2009; Pyykkönen et al., 2010; Koster et al., 2011; Hartshorne et al., 2015)[3]. For instance, Song and Fisher (2005) found that

[3]Some studies have tested children's comprehension of intra-sentential anaphora. These are sentences in which the referent of the pronoun is inside the same sentence in which the pronoun appears (e.g., Sekerina et al., 2004; van Rij et al., 2010; Clackson et al., 2011). Here we concentrate only on extra-sentential anaphora, where the referent of the pronoun is in the extra-sentential or extra-linguistic (visual) context. This is the relevant case for the present study.

3-year-olds, tested with a preferential-looking paradigm, looked more to the correct referent figure of a pronoun when it was made prominent in the discourse (in the preceding context it was the first-mentioned figure in a subject position and pronominalized once), than when the referent was not prominent. Children in Koster et al.'s (2011) study interpreted the pronoun as referring to the first-mentioned character in a context story, both when this character was consistently the discourse topic and when there was a shift in the topic of the story. Production studies also suggest that children are sensitive to referential properties of pronouns, as well as to the extra-sentential or extra-linguistic context, when they choose which referring expression to utter (see Serratrice, 2013 and references therein). Together, these studies suggest that, from early on, children are sensitive to discourse properties of pronouns such as topicality or order-of-mention. It appears that children can use these properties in order to construct a plausible discourse model and, based on that model, derive expectations regarding the usage of the referring expressions they encounter in the linguistic input (see a related discussion in Trueswell et al., 2011).

According to Goodluck (2010), who discusses data in contradiction with Friedmann et al.'s (2009) approach, children's performance on complex structures is determined by both syntactic and discourse accessibility operations (see also Goodluck, 1990, 2005 and Avrutin, 2000). Whereas the RM account predicts difficulties with object-extracted wh-questions in which both the moved constituent and the intervening one are full DPs (*Which lion did the zebra kick?*), Goodluck (2005) found that children perform more accurately when the moved constituent is a more generic name (*Which animal did the zebra kick?*). In explaining the data, Goodluck suggests that children's difficulty with object *which*-questions is related both to the syntactic factor of distance (*which lion/animal* is extracted from the more distant position as the object of the verb *kick*) and to the discourse factor of set-restriction (to interpret *which lion*, the child has to restrict the set of given lions and understand which one she is asked about; this operation is less costly when *lion* is replaced with the more generic *animal*). Although Goodluck's (2010, p. 1520) proposal is made in relation to structures that are slightly different from the ones dealt with here, the relevance of her work lies in the idea that "[. . .] children appear to have difficulty in general with grammatical phenomena that require access to discourse."

The Present Study

To summarize the goal of the present study, we test the prediction that ORs with different embedded pronominal subjects are easier than ORs with two full DPs. Moreover, no difference is predicted between the conditions with pronouns. We used right-branching ORs with various referring expressions in the embedded subject position. ORs with an embedded 1st-person pronoun (7) and with 3rd-person pronoun (8) were compared to a baseline condition of ORs in which both the head noun and the embedded subject are full DPs (9)[4]. Note that these ORs differ with respect to the referring expression that occupies the embedded subject position (in bold). Hence, we expect differences in performance on the ORs to reflect effects caused by these referring expressions.

(7) OR + 1pro: Welche Farbe hat der Hase, den **ich** jage?
what color has the bunny who **I** chase

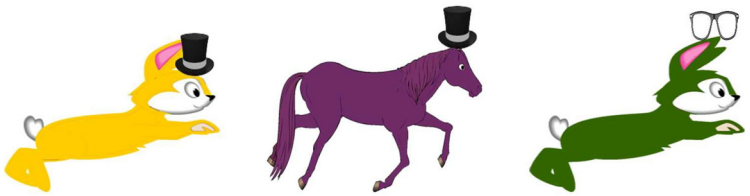
(8) OR + 3pro: Welche Farbe hat der Hase, den **es** jagt?
what color has the bunny who **it** chases

(9) OR + 2DP: Welche Farbe hat der Hase, den **das Pferd** jagt?
what color has the bunny who **the horse** chases

Previous studies on children's OR comprehension have used only off-line methods. Here, we designed a visual-world experiment (Tanenhaus et al., 1995) and measured both off-line response accuracy and on-line eye-gaze during the inspection of a visual scene that accompanied each test sentence. The off-line accuracy was collected as a measure of explicit comprehension; the on-line eye-gaze as a measure of implicit parsing strategies. Many studies using on-line measures (e.g., eye-tracking) have found evidence for early processing of complex structures and/or a more fine-grained performance pattern that usually remains hidden in the explicit response (Brandt-Kobele and Höhle, 2010; Adani and Fritzsche, 2015). Thus, on-line gaze measures are arguably more sensitive in testing child language, yielding results that suggest that children might implicity process a structure accurately even when their explicit response is inaccurate. For this reason, and since previous studies have found difficulties with ORs that persist until late in development (e.g., Friedmann et al., 2009; Arosio et al., 2012; Adani et al., 2014), we tested children at age 5. If the on-line eye-gaze measure is indeed more sensitive than the off-line response accuracy we might find evidence for correct processing of the harder condition(s) even as early as this age.

Let us now summarize the predictions regarding children's performance on the three conditions and the possible relation to language and memory abilities. The initial prediction is that children will be more accurate on OR + 1pro and OR + 3pro than on OR + 2DP, and there should be no difference between performance on OR + 1pro and OR + 3pro. However, if the different ways with which the 1st- and the 3rd-person pronouns establish reference influence children's performance, as found with adults (Warren and Gibson, 2002; Carminati, 2005), children should be more accurate on OR + 1pro than on OR + 3pro. We have already mentioned that stronger grammatical skills improve children's performance on two of the conditions. Given previous studies (Kidd, 2013), we might expect to find also an impact of memory that shows that stronger memory capacity improves performance on the task.

[4]In addition to these three conditions, we also tested a fourth condition in which the head noun was a demonstrative pronoun and the embedded subject was a full DP (*Welche Farbe hat der, den das Pferd jagt?* 'What color has that (the one) that the horse is chasing?'). The predictions regarding this condition are not straightforward, since existing literature is not explicit about whether such a demonstrative bears the [+NP] feature or not. Moreover, unlike this condition, all the others differed minimally by the referring expression in the embedded subject position. Upon suggestion from the two reviewers, we will neither present nor discuss the data from this condition.



Preamble: *Hier sind zwei Hasen. Der eine Hase ist gelb, und der andere Hase ist grün. Und hier ist ihr Freund, das Pferd. Das Pferd ist lila* ('Here are two bunnies. The one bunny is yellow and the other bunny is green. And here is their friend, the horse. The horse is purple').

Test sentence: *Guck mal! Welche Farbe hat der Hase, den das Pferd jagt?* ('Look! What color has the bunny who the horse chases?')

FIGURE 1 | Example of a visual scene, a preamble and a test sentence.

We might also find that language and memory abilities modulate children's performance differently. This would result in different patterns of interaction between language/memory and response accuracy/eye-gaze.

Regarding the specific pattern expected in the two kinds of data we have collected, a higher proportion of correct responses (i.e., naming the color of the correct figure) will express a more accurate off-line performance. With respect to the eye-gaze data, there are several possibilities. We measure the proportion of looks to the target figure in the visual scene that accompanies each test sentence, within a time window defined in advance for the analysis. Accurate processing of the sentence within the analysis window will be expressed either by earlier looks to the target figure, or by longer looks to the target (higher proportion of target looks), or both. Therefore, the initial predictions regarding the performance pattern in the accuracy data and the eye-gaze data roughly correspond. However, we might find evidence for correct processing of the sentences, or a more fine-grained performance pattern, only in the eye-gaze data.

Materials and Methods

Participants

Forty-seven 5-years-old children (24 females, age range 5.0–5.11, $M = 5.5$) participated in the study. All children are growing up as monolingual speakers of German and none has reported history of linguistic, hearing or other cognitive developmental disorders. Parents gave their consent for the participation of their children. The study, approved by the ethics commission of the University of Potsdam, was successfully piloted with a group of university students.

Material

Visual Stimuli

In a setup inspired by Arnon (2005, 2010) and Adani (2011) participants watched in each trial an animated video with two identical animals on the sides (target and distractor animals) and a third different animal in the middle (middle animal). Each of these three regions of interest had the same size of 436 × 400 pixels. An example of a visual scene is provided in **Figure 1**. Employing two verbs, 'chase' and 'tickle,' the three animals in the scene were chasing each other on half of the trials and tickling one another with a feather on the other half. Each of the animals in the scene was colored differently. The three colors were combined such that similar colors did not appear within the same video, in order to facilitate color distinction and recognition (Pitchford and Mullen, 2003). Each of the animals carried a small object (hat, glasses, flower or heart–all clip art images) that was relevant for the fillers, but not for the experimental items. The target animal (i.e., the referent of the OR head noun) could be one of four masculine nouns–bear, bunny, lion, or monkey–each of which appeared an equal number of times as target, and in a balanced manner across conditions. The middle animal was on some trials a neuter noun (horse, camel, zebra, or sheep) and on others a feminine noun (duck, cow, cat, or mouse). In the OR + 1pro condition, the middle animal was always the dog, established as referent for the 1st-person pronoun in an introduction story prior to the experiment (see Procedure). The direction of the scene was in half of the trials from left to right and in the other half from right to left. Depending on the action direction, the target animal was always either on the left or on the right side of the scene, but never in the middle. In the ORs, the target animal was always the last animal in the row; in the fillers, it was always the first animal in the row, to prevent participants from anticipating the side on which the target appeared.

Linguistic Stimuli

The design consisted of three experimental conditions [examples (7)–(9) in the Introduction], with seven trials in each condition, and 12 fillers (e.g., *Welche Farbe hat der Hase mit dem Hut?* 'What color is the bunny with the hat?'). Piloting the experiment before the actual testing revealed that, with this amount of items, the duration of the experiment (∼20 min) was adequate for 5-year-olds. The displayed videos were accompanied by the test sentences that were pre-recorded with a female German native speaker and integrated into the video file. These were questions about the color of one animal in the scene to be identified through a relative clause (in experimental items) or a small object (in fillers). Two lists were constructed, each containing a different pseudo-randomized order of the items. Half of the participants were exposed to the first list, and the other half were exposed

to the second list. The full list of items is provided in the online supplementary material.

Since all the target animals (i.e., the OR head noun) were singular masculine nouns, the relative pronoun in all the ORs was always unambiguously accusative case-marked (*den*, 'who$_{_ACC_MASC}$'). This way, the sentence is revealed to be an OR already upon encountering the relative pronoun and children might be facilitated in processing the sentence (Arosio et al., 2012). However, in order for children to be able to make use of this information, they have to be able to recognize the accusative case-marking on the relative pronoun. In particular, they have to be able to distinguish the accusative case-marked *den* from the nominative case-marked *der*. If children cannot tell apart the two minimally differing case-markings they might erroneously understand the sentence as a SR (e.g., *Welche Farbe hat der Hase, der das Pferd jagt?* 'What color has the bunny who$_{_NOM_MASC}$ the horse chases?'). This might mask the comprehension difficulties children typically have with the syntactic structure of the OR as such. In order to determine whether children were able to discern between the two case-markings, we looked at their performance on one of the language tests that were administered (from the TSVK battery, Siegmüller et al., 2010): the test on the comprehension of OVS sentences, which are grammatical but non-canonical in German. Successful performance on this test requires the distinction between nominative (*der*), accusative (*den*) and dative case-marking (*dem*), in order to understand that the pre-verbal noun is an accusative- or dative-marked object and that the post-verbal noun is a nominative-marked subject. When looking at the performance on this test it appears that 37 out of 41 children scored at or above 50% (answering correctly six or more out of the 12 questions in the test). Scatterplots showing the relation between individual performance on this test and the overall performance in the experiment (both in terms of off-line accuracy and on-line eye-gaze) are provided in the online supplementary material. Additional evidence that children in our study were able to tell apart nominative and accusative case-marking stems from independent studies that show that children as old as 4.6 can already distinguish nominative and accusative case-marking in German (Grünloh et al., 2011)[5].

Memory

We administered to the children a forward span test and a backward version of the same test. The sequences for the forward span test were taken from the Intelligence and Development Scales battery (Grob et al., 2009). The forward span test was used to measure verbal short-term memory. To measure verbal working memory, we used the same sequences in a backward span test which is typically taken to measure this type of memory capacity. The sequences in the two memory tasks were of increasing length, ranging from 2 to 7 items in each sequence, and containing either digits or letters (for instance, 5-3-8 or C-O-G). For each sequence length (of two items, three items, and so on) there was one sequence of digits and one sequence of letters.

Language

The language tests were three subtests from Siegmüller et al.'s (2010) standardized battery for receptive grammatical abilities in German: subtest 3 for the comprehension of SVO and OVS sentences (e.g., *Die kinder zeichnet der Mann* 'The$_{_ACC}$ children draws the$_{_NOM}$ man'); subtest 5 for the comprehension of sentences containing reflexives and pronouns (*Der Papa wäscht ihn* 'The$_{_NOM}$ father washes him$_{_ACC}$'); and subtest 6 for the comprehension of various types of relative clauses (right-branching SR: *Den Hasen schiebt der Esel, der weint* 'The$_{_ACC}$ bunny pushes the$_{_NOM}$ donkey that$_{_NOM}$ cries'; center-embedded OR: *Der Mann, den der Indianer trägt, liest* 'The$_{_NOM}$ man, that$_{_ACC}$ the$_{_NOM}$ Indian carries, reads'). In all these tests, the task is to point to one picture out of three that best corresponds to a sentence read aloud by the experimenter.

Procedure

The experiment was carried out at a university lab, in a quiet and child-friendly room. Participants were seated at a distance of 55–70 cm from a DELL laptop (screen resolution 1600 × 900, white background), connected to an SMI RED-m eye-tracker (sample rate 60 Hz). The experiment was run over the SMI Experiment Center software. An experimenter sat next to the participant, observing the tracking quality on a separate monitor and moving from one trial to the next, or repeating a trial if necessary, by pressing keys on an external keyboard. The experimenter also registered by hand the participant's verbal response in each trial.

In an introduction video, displayed prior to the experiment, Nellie the dog appeared and explained she would like to have the child's help in learning the color names. She explained the task and gave three example questions that served as warm-up trials. Participants received feedback on their responses to the practice trials, but not during the actual experiment. After the warm-up items, Nellie showed and named all the animals as well as the actions (chasing and tickling) that would appear in the game. The story teller also said she would appear every now and then and play with her friends. This, together with the appearance of the dog as the middle animal in the relevant trials, established the referent for the 1st-person pronoun and made its usage felicitous.

In the experiment, each trial started with a preamble video in which the animals of the scene were presented and their colors were named. The referent of the 3rd-person pronoun was stressed prosodically in the preamble, in order to make it more salient in the discourse. The test question followed the preamble video immediately (**Figure 1** shows an example of a visual scene with the preamble text and the test sentence accompanying it. An example of a preamble text and a test sentence for each of the conditions, as well as a video exemplifying a trial, can be found in the online supplementary material.). Upon hearing the question about the color of one of the animals, participants answered and the experimenter noted their response on a sheet. In case of no response the experimenter offered the participant to listen again to the question. In such cases, both the preamble and the

[5]To be sure, we performed all the analyses after excluding the four children who scored lower than 50% on the test for comprehension of OVS sentences. The results were qualitatively similar to those of the analysis in which these children are included. We therefore report the results from the analysis that includes all children.

test question were replayed and only the second response was counted in the analysis. A short break was taken after every 10 items. The entire duration of the experiment was approximately 20 min. Children, who were generally engaged and happy to participate, received stickers as a reward.

The forward and backward span tasks and the language tests were administered in a separate session, 1–3 weeks after the first appointment, at the same room at the university lab. The instructions for the forward span task were given following the protocol of this test (IDS, Grob et al., 2009). The instructions for the backward span task were based on those given in another such test that has norms from older children (HAWIK, Petermann and Petermann, 2008). In the forward span task, the experimenter read to the children the sequences of digits and letters and the child was required to repeat each sequence in the order in which the items had been presented. In the backward span task, the child heard the same sequences read by the experimenter and was instructed to repeat each sequence in the exact opposite order. The task was interrupted if the child failed to correctly repeat three consecutive sequences. The order of testing was the same for all children: the forward digit span test was administered first, then the backward digit span test, followed by the three language tests [comprehension of (a) OVS sentences; (b) pronouns and reflexives; and (c) relative clauses].

Results

We analyzed the data using the *lme4* package (Bates et al., 2014) in the R environment (R Development Core Team, 2014). The categorical accuracy data were analyzed with logit mixed models (Jaeger, 2008). The eye-tracking data were analyzed using linear mixed models with empirical logit as dependent variable (Barr, 2008). The eye-gaze plots present the data after having removed the individual differences from the dependent variable, based on the outcome of the linear mixed model. This was done using the *remef* function (Hohenstein and Kliegl, 2013). The plots therefore present the results on which the statistical inferences are based, that is, the ones that are derived from the statistical model. Importantly, in the case of the data presented here, plotting the partial effects yielded patterns qualitatively similar to those of the observed data. This means that removing the individual differences did not alter the general pattern in the data. For each of the eye-gaze plots, a corresponding figure showing the observed data is provided in the online supplementary material, for the sake of comparison. Memory Score (average score on the two span tests) and Language Score (average score on the three language tests) were inserted into the mixed-effects model analysis as continuous covariates, without splitting the group of participants. However, for the sake of presenting the data (either in a plot or in a table), the group was divided into children who scored higher vs. those who scored lower on the tests. This division was done with a median split. Scatterplots showing the individual performance pattern (for both the accuracy and the eye-tracking data) in relation to the average score on the memory and language tests can be found in the online supplementary material. In this section, we report the most relevant results of the analyses. The complete output of each model is listed in the online supplementary material.

The data from six children who did not do the memory and language tests were excluded, so the analysis of the accuracy data is based on 41 children. For two among these, eye-tracking failed due to technical problems during the testing session. Thus, the analysis of the eye-tracking data is based on 39 children. In the eye-tracking data analysis, we excluded 35 trials (2.2% of the total trials available) in which there was more than 50% data loss. The excluded items were distributed across all conditions and several participants. Prior to the analysis, we checked whether the participants performed similarly on trials with the verb *jagen* 'chase' and on those with the verb *kitzeln* 'tickle.' There was no substantial difference in the performance on trials involving these two actions, neither in terms of response accuracy nor in terms of eye-gaze. Hence, all trials were analyzed together.

Accuracy

Response accuracy was calculated based on the color named by the participants (Arnon, 2010). Naming the color of the target animal was scored as 1; otherwise as 0. Without taking into account the individual differences of language and memory abilities, children performed on the OR + 1pro condition 97% (SE = 0.03) accurately, on the OR + 2DP condition 47% (SE = 0.02) and on the OR + 3pro condition 44% (SE = 0.03). These accuracy percentages were compared to chance level using one-sample *t*-tests (chance level was set at 0.5 since, although there were three regions of interest in the visual scene, children never named the color of the middle animal, indicating that they never considered it a possible answer). Only performance on the OR + 1pro condition was significantly above chance ($t = 43.06$). On the OR + 2DP and OR + 3pro conditions, performance was at chance ($t = -0.59$ and $t = -1.16$, respectively).

The results look different when language and memory abilities are considered. **Figure 2** shows the pattern of relation between children's scores on the language and memory tests, and how it is manifested in their performance on each of the three conditions. The ceiling performance on the OR + 1pro condition was not influenced by language and memory abilities. The pattern that emerges in the OR + 2DP condition is similar to that in the OR + 3pro condition. A lower score on the language tests determined a below-chance performance on these two conditions, whereas a higher score on the language tests determined a more accurate performance on them.

The accuracy data were fit into a logit mixed model, including Condition as fixed factor, Language Score and Memory Score as two continuous covariates (without splitting the participant group) and intercepts for random effects of subjects and items. The OR + 1pro condition was excluded from the analysis to avoid the impact of extreme differences in task performance on the model outcome. All the terms that contain an interaction between Language and Memory were included, since these two covariates did not correlate significantly ($r = 0.08$, $t = 0.45$). A table of correlations between the language measure, the memory measure and response accuracy is provided in the online

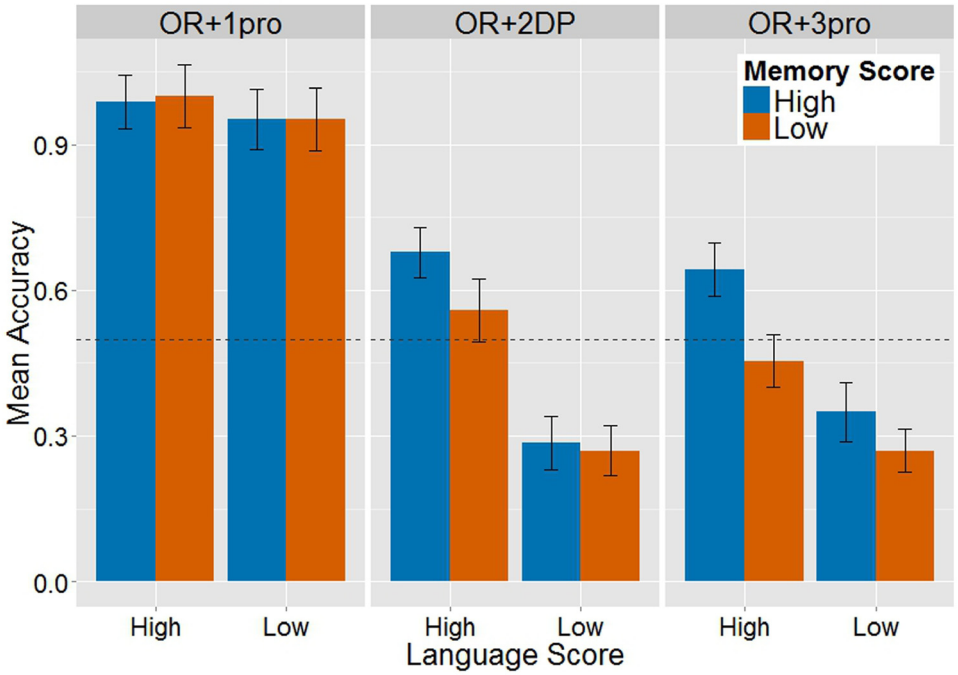


FIGURE 2 | Mean response accuracy (±1 SE) on the three conditions, in relation to children's scores on the language tests (on the *x*-axis) and on the memory tests (blue = High Score; orange = Low Score). The horizontal dashed line marks the chance level of 0.5.

supplementary material. The main effect Condition was not statistically significant (coef = −0.12, SE = 0.49, $z = -0.25$, $p = 0.81$), confirming that performance on OR + 2DP and OR + 3pro was overall similar. The main effect Language Score was significant (coef = 0.36, SE = 0.16, $z = 2.26$, $p = 0.02$), and so was the interaction Condition by Language Score (coef = −0.31, SE = 0.13, $z = -2.34$, $p = 0.02$). This interaction reflects the fact that, whereas performance on the OR + 2DP and OR + 3pro conditions was the same in children with lower language scores, children with higher language scores were significantly more accurate on OR + 2DP than on OR + 3pro. None of the terms that include Memory Score (main effect Memory and the interactions Condition by Memory, Language by Memory as well as Condition by Language by Memory) was statistically significant. Hence, we see that children's performance on OR + 2DP and OR + 3pro in the off-line data is modulated by language, but not by memory capacity.

Eye-Tracking

Figure 3 shows, for each of the three conditions, the proportion of target looks of children with high and low scores on the memory tests, broken by their scores on the language tests in order to see the relation between the two cognitive measures. The plot shows the data within the relevant time window, defined *a priori* for the analysis, rather than for the entire trial duration. This window starts at the offset of the relative pronoun *den* (plus 200 ms, the average time span necessary for programming and executing an eye movement; Trueswell, 2008). Note that the part that precedes the relative pronoun (*Welche Farbe hat der Hase,...* 'What color has the bunny...') is ambiguous about whether the sentence is a SR or an OR. However, based on the unambiguously accusative case-marked relative pronoun, it is already possible (and, indeed, very likely for adult speakers at least) to correctly predict that the sentence will turn out to be an OR. For these reasons, the beginning of the critical time window has been set at the beginning of the critical information in the sentence, that is, after the relative pronoun has been processed. This window ends after the 2-s long silence that followed the test question.

Within this time window, the effects we are interested in might start from the onset of the embedded subject DP onward, while the embedded full DP or pronoun and the verb are processed. Another (perhaps more plausible) possibility is that the effects emerge also in the 2-s long silence following the test sentence. In other words, children might continue to process the structure even after the sentence offset (Brandt-Kobele and Höhle, 2010; Adani and Fritzsche, 2015). Importantly, by including the post-sentential silence in the analysis time window we account for effects that might occur upon processing the verb, which is the very last word in the sentence. This is relevant in the light of studies with adults that predict the effect to occur at the verb, the point in which the filler-gap dependency is resolved (e.g., Gibson, 2000; Gordon et al., 2001, 2002; Warren and Gibson, 2002; Lewis et al., 2006; O'Grady, 2011).

Within the critical time window, which was approximately 2800 ms long, the dependent variable was the proportion of looks to the target figure, calculated as looks to the target animal

divided by looks to all the three animals in the visual scene. An accurate processing of the sentence in terms of eye-gaze might be expressed by faster looks to the target (earlier increase in proportion of target looks, or PTL), by more target looks (higher PTL), or by both. Note that, in the analysis procedure adopted here (Barr, 2008), Time is included in the model as a continuous covariate. Therefore, the analysis does not provide information about the specific point in which the effect occurs. For this reason, we will not be able to say how long exactly after the embedded subject DP or the embedded verb have been processed the effect starts. However, the advantage in such an analysis is that the time-related information is obtained in its entirety, without the necessity to cut time into chunks and lose information about the timely course of the gaze pattern. The time-related information is expressed here in the form of significant interactions with the Time covariate. For instance, a significant interaction Condition by Time would mean that, over time (without knowing where exactly during the analyzed window), target looks in one condition increase more than in another condition. For the analysis, each of the pronoun conditions was compared to the baseline condition with two full DPs, using sliding contrast specification (OR + 1pro vs. OR + 2DP vs. OR + 3pro). The plot and analysis of the eye-gaze data include all the trials in the experiment, independently of whether they were answered correctly or incorrectly.

Let us turn to the gaze pattern shown in **Figure 3**. In the OR + 1pro condition, the increase in target looks is faster and the PTL is higher (peaking around 1200 ms into the critical time window) than in the other two conditions, reflecting what we find in the accuracy data. Individual differences in language and memory skills do not appear to affect this pattern. In the OR + 2DP condition, children with a low score on the language tests look less to the target independently of their memory score (lower middle panel in **Figure 3**). Children with a higher language score (upper middle panel) look faster to the target when their memory score is high (culminating at about 1500 ms), as compared to when their memory score is low. These high-language but low-memory children eventually look to the target like their high-memory peers, but at a later point (around 1800 ms). In the OR + 3pro condition, children with a low language score again look less to the target independently of their memory score (lower right panel). However, a clear difference emerges between high-memory and low-memory children when their language score is high (upper right panel). Here, high-memory children look to the target faster and more than their low-memory peers.

Following Barr's (2008) procedure for the analysis of eye-tracking data in the visual-world paradigm, we performed only the by-subject analysis, aggregating the data across items. This was done due to the relatively small number of items per condition. The proportion of target looks was transformed to an empirical logit and used as the dependent variable in the model. Time, divided into 50 ms long bins, was centered around the point in which target looks started to increase when all conditions are collapsed together, based on a Grand Mean plot. We then fit a linear mixed model including Condition as fixed factor, Time as

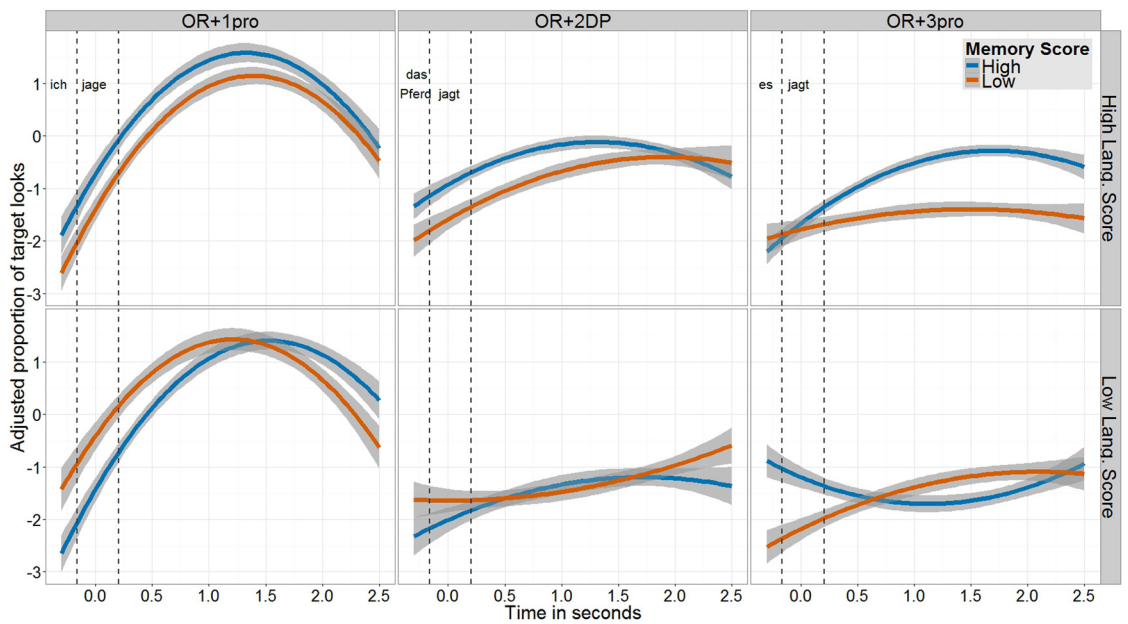


FIGURE 3 | Proportion of target looks (transformed to empirical logit and adjusted after the removal of individual differences) within the time window relevant for analysis, shown separately for each condition, divided by children's score on the memory tests (blue line = High Score; orange line = Low Score) and broken by their score on the language tests (top row = High Score; bottom row = Low Score). On the *x*-axis Time ranges from the offset of the relative pronoun until the end of the 2-s long silence that followed the sentence. Two vertical dashed lines mark the critical chunks in the analysis window: (1) embedded subject DP (*ich* 'I'; *das Pferd* 'the horse'; *es* 'it'); (2) embedded verb (*jage/t* 'chase/s'); (3) post-sentential silence. The analysis of the eye-gaze data was performed on the entire time window shown in the plot (chunks 1–3).

covariate with linear and quadratic polynomials, Language Score and Memory Score as additional continuous covariates (without group splitting) and an intercept for the random effect of subjects. As in the model for the accuracy data, all the terms that contain an interaction between Language and Memory were included as well, due to the lack of correlation between the two measures. The inclusion of a quadratic term for Time was justified by a comparison to a model with a linear term only ($\chi^2 = 726.3$, difference in Df = 12, $p < 0.001$).

The main effect Condition was significant for both comparisons, but in opposite directions: PTL in the OR + 1pro condition were significantly greater than those in the OR + 2DP condition (coef = −0.82, SE = 0.03, $t = -30.88$); PTL in the OR + 2DP condition were significantly greater than those in the OR + 3pro condition (coef = −0.25, SE = 0.03, $t = -9.46$). These effects mean that children looked to the target in OR + 1pro trials overall longer than in OR + 2DP trials, and in these longer than in OR + 3pro trials. The former effect reflects what we find in the accuracy data, but the advantage of OR + 2DP over OR + 3pro in terms of eye-gaze is absent in the accuracy data. Both the main effect of Language (coef = 0.06, SE = 0.03, $t = 1.98$) and the main effect of Memory (coef = 0.09, SE = 0.05, $t = 1.87$) were only marginally significant. Also the interaction Language by Memory was not statistically significant (coef = 0.07, SE = 0.04, $t = 1.73$). Most importantly, all the four-way interactions were significant. For the comparison OR + 1pro vs. OR + 2DP, the interaction Time by Condition by Language by Memory was significant (for the quadratic term of Time: coef = 3.88, SE = 1.82, $t = 2.13$). This effect reflects the pattern observed in the two middle and the two left panels of **Figure 3**. No individual differences in language and memory emerge in the performance on the OR + 1pro condition, whereas differences do emerge in the OR + 2DP condition depending on language and memory scores. Also for the comparison OR + 2DP vs. OR + 3pro, the interaction Time by Condition by Language by Memory was significant (for the linear term of Time: coef = 8.41, SE = 1.80, $t = 4.66$; for the quadratic term of Time: coef = −6.39, SE = 1.76, $t = -3.63$). This effect reflects what we see in the two middle and the two right panels of **Figure 3**. When language score is low, the gaze pattern in the two conditions is the same independently of the memory score. But when language score is high, the differences between high-memory and low-memory children are more pronounced in the OR + 3pro condition than in the OR + 2DP condition: only in the latter the low-memory children eventually look to the target like their high-memory peers, albeit later.

Looks to Distractor

Before discussing the results, let us examine the pattern of children's looks to the distractor animal. Recall that, in their off-line responses on incorrect trials, children named the color of the distractor animal, never that of the middle animal. **Figure 4** shows, for each of the three conditions, the proportion of distractor looks in children with high and low scores on the memory tests, broken by their language scores (again, we plot here the partial effects; the corresponding plot showing the

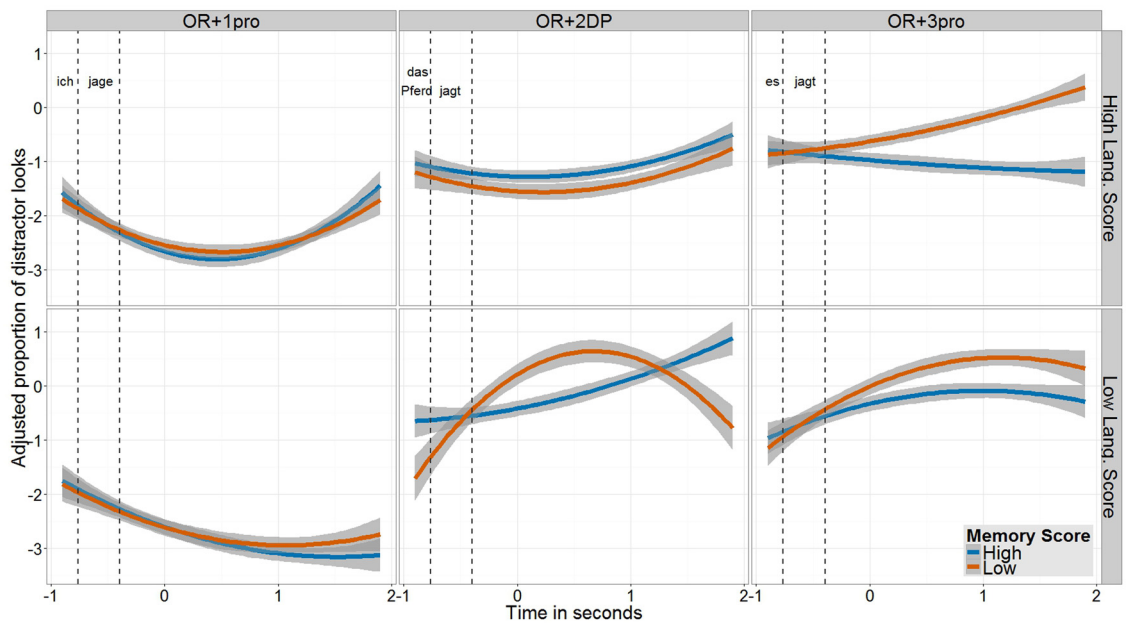


FIGURE 4 | Proportion of looks to the distractor figure (transformed to empirical logit and adjusted after the removal of individual differences) within the time window relevant for analysis, shown separately for each condition, divided by children's score on the memory tests (blue line = High Score; orange line = Low Score) and broken by their score on the language tests (top row = High Score; bottom row = Low Score). On the *x*-axis Time ranges from the offset of the relative pronoun until the end of the 2-s long silence that followed the sentence. Two vertical dashed lines mark the critical chunks in the analysis window: (1) embedded subject DP (*ich* 'I'; *das Pferd* 'the horse'; *es* 'it'); (2) embedded verb (*jage/t* 'chase/s'); (3) post-sentential silence.

observed data is provided in the online supplementary material). As expected, and reflecting children's off-line responses, on the OR + 1pro condition their looks to the distractor are very low. By contrast, on the OR + 2DP and OR + 3pro conditions, the proportion of distractor looks throughout the critical time window is very high, mostly for children with lower memory scores. That is, children's errors were expressed by their systematic (off-line as well as on-line) interpretation of the OR as a SR, treating the DP head as the subject rather than the object of the embedded clause. This pattern of error is typically found in studies on children's comprehension of relative clauses.

Discussion

The aim of the study was to test the effects of various pronoun types on children's processing of ORs. We took as reference condition ORs with a full DP head and an embedded full DP subject, which are typically hard for children, and manipulated the embedded subject using personal pronouns. The three OR types were structured with a masculine noun as DP head, which had the advantage of facilitating, at least potentially, children's comprehension. This was achievable due to the possibility to recognize the sentence as an OR rather early in the sentence, upon processing the accusative case-marking on the relative pronoun (*den*). There is evidence from previous studies on relative clause comprehension in German (Arosio et al., 2012) that children are facilitated when the relative clause (whether a SR or an OR) is disambiguated by case (as in our stimuli), as compared to when it is disambiguated by a singular or plural number-marking on the embedded verb (in our stimuli, the verb was always marked with singular). Another characteristic of the three conditions we tested is that they differ with respect to the referring expression in the embedded subject position–full DP, 1st- or 3rd-person pronoun. We therefore expect these referring expressions to trigger effects in task performance, if their referential properties play a role in determining OR processing. The initial prediction, as made by Friedmann et al. (2009) and by other accounts, is that ORs with embedded pronominal subjects are more accurately comprehended than ORs with two full DPs, independently of the pronoun type. Our findings support this prediction only partially.

First, we find that children are more accurate on ORs with an embedded 1st-person pronoun than ORs with two full DPs, both in terms of off-line accuracy and in terms of on-line eye-gaze, where we find more target looks in the OR + 1pro than in the OR + 2DP condition. This finding supports the initial prediction. It is also in line with other studies, both with children and with adults, showing that a 1st- or 2nd-person pronoun in the embedded subject position makes the OR easier to process (Gordon et al., 2001; Warren and Gibson, 2002, 2005; Arnon, 2010).

We also find that ORs with 1st-person pronoun are more accurately processed (again, both off-line and on-line) than ORs with 3rd-person pronoun. This result is not in line with the RM account, since the prediction is that different pronoun types in the embedded subject position facilitate ORs to an equal extent. The reason is that in both cases the full DP head, which contains the [+NP] feature, crosses an intervening pronoun, a constituent that lacks the [+NP] feature. This result appears to disagree also with other accounts that predict facilitated performance on ORs with an embedded pronoun, independently of the pronoun type (e.g., Gordon et al., 2001; Lewis et al., 2006). The pronoun asymmetry suggests that defining the (dis)similarity between the DP head and the embedded subject DP only in terms of 'lexical restriction,' that is, in terms of a full DP vs. a personal pronoun, is not sufficient. This pronoun asymmetry is in line, however, with theoretical accounts on referential properties of pronouns (Heim, 1991; Recanati, 1993; Erteschik-Shir, 1997; Ariel, 2001; Legendre and Smolensky, 2012) as well as with previous experimental studies with adults. Both Warren and Gibson (2002) and Carminati (2005) found that the presence of a 1st-person pronoun facilitates adults' sentence processing more than the presence of a 3rd-person pronoun. These studies explain such an asymmetry in terms of the different referential properties of the pronouns. Since discourse referents of 1st-person pronouns are accessed directly, these pronouns are less costly for processing than 3rd-person pronouns, which need to be resolved via an antecedent (in the sentential or extra-sentential context), before the discourse referent of the pronoun is accessed. This is also the case in the present study: the discourse referent of the 3rd-person pronoun is accessed only after the pronoun has been resolved via an antecedent, which had to be retrieved from the linguistic context provided in the preamble video before the trial. Hence, the presence of the pronoun in itself does not necessarily facilitate OR processing. It seems that only pronouns that relate to their discourse referents directly, like 1st-person pronouns, do so[6]. The facilitation found by Friedmann et al. (2009) with Hebrew ORs containing an embedded arbitrary *pro* subject (example 6 in the *Introduction*) can be explained on similar terms. The Hebrew arbitrary *pro* is used when the agent of the action remains unspecified. It might well be that the facilitation was due to the discourse properties of *pro*–the fact that it does not relate to any specific discourse referent, thus reducing processing cost–rather than to its property of lacking the [+NP] feature, as suggested by the authors.

A third pattern, that emerges in the eye-gaze data, is that ORs with a 3rd-person pronoun are actually harder for children than ORs with two full DPs. This finding is not in line with the prediction that any kind of pronoun in the embedded subject position facilitates OR comprehension (e.g., Gordon et al., 2001; Friedmann et al., 2009; Rizzi, 2013). It can be explained, again, if the referential properties of the referring expressions are taken

[6]Recall that the middle animal in the visual scenes accompanying the OR + 1pro condition was always the dog, the narrator. One reviewer pointed out that children's high performance on this condition might reflect their familiarity with this animal, rather than the effect caused by the pronoun itself. We have already addressed this issue in a follow-up study, yet to be published. Using similar material and methodology, we tested children on different types of relatives (SRs and ORs), in which the figure of the narrator appeared in various experimental conditions and in some fillers. In this setup, it was impossible to anticipate the type of sentence based on the visual presence of the narrator. Importantly, the results show that the 1st-person pronoun advantage over the 3rd-person pronoun persists, similarly to what we find in the present study.

into account. A 3rd-person pronoun can be interpreted only after it has been related to an antecedent, which needs to be located and retrieved from the linguistic or extra-linguistic context. This is not the case with a full DP, whose discourse referent is derived from its lexical meaning and accessed directly. Note that, just like in an OR with a 1st-person pronoun, also in an OR with 3rd-person pronoun the DP head crosses an intervening pronoun. The fact that the former condition is easier than the latter, compared to the baseline with two full DPs, supports further the claim that the presence of the pronoun on its own cannot account for children's performance. Rather, the type of pronoun–and more precisely, the referential properties of that pronoun–appear to play a major role in facilitating or not facilitating the processing of the OR.

Interestingly, Goodluck (2005, 2010) managed to separate intervention locality effects from complex discourse accessibility operations. Goodluck (2005) manipulated the discourse accessibility operation in object-extracted wh-questions by making it more demanding (*Which lion did the zebra kiss?*) or less demanding (*Which animal did the zebra kiss?*). Crucially, in both cases, the intervention locality effect was present (in both sentences, both the moved object DP and the intervening subject DP are lexically restricted). The fact that children were more accurate on the *which-animal* question than on the *which-lion* led the author to conclude that discourse accessibility determines children's performance on the structure independently of the syntactic complexity. This is reminiscent of what we find in the two pronoun conditions. Both in OR+1pro and in OR + 3pro, the (reduced) syntactic complexity is kept constant due to the embedded pronoun. Therefore, children' higher accuracy rate on OR + 1pro than on OR + 3pro is likely due to the different referential properties of the pronouns. In other words, the direct discourse accessibility in the case of the 1st-person pronoun makes this condition easier than the 3rd-person pronoun condition, in which discourse accessibility is indirect and therefore more demanding.

Note that the advantage of the OR + 2DP condition over OR + 3pro, in terms of main effect, is found only in the on-line eye-gaze data. An even more crucial finding is that the effects of memory only emerge in the on-line data, whereas they remain hidden when looking at the off-line accuracy data. These findings join a growing body of studies that show that children's performance sometimes appears different when tested by means of explicit or implicit responses. Specifically, measures of implicit processing (such as eye-tracking) often suggest that children accurately parse ORs even though their explicit performance on the same ORs remains poor (Adani and Fritzsche, 2015; see also discussion in Brandt-Kobele and Höhle, 2010). In the present study we show that children looked faster or longer to the target figure in conditions that they processed more accurately than in conditions that were harder for them. In other words, when children correctly processed a sentence their attention on the target figure was more stable in comparison to harder sentences.

These eye-gaze effects were found within the 2800 ms long time window defined *a priori* for the analysis. A widespread assumption, supported by evidence from on-line processing studies with adults, is that such effects occur upon processing the embedded verb of an OR, the site in which the filler-gap dependency is resolved (e.g., Gibson, 2000; Gordon et al., 2001, 2002; Warren and Gibson, 2002; Lewis et al., 2006; O'Grady, 2011). Although Friedmann et al. (2009) do not make specific predictions regarding the exact point in which intervention effects occur, it seems they do so in subsequent work (Belletti et al., 2012), suggesting that intervention effects are detectable only when the two relevant DPs (the head noun and the embedded subject in an OR) are similar in terms of morphological features that are overtly marked on the embedded verb. Hence, it seems that also according to the RM account intervention effects in ORs are expected to occur at the embedded verb. This idea is entertained also in Franck et al. (2015).

Analyzing the eye-gaze data in the entire time window from the offset of the relative pronoun until the end of the post-sentential silence does not allow the detection of time-locked effects. Nevertheless, it had several motivations and some evident advantages. First, the part of the sentence that precedes the relative pronoun, which was equal in the three conditions, is not informative enough to guide the participants toward the identification of the relevant referent. We therefore do not expect any gaze pattern prior to hearing the relative pronoun to be driven by the linguistic input. Second, processing the unambiguously accusative case-marked relative pronoun is virtually enough to be able to identify the sentence as an OR and thus the correct referent. Even though we do not expect to find evidence for such rapid processing in 5-year-olds, the crucial point is that the relative pronoun is the first informative point in the sentence. Third, young children might be slow in processing the OR, and effects stemming from their eye-gaze might well emerge after the critical information has been processed. Several visual-world studies have even found effects occurring after the sentence ended (e.g., Brandt-Kobele and Höhle, 2010; Adani and Fritzsche, 2015). Crucially, the embedded verb in our stimuli is the last word in the sentence. Thus, post-sentential effects might be driven (also) by the filler-gap dependency resolution at the verb, as predicted, for instance, by Gibson (2000), Gordon et al. (2001, 2002), Warren and Gibson (2002), Lewis et al. (2006), O'Grady (2011) and other account. Finally, following Barr's (2008) analysis procedure, the inclusion of Time as a continuous covariate appears to be more appropriate in a linear mixed-effects model analysis. The main reason is that the effect of time (the change in gaze pattern throughout the duration of the trial) is captured in its entirety, whereas by cutting it into chunks some information about the time course of the gaze pattern is lost.

Concerning language and memory abilities, we have looked at the role of children's memory capacity in their OR processing and at its relation to the role of their language skills. The goal was to test whether effects which are due to language and memory depend on each other or not and, if they do, in what manner. We had previously shown that, on the two harder conditions (OR + 2DP and OR + 3pro), children with stronger language abilities are significantly more accurate than children with weaker language skills (Haendler et al., 2015). Given the linguistic material used in the three administered subtests, we reasoned that

stronger language or grammatical skills meant a stronger ability to compute movement-derived structures (subtests on sentences with canonical and non-canonical word order) and a stronger ability in discourse accessibility operations (subtest on reflexives and pronouns). It is therefore not surprising that children who had a higher average score on these tests were more accurate on ORs that were more difficult in terms of computing the syntactic movement (OR + 2DP) and on ORs that were more difficult in terms of discourse accessibility (OR + 3pro). On the OR + 1pro condition, in which both the computation of the syntactic movement and discourse accessibility are facilitated, all children were accurate independently of their score on the grammatical tests.

In the present study, adding memory abilities to the picture reveals a more fine-grained pattern in the effects of language skills previously found. The analysis shows that language and memory have independent, additive effects that vary in relation to the experimental conditions. Children are most accurate on the OR + 1pro condition, but neither their response accuracy nor their eye-gaze are influenced by individual differences in language and memory abilities. Individual differences in language and memory do affect, however, performance on the OR + 2DP and OR + 3pro conditions, but the effects of memory are observable only in the eye-gaze data, as mentioned earlier. Whether children with weaker grammatical skills have stronger or weaker memory does not seem to affect their performance substantially. By contrast, the gaze pattern of children with stronger grammatical skills clearly changes depending on their memory capacity. In the OR + 2DP condition, low-memory (and high-language) children look to the target like their high-memory peers, but later, suggesting an accurate albeit delayed processing of the sentence. In the OR + 3pro condition, low-memory (and high-language) children look to the target less than their high-memory peers up to the end of the trial, showing no evidence of correct processing of the sentence. **Table 1** summarizes these findings in a schematic way.

To account for these results, we will now explain what might cause the qualitative differences among the conditions and how language and memory abilities might play a role in creating the effects we find. The three conditions are similar in their syntactic structure, in the sense that they are all ORs in which the DP head moves from the embedded object position. Processing this movement, and resolving the filler-gap dependency, is assumed to be facilitated in the two pronoun conditions. According to the RM account, the syntactic complexity of OR + 1pro and OR + 3pro is reduced due to the attenuation of the intervention locality effect, since the full DP head crosses an intervening pronoun rather than another full DP (Friedmann et al., 2009; Rizzi, 2013). The syntactic complexity of ORs with pronouns is reduced also from the perspective of the integration cost metric account (Gibson, 1998, 2000; Warren and Gibson, 2002, 2005) and according to the similarity-based and cue-based interference approach (Gordon et al., 2001, 2002, 2004; Lewis and Vasishth, 2005; Lewis et al., 2006; van Dyke and McElree, 2006, 2011). All these accounts argue that facilitated processing of ORs with embedded pronouns is due to reduced burden on memory resources (see also Sheppard et al., 2015). The three conditions differ, however, with respect to the referring expression in the embedded subject position: these referring expressions require different levels of processing cost in terms of discourse accessibility. The 1st-person pronoun and the full DP relate to their discourse referents directly, deriving them from their lexical meanings, whereas the 3rd-person pronoun relates to its discourse referent indirectly, deriving it from the meaning of the antecedent to which it relates. This implies that referring expressions (such as 1st-person pronouns and full DPs) whose discourse referent is accessed directly overload memory resources less than referring expressions (such as 3rd-person pronouns) whose discourse referent has to be retrieved from the previously encoded context (Warren and Gibson, 2002; van Rij et al., 2013).

These syntactic and discourse characteristics of the conditions appear to explain the pattern we find in the data. In particular, they might account for the role of memory capacity and its additive effects to those of language skills. Language skills, as defined by the average score on the three language tests, appear to be the underlying constraint on children's performance. If children score low on these tests–in other words, if they are less proficient in processing movement-derived structures and in accessing discourse (these are the two relevant operations assessed by the language tests, as we have seen)–then we find no evidence for accurate processing of the two conditions that are hard either due to syntactic movement (OR + 2DP, in which a full DP moves over another full DP) or due to discourse accessibility (OR + 3pro, in which accessing the discourse referent of the 3rd-person pronoun is more demanding). It seems that, in the case of low-language children, some basic grammatical skills are weaker and therefore their memory capacity does not make any difference. Not surprisingly, even low-language children succeed on the OR + 1pro condition, which is less demanding both in terms of its syntactic movement and in terms of discourse accessibility. But also here memory capacity does not make any difference: this condition is equally easy for all children independently of their memory skills. What happens in children who score high on the three language tests? Just like their low-language peers, they perform at ceiling on the easiest OR + 1pro condition, independently of their memory capacity. A different pattern, modulated by memory, emerges in the two harder conditions (OR + 2DP and OR + 3pro). In OR + 2DP, high-memory children correctly process the structure, whereas low-memory children do so as well, but rather late. In OR + 3pro, there is evidence that only high-memory children

TABLE 1 | A summary of the cases in which we find evidence for accurate processing (in terms of on-line target looks) of the different conditions, depending on language, and memory abilities.

		OR + 1pro	OR + 2DP	OR + 3pro
High-language	High-memory	YES	YES	YES
	Low-memory	YES	YES, but late	NO
Low-language	High-memory	YES	NO	NO
	Low-memory	YES	NO	NO

YES, there is such evidence; NO, there is no such evidence.

correctly process the structure, whereas low-memory children are substantially less accurate.

Thus, memory capacity appears to be crucial when discourse accessibility is demanding (as when 3rd-person pronouns need to be resolved), but only if general linguistic abilities, such as computing syntactic movements and accessing discourse referents of pronouns and reflexives, are sufficiently strong. In the OR + 2DP condition, in which retrieving the referent of a full DP is less costly, even low-memory children eventually look to the target, although later than their high-memory peers. In the OR + 3pro condition, in which the retrieval of the referent of the 3rd-person pronoun is more costly, low-memory children do not catch up with their high-memory peers and there is no evidence that they accurately process the structure.

Our findings resemble, at least partly, those of Warren and Gibson (2002), who elaborate on the idea that memory resources are crucial for processing structures that require both filler-gap dependency resolution and accessing discourse referents of various referring expressions. These authors found the same asymmetry between 1st-person pronouns and 3rd-person pronouns, with the former facilitating OR processing more than the latter, an asymmetry which is explained in the light of Gibson's (1998, 2000) integration cost metric. According to the authors, the processing cost of a certain structure increases with the number of discourse referents that intervene between the filler and the gap site in which it is integrated. The reason is that each of the intervening discourse referents has to be integrated as well, thus reducing the memory resources available to process the structure. When one of the intervening discourse referents is a 1st-person pronoun, whose integration is done straightforwardly, the available memory resources are less burdened than in the case in which the intervening constituent is a 3rd-person pronoun, whose integration is more costly. Note, however, that in Warren and Gibson (2002) adults judged ORs with an embedded 3rd-person pronoun as less complex than ORs with two full DPs. This pattern is unlike what we find with children. In the present study, OR + 3pro appears to be the condition on which memory has the strongest impact. Given that children's memory abilities are underdeveloped, compared to adults,' it is not surprising that children with weaker memory skills struggle while processing ORs with an embedded 3rd-person pronoun, even if their ability to perform on the language tests we used is already strong.

Conclusion

Our data support only in part a purely syntax-based account such as Friedmann et al.'s (2009), or the similarity-/cue-based interference accounts of relative clause processing. While we do find that an embedded 1st-person pronoun facilitates OR processing, we also find that an embedded 3rd-person pronoun does not. It appears that OR processing is constrained not only by the syntactic complexity of the structure, but also by the referential properties of the involved constituents. Both require memory resources and might thus determine difficulties in processing the OR, as has been suggested for adults. The results suggest that both language and memory abilities play a role in modulating these syntactic and discourse accessibility constraints, and that they do so in an independent, additive fashion.

Acknowledgments

We thank: the participating children and their parents; Julia Billerbeck and Isabell Keßlau for their help with material preparation and data collection; the two reviewers for insightful comments and helpful suggestions. The research presented here was supported by a full scholarship granted to the first author from the *Ernst Ludwig Ehrlich Studienwerk* (Grant PF123) and by the SFB 632 "Information Structure" project, funded by the *Deutsche Forschungsgemeinschaft* (German Research Foundation). The publication of the paper was funded by the *Deutsche Forschungsgemeinschaft* and the *Open Access Publishing Fund* of the University of Potsdam, which are gratefully acknowledged.

Supplementary Material

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpsyg.2015.00860

References

Adani, F. (2011). Rethinking the acquisition of relative clauses in Italian: towards a grammatically based account. *J. Child Lang.* 38, 141–165. doi: 10.1017/S0305000909990250

Adani, F., Forgiarini, M., van der Lely, H. K. J., and Guasti, M. T. (2014). Number dissimilarities facilitate the comprehension of relative clauses in children with (Grammatical) Specific Language Impairment. *J. Child Lang.* 41, 811–841. doi: 10.1017/S0305000913000184

Adani, F., and Fritzsche, T. (2015). "On the relation between implicit and explicit measures of child language development: evidence from relative clause processing in 4-year-olds and adults," in *Proceedings of the 39th Annual Boston University Conference on Language Development*. Somerville, MA: Cascadilla Press.

Adani, F., van der Lely, H. K. J., Forgiarini, M., and Guasti, M. T. (2010). Grammatical feature dissimilarities make relative clauses easier: a comprehension study with Italian children. *Lingua* 120, 2148–2166. doi: 10.1016/j.lingua.2010.03.018

Ariel, M. (2001). "Accessibility theory: an overview," in *Text Representation,* eds T. Sanders, J. Schilperhood, and W. Spooren (Amsterdam: John Benjamins Publishing Company), 29–87. doi: 10.1075/hcp.8.04ari

Arnon, I. (2005). "Relative clause acquisition in Hebrew: towards a processing-oriented account," in *Proceedings of the 29th Annual Boston University Conference on Language Development*, Somerville, MA: Cascadilla Press.

Arnon, I. (2010). Rethinking child difficulty: the effect of NP type on children's processing of relative clauses in Hebrew. *J. Child Lang.* 37, 27–57. doi: 10.1017/S030500090900943X

Arosio, F., Guasti, M. T., and Stucchi, N. (2011). Disambiguating information and memory resources in children's processing of Italian relative clauses. *J. Psycholinguist. Res.* 40, 137–154. doi: 10.1007/s10936-010-9160-0

Arosio, F., Yatsushiro, K., Forgiarini, M., and Guasti, M. T. (2012). Morphological information and memory resources in children's processing of relative clauses in German. *Lang. Dev.* 8, 340–364. doi: 10.1080/15475441.2011.634691

Avrutin, S. (2000). "Comprehension of wh-questions by children and Broca's aphasics," in *Language and Brain: Representation and Processing,* eds Y. Grodzinsky, L. Shapiro, and D. Swinney (San Diego, CA: Academic Press), 295–312. doi: 10.1016/B978-012304260-6/50017-7

Baddeley, A. D. (1986). *Working Memory*. Oxford: Clarendon Press.

Baddeley, A. D., Eysenck, M. W., and Anderson, M. C. (2009). *Memory*. New York, NY: Psychology Press.

Barr, D. J. (2008). Analyzing 'visual world' eyetracking data using multilevel logistic regression. *J. Mem. Lang.* 59, 457–474. doi: 10.1016/j.jml.2007.09.002

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2014). *lme4: Linear Mixed-Effects Models using Eigen and S4. R Package Version 1.1-6*. Available at: http://CRAN-R-project.org/package~$=$~lme4

Belletti, A., Friedmann, N., Brunato, D., and Rizzi, L. (2012). Does gender make a difference? Comparing the effect of gender on children's comprehension of relative clauses in Hebrew and Italian. *Lingua* 122, 1053–1069. doi: 10.1016/j.lingua.2012.02.007

Booth, J. R., MacWhinney, B., and Haraskai, Y. (2000). Developmental differences in visual and auditory processing of complex sentences. *Child Dev.* 71, 981–1003. doi: 10.1111/1467-8624.00203

Boyle, W., Lindell, A. K., and Kidd, E. (2013). Investigating the role of verbal working memory in young children's sentence comprehension. *Lang. Learn.* 63, 211–242. doi: 10.1111/lang.12003

Brandt, S., Kidd, E., Lieven, E., and Tomasello, M. (2009). The discourse bases of relativization: an investigation of young German and English-speaking children's comprehension of relative clauses. *Cogn. Linguist.* 20, 539–570. doi: 10.1515/COGL.2009.024

Brandt-Kobele, O., and Höhle, B. (2010). What asymmetries within comprehension reveal about asymmetries between comprehension and production: the case of verb inflection in language acquisition. *Lingua* 120, 1910–1925. doi: 10.1016/j.lingua.2010.02.008

Brener, R. (1983). Learning the deictic meaning of third person pronouns. *J. Psycholinguist. Res.* 12, 235–262. doi: 10.1007/BF01067669

Carminati, M. N. (2005). Processing reflexes of the Feature Hierarchy (Person > Number > Gender) and implications for linguistic theory. *Lingua* 115, 259–285. doi: 10.1016/j.lingua.2003.10.006

Clackson, K., Felser, C., and Clahsen, H. (2011). Children's processing of reflexives and pronouns in English: evidence from eye-movements during listening. *J. Mem. Lang.* 65, 128–144. doi: 10.1016/j.jml.2011.04.007

Engel de Abreu, P. M. J., Gathercole, S. E., and Martin, R. (2011). Disentangling the relationship between working memory and language: the roles of short-term storage and cognitive control. *Learn. Individ. Dif.* 21, 569–574. doi: 10.1016/j.lindif.2011.06.002

Erteschik-Shir, N. (1997). *The Dynamics of Focus Structure*. Cambridge: Cambridge University Press.

Franck, J., Colonna, S., and Rizzi, L. (2015). Task-dependency and structure-dependency in number interference effects in sentence comprehension. *Front. Psychol.* 6:349. doi: 10.3389/fpsyg.2015.00349

Friedmann, N., Belletti, A., and Rizzi, L. (2009). Relativized relatives: types of intervention in the acquisition of A-bar dependencies. *Lingua* 119, 67–88. doi: 10.1016/j.lingua.2008.09.002

Fukumura, K., and van Gompel, R. P. G. (2012). Producing pronouns and definite noun phrases: do speakers use the addresse's discourse model? *Cogn. Sci.* 36, 1289–1311. doi: 10.1111/j.1551-6709.2012.01255.x

Gibson, E. (1998). Syntactic complexity: locality of syntactic dependencies. *Cognition* 68, 1–76. doi: 10.1016/S0010-0277(98)00034-1

Gibson, E. (2000). "The dependency locality theory: a distance-based theory of linguistic complexity," in *Image, Language, Brain,* eds Y. Miyashita, A. P. Marantz, and W. O'Neil (Cambrdige, MA: The MIT Press), 95–126.

Girouard, P. C., Ricard, M., and Décarie, T. G. (1997). The acquisition of personal pronouns in French-speaking and English-speaking children. *J. Child Lang.* 24, 311–326. doi: 10.1017/S030500099700305X

Goodluck, H. (1990). "Knowledge integration in processing and acquisition," in *Language Processing and Language Acquisition,* eds L. Frazier and J. de Villiers (Dordrecth: Kluwer Academic Publishers), 369–382.

Goodluck, H. (2005). "Discourse-linking and question formation: comprehension effects in children and Brocas aphasics," in *UG and External Systems: Language, Brain and Computation,* ed. A. M. Di Sciullo (Amsterdam: John Benjamins Publishing Company). doi: 10.1075/la.75.12goo

Goodluck, H. (2010). Object extraction is not subject to child relativized minimality. *Lingua* 120, 1516–1521. doi: 10.1016/j.lingua.2009.10.005

Gordon, P. C., Hendrick, R., and Johnson, M. (2001). Memory interference during language processing. *J. Exp. Psychol. Learn. Mem. Cogn.* 27, 1411–1423. doi: 10.1037/0278-7393.27.6.1411

Gordon, P. C., Hendrick, R., and Johnson, M. (2004). Effects of noun phrase type on sentence complexity. *J. Mem. Lang.* 51, 97–114. doi: 10.1016/j.jml.2004.02.003

Gordon, P. C., Hendrick, R., and Levine, W. H. (2002). Memory-load interference in syntactic processing. *Psychol. Sci.* 13, 425–430. doi: 10.1111/1467-9280.00475

Grillo, N. (2005). "Minimality effects in agrammatic comprehension," in *Proceedings of ConSOLE XIII*, eds S. Blaho, E. Schoorlemmer, and L. Vicente (106–120). Available at: http://www.sole.leidenuniv.nl/

Grillo, N. (2009). Generalized minimality: feature impoverishment and comprehension deficits in agrammatism. *Lingua* 119, 1426–1443. doi: 10.1016/j.lingua.2008.04.003

Grob, A., Meyer, C. S., and Hagmann-von Arx, P. (2009). *Intelligence and Development Scales: Intelligenz und Entwicklungsskalen für Kinder von 5-10 Jahren*. Bern: Verlag Hans Huber.

Grünloh, T., Lieven, E., and Tomasello, M. (2011). German children use prosody to identify participant roles in transitive sentences. *Cogn. Linguist.* 22, 393–419. doi: 10.1515/cogl.2011.015

Gutierrez-Mangado, M. (2011). Children's comprehension of relative clauses in an ergative language: the case of Basque. *Lang. Acquis.* 18, 176–201. doi: 10.1080/10489223.2011.580674

Haendler, Y., Kliegl, R., and Adani, F. (2015). "The impact of pronoun type and grammatical skills on children's processing of object relative clauses," in *Proceedings of the 39th Annual Boston University Conference on Language Development*, Somerville, MA: Cascadilla Press.

Hartshorne, J. K., Nappa, R., and Snedeker, J. (2015). Development of the first-mention bias. *J. Child Lang.* 42, 423–446. doi: 10.1017/S0305000914000075

Heim, I. (1991). "Artikel und Definitheit," in *Semantik: Ein internationales Handbuch derzeitgenoessischen Forschung*, eds A. von Stechow and D. Wunderlich (Berlin: De Gruyter), 487–535.

Hohenstein, S., and Kliegl, R. (2013). Semantic preview benefit during reading. *J. Exp. Psychol. Learn. Mem. Cogn.* 40, 166–190. doi: 10.1037/a0033670

Jaeger, F. T. (2008). Categorical data analysis: away from ANOVA (transformation or not) and towards Logit Mixed Models. *J. Mem. Lang.* 59, 434–446. doi: 10.1016/j.jml.2007.11.007

Kidd, E. (2013). The role of verbal working memory in children's sentence comprehension: a critical review. *Top. Lang. Disord.* 33, 208–223. doi: 10.1097/TLD.0b013e31829d623e

Kidd, E., Brandt, S., Lieven, E., and Tomasello, M. (2007). Object relatives made easy: a cross-linguistic comparison of the constraints influencing young children's processing of relative clauses. *Lang. Cogn. Process.* 22, 860–897. doi: 10.1080/01690960601155284

Koster, C., Hoeks, J. C. J., and Hendriks, P. (2011). "Comprehension and production of subject pronouns," in *Production-Comprehension Asymmetries in Child Language*, eds A. Grimm, A. Müller, C. Hamman, and E. Ruigendijk (Berlin: De Guyter), 99–122.

Legendre, G., Barrière, I., Goyet, L., and Nazzi, T. (2011). "On the acquisition of implicated presuppositions: evidence from French personal pronouns," in *Selected Proceedings of the 4th Conference on Generative Approaches to Language Acquisition North America (GALANA),* eds M. Pirvulescu, M. C. Cuervo, A. T. Pérez-Leroux, J. Steele, and N. Strik (Somerville, MA: Cascadilla Proceedings Project), 150–162.

Legendre, G., and Smolensky, P. (2012). On the asymmetrical difficulty of acquiring person reference in French: production versus comprehension. *J. Logic Lang. Inform.* 21, 7–30. doi: 10.1007/s10849-011-9150-0

Lewis, R. L., and Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cogn. Sci.* 29, 375–419. doi: 10.1207/s15516709cog0000_25

Lewis, R. L., Vasishth, S., and Van Dyke, J. A. (2006). Computational principles of working memory in sentence comprehension. *Trends Cogn. Sci. (Regul. Ed.)* 10, 447–454. doi: 10.1016/j.tics.2006.08.007

Montgomery, J. W., and Evans, J. L. (2009). Complex sentence processing and working memory in children with specific language impairment. *J. Speech Lang. Hear. Res.* 52, 269–288. doi: 10.1044/1092-4388(2008/07-0116)

Montgomery, J. W., Magimairaj, B., and O'Malley, M. (2008). The role of working memory in typically developing children's complex sentence comprehension. *J. Psycholinguist. Res.* 37, 331–354. doi: 10.1007/s10936-008-9077-z

O'Grady, W. (2011). "Relative clauses: processing and acquisition," in *The Acquisition of Relative Clauses. Processing, Typology and Function*, ed. E. Kidd (Amsterdam: John Benjamins Publishing Company), 13–38. doi: 10.1075/tilar.8.03gra

Petermann, F., and Petermann, U. (2008). *Hamburg-Wechsler-Intelligenztest für Kinder IV*. Bern: Verlag Hans Huber.

Pitchford, N., and Mullen, K. (2003). The development of conceptual colour categories in pre-school children: influence of perceptual categorization. *Vis. Cogn.* 10, 51–77. doi: 10.1080/713756669

Pyykkönen, P., Matthews, D., and Järvikivi, J. (2010). Three-year-olds are sensitive to semantic prominence during online language comprehension: a visual world study of pronoun resolution. *Lang. Cogn. Process.* 25, 115–129. doi: 10.1080/01690960902944014

R Development Core Team. (2014). *R: A Language and Environment for Statistical Computing. R Foundation and Statistical Computing*. Vienna: R Development Core Team. Available at: http://www.R-project.org/

Recanati, F. (1993). *Direct Reference: From Language to Thought*. Oxford: Blackwell Publishers.

Rizzi, L. (1990). *Relativized Minimality*. Cambridge, MA: The MIT Press.

Rizzi, L. (2013). Locality. *Lingua* 130, 169–186. doi: 10.1016/j.lingua.2012.12.002

Roberts, L., Marinis, T., Felser, C., and Clahsen, H. (2007). Antecedent priming at trace positions in children's sentence processing. *J. Psycholinguist. Res.* 36, 175–188. doi: 10.1007/s10936-006-9038-3

Sekerina, I. A., Stromswold, K., and Hestvik, A. (2004). How do adults and children process referentially ambiguous pronouns. *J. Child Lang.* 31, 123–152. doi: 10.1017/S0305000903005890

Serratrice, L. (2013). The role of number of referents and animacy in children's use of pronouns. *J. Pragmat.* 56, 31–42. doi: 10.1016/j.pragma.2013.05.008

Sheppard, S. M., Walenski, M., Love, T., and Shapiro, L. P. (2015). The auditory comprehension of Wh-questions in aphasia: support for the intervener hypothesis. *J. Speech Lang. Hear. Res.* doi: 10.1044/2015_JSLHR-L-14-0099 [Epub ahead of print].

Siegmüller, J., Kauschke, C., van Minnen, S., and Bittner, D. (2010). *Test zum Satzverstehen von Kindern: Eine profilorientierte Diagnostik der Syntax*. Munich: Urban & Fischer/Elsevier.

Song, H.-J., and Fisher, C. (2005). Who's "she"? Discourse prominence influences preschoolers' comprehension of pronouns. *J. Mem. Lang.* 52, 29–57. doi: 10.1016/j.jml.2004.06.012

Song, H.-J., and Fisher, C. (2007). Discourse prominence effects on 2.5-year-old children's interpretation of pronouns. *Lingua* 117, 1959–1987. doi: 10.1016/j.lingua.2006.11.011

Spenader, J., Smits, E.-J., and Hendriks, P. (2009). Coherent discourse solves the pronoun interpretation problem. *J. Child Lang.* 36, 23–52. doi: 10.1017/S0305000908008854

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., and Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language recognition. *Sciene* 268, 1632–1634. doi: 10.1126/science.7777863

Trueswell, J. (2008). "Using eye movements as a developmental measure within psycholinguistics," in *Developmental Psycholinguistics: On-line Methods in Children's Language Processing*, eds I. Sekerina, E. M. Fernández, and H. Clahsen (Amsterdam: John Benjamins Publishing Company), 73–96. doi: 10.1075/lald.44.05tru

Trueswell, J. C., Papafragou, A., and Choi, Y. (2011). "Referential and syntactic processes: what develops?" in *The Processing and Acquisition of Reference*, eds E. Gibson and N. J. Pearlmutter (Cambridge, MA: The MIT Press), 65–108. doi: 10.7551/mitpress/9780262015127.003.0004

van Dyke, J. A., and McElree, B. (2006). Retrieval interference in sentence comprehension. *J. Mem. Lang.* 55, 157–166. doi: 10.1016/j.jml.2006.03.007

van Dyke, J. A., and McElree, B. (2011). Cue-dependent interference in comprehension. *J. Mem. Lang.* 65, 247–263. doi: 10.1016/j.jml.2011.05.002

van Rij, J., van Rijn, H., and Hendriks, P. (2010). Cognitive architectures and language acquisition: a case study in pronoun comprehension. *J. Child Lang.* 37, 731–766. doi: 10.1017/S0305000909990560

van Rij, J., van Rijn, H., and Hendriks, P. (2013). How WM load influences linguistic processing in adults: a computational model of pronoun interpretation in discourse. *Top. Cogn. Sci.* 5, 564–580. doi: 10.1111/tops.12029

Wagers, M. W., and Phillips, C. (2014). Going the distance: memory and control processes in active dependency construction. *Q. J. Exp. Psychol.* 67, 1274–1304. doi: 10.1080/17470218.2013.858363

Warren, T., and Gibson, E. (2002). The influence of referential processing on sentence complexity. *Cognition* 85, 79–112. doi: 10.1016/S0010-0277(02)00087-2

Warren, T., and Gibson, E. (2005). Effects of NP type in reading cleft sentences in English. *Lang. Cogn. Process.* 20, 751–767. doi: 10.1080/01690960500051055

Weighall, A. R., and Altmann, G. T. M. (2011). The role of working memory and contextual constraints in children's processing of relative clauses. *J. Child Lang.* 38, 579–605. doi: 10.1017/S0305000910000267

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2015 Haendler, Kliegl and Adani. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

METHODS
published: 14 June 2016
doi: 10.3389/fpsyg.2016.00873

Using the Visual World Paradigm to Study Retrieval Interference in Spoken Language Comprehension

Irina A. Sekerina[1, 2, 3]*, *Luca Campanelli*[4] *and Julie A. Van Dyke*[5]

[1] Department of Psychology, College of Staten Island, City University of New York, Staten Island, NY, USA, [2] Linguistics Program, The Graduate Center, City University of New York, NY, USA, [3] Neurolinguistics Laboratory, National Research University Higher School of Economics, Moscow, Russia, [4] Speech-Language-Hearing Sciences, The Graduate Center, City University of New York, NY, USA, [5] Haskins Laboratories, New Haven, CT, USA

The cue-based retrieval theory (Lewis et al., 2006) predicts that interference from similar distractors should create difficulty for argument integration, however this hypothesis has only been examined in the written modality. The current study uses the Visual World Paradigm (VWP) to assess its feasibility to study retrieval interference arising from distractors present in a visual display during spoken language comprehension. The study aims to extend findings from Van Dyke and McElree (2006), which utilized a dual-task paradigm with written sentences in which they manipulated the relationship between extra-sentential distractors and the semantic retrieval cues from a verb, to the spoken modality. Results indicate that retrieval interference effects do occur in the spoken modality, manifesting immediately upon encountering the verbal retrieval cue for inaccurate trials when the distractors are present in the visual field. We also observed indicators of repair processes in trials containing semantic distractors, which were ultimately answered correctly. We conclude that the VWP is a useful tool for investigating retrieval interference effects, including both the online effects of distractors and their after-effects, when repair is initiated. This work paves the way for further studies of retrieval interference in the spoken modality, which is especially significant for examining the phenomenon in pre-reading children, non-reading adults (e.g., people with aphasia), and spoken language bilinguals.

Keywords: memory retrieval, spoken language comprehension, visual world paradigm, eye-tracking, cleft sentences

OPEN ACCESS

Edited by:
Matthew Wagers, University of California, Santa Cruz, USA

Reviewed by:
Kaili Clackson, University of Cambridge, UK
Jennifer E. Mack, Northwestern University, USA

***Correspondence:**
Irina A. Sekerina
irina.sekerina@csi.cuny.edu

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: *31 January 2016*
Accepted: *26 May 2016*
Published: *14 June 2016*

Citation:
Sekerina IA, Campanelli L and Van Dyke JA (2016) Using the Visual World Paradigm to Study Retrieval Interference in Spoken Language Comprehension. Front. Psychol. 7:873. doi: 10.3389/fpsyg.2016.00873

INTRODUCTION

Memory processes are crucial for language comprehension, especially the ability to store linguistic constituents and retrieve them later (perhaps much later) to combine with new information. For example, it is quite common for linguistically dependent information to be separated by a considerable distance. An example of such a construction is in (1), where a dependent constituent, *the girl,* is separated from the verb *smelled* by two relative clauses.

(1) The girl who walked with the cute little boy that wore the striped shirt smelled the flowers.

Consequently, a clear understanding of the memory processes that support accurate comprehension is critical to any psycholinguistic model of language use. In this paper, we

present a novel application of the Visual World eye-tracking Paradigm (VWP; Altmann, 2004; Trueswell and Tanenhaus, 2005) for studying these memory retrieval processes in spoken language comprehension. The particular novelty of the current study is to test the VWP against the logic of the dual-task paradigm, which has been used previously (Van Dyke and McElree, 2006, 2011) as a means of explicitly manipulating the contents of memory, and arguing specifically for retrieval interference (as opposed to encoding interference) in processing of spoken sentences with syntactic dependencies.

The Cue-Based Retrieval Theory (CBRT)

Several theories have been proposed to explain why establishing memory-dependent linguistic relationships as in (1) is challenging, even for monolingual adult speakers (see Levy et al., 2013 for a review). One of the most cited is *the Cue-Based Retrieval Theory* (*CBRT;* Gordon et al., 2002; McElree et al., 2003; Van Dyke and Lewis, 2003; Lewis et al., 2006; Van Dyke and McElree, 2006, 2011; Van Dyke and Johns, 2012; Van Dyke et al., 2014) which is grounded in a large body of empirical research pointing to a severely limited active memory capacity even for skilled monolingual readers, accompanied by a fast, associative retrieval mechanism that uses cues to access memory directly (reviewed in McElree, 2006). A central prediction of the CBRT is that interference effects will arise whenever retrieval cues necessary for identifying a distant dependent are ambiguous. It is this interference that creates comprehension difficulty. For example, in (1) the verb *smelled* selects for an animate subject, and there are two such NPs that fit these cues (*the girl* and *the cute little boy*). The second NP serves as a distractor for retrieving the target subject, resulting in longer reading times at *smelled* and lower accuracy to comprehension questions (Van Dyke, 2007).

In order to distinguish the retrieval account of the CBRT from accounts emphasizing costs associated with storing multiple similar items (e.g., Gordon et al., 2002). Van Dyke and McElree (2006) directly manipulated the relationship between the contents of memory and the cues available at retrieval. To do this, they utilized a dual task paradigm in which they asked participants to read written sentences like (2) in a phrase-by-phrase manner while performing a simultaneous memory load task.

(2) It was the button that the maid who returned from vacation spotted in the early morning.

On high memory load trials, participants were asked to remember a list of three words (i.e., KEY-PEN-EARRING) and then read the sentence in (2). The manipulation of interest was when the verb *spotted* was replaced with *sewed*; in the *spotted* case, all of the words from the memory list could serve as the verb's object, but only a *button* is *sew*-able. The authors observed increased interference effects from the words in the memory list in the form of longer reading times at the verb *spotted* (578 ms), but not at the verb *sewed* (540 ms). This difference disappeared when the memory list was not presented (564 vs. 567 ms, respectively), demonstrating that the reading time difference was *not* simply related to a difference in the semantic association between the verb and the clefted NP. Interference was due to the match between the distractors in the memory list and the semantic retrieval cues from the verb that specify the target referent (*the button*), i.e., an object that can be spotted.

This type of interference has now been demonstrated not only in measures of reading speed, but also in comprehension accuracy and grammaticality judgments, and in a variety of linguistic constructions; it takes place whether the intruders occur before (proactive interference) or after (retroactive interference) the retrieval target (Van Dyke and McElree, 2006, 2011; Martin and McElree, 2009); whether the intruder is syntactically, semantically, or referentially similar (Gordon et al., 2002; Van Dyke and Lewis, 2003); or even when the intruder is unlicensed in the grammatical construction (Van Dyke, 2007; Vasishth et al., 2008; Martin et al., 2012). Finally, sensitivity to interference appears to be modulated by individual differences in cognitive abilities (Van Dyke et al., 2014).

Written vs. Spoken Modality

The evidence associated with the CBRT is robust, but so far, it has been restricted to the reading modality. Hence, the role of retrieval interference in *spoken language comprehension* remains unknown. Speech contains a variety of spoken cues, but there is little evidence about how spoken cues are considered by the retrieval mechanism, and what priority they may receive vis-à-vis other cues (e.g., semantic, syntactic). This issue is critical because speech cues play a primary role in memory encoding (Baddeley, 1966, 2012; Liberman et al., 1972), creating the possibility that input modality may be an important means for modulating effects of retrieval interference.

Modality effects have been found elsewhere in the literature. Using a self-paced listening paradigm, contrasted with a self-paced reading paradigm, older adults have been found to take longer to read relative clauses than to listen to them (Waters and Caplan, 2005; Caplan et al., 2011). Further, a study with cleft sentences of the sort investigated here (DeDe, 2013) examined whether input modality and syntactic complexity interact in healthy younger and older adults and people with aphasia. As in the studies conducted by Caplan and colleagues, DeDe found that the processing time for healthy controls was longer in the self-paced reading experiment than in the self-paced listening one, and this effect was only observable on the verb. She concluded that "...listening may exert fewer processing demands because it is a more natural and over-practiced skill than reading" (p. 11).

In contrast, neuroimaging studies have found small, but consistent modality differences in word (Chee et al., 1999) and sentence processing (Michael et al., 2001; Rüschemeyer et al., 2006), with listening being more resource-demanding. For example, Michael and colleagues compared subject and object relative clauses and found increased hemodynamic response in listening to object relatives in the auditory modality, but not while reading. A possible explanation for this difference, offered by Chee and colleagues, points to the greater reliance on working memory in spoken language comprehension (but see Van Dyke et al., 2014 for an alternative view). Hence, examining retrieval interference in the spoken modality and the specific role of speech cues is an important means of advancing the CBRT.

APPLYING THE VWP TO STUDY RETRIEVAL INTERFERENCE IN SPOKEN LANGUAGE COMPREHENSION

In all of the aforementioned studies that tested the CBRT, sentences with filler-gap dependencies were presented to participants in the written form and, therefore, effects of distractors were indirectly inferred from differences in reading times at the verb (*spotted* took longer to read than *sewed*), contrasted with similar conditions containing no extra-sentential distractors. The current study seeks to determine whether retrieval interference effects can be found in the spoken modality.

The Visual World Paradigm

The Visual World eye-tracking Paradigm (VWP) is well-suited for addressing these questions because VWP experiments measure overt looking to multiple, clearly separable referents (represented as pictures or real objects) called *Target, Competitor*, and *Distractors*. Hence, it provides a straightforward measure of competition between referents while listening. For example in **Figure 1**, the key manipulation involves the relationship between the four pictures and the main verb in the spoken sentence. As in the original study (Van Dyke and McElree, 2006), we expect that the semantic properties of the verb will guide the search for a filler for the gapped object (trace position), so that when the verb is *sewed* there will be more looks to the button than to any other picture, whereas the verb *spotted* will support looks to any of the four pictures, which are all objects that could be spotted.

This prediction is similar to classic findings in the VWP literature, where properties of a verb enable participants to anticipate what will be referred to post-verbally (e.g., Kamide et al., 2003; Huettig and Altmann, 2005). For example, Altmann and Kamide (2004) presented participants with four pictures (i.e., a cake, toy train, toy car, and ball) while requiring them to listen to either of the two spoken sentences, (3a) or (3b):

(3) a. The boy will eat the cake.
b. The boy will move the cake.

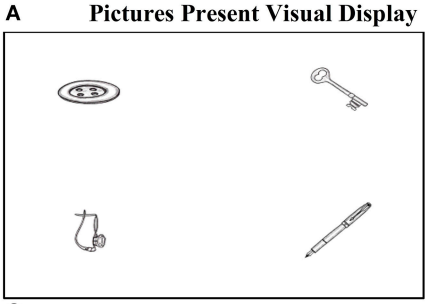
They found that the participants were much more likely to launch eye movements to the cake in (3a) than (3b) and that this happened before the onset of the word *cake*. They interpreted these results as evidence that semantic properties of the verb are used immediately (and incrementally) to guide subsequent integrative processing.

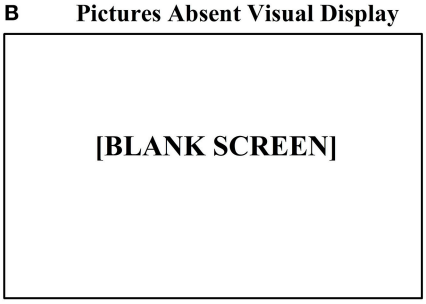
There are several previous VWP studies that investigated processing of memory-dependent linguistic relations in sentences with syntactic dependencies. In these studies, the visual display always included four pictures of the referents explicitly named in the preamble and experimental instruction. Sussman and Sedivy (2003) tested unimpaired adults and established that in oblique object *Wh*-questions (e.g., *What did Jody squash the spider with*?), the *wh*-filler *what* triggered an increase in anticipatory fixations to the potential argument of the verb (i.e., the spider) during the verb despite the fact that the gap was filled. At the preposition, the participants quickly switched to the correct referent (i.e., the shoe). Dickey et al. (2007) simplified the object *Wh*-questions used in Sussman and Sedivy's experiment by removing the oblique object (e.g., *Who did the boy kiss that day at school?*) and compared eye movements of control adults with those of people with aphasia who had difficulties with comprehension of sentences with syntactic dependencies. Based on eye-movement patterns of people with aphasia in the incorrectly answered questions, they argued that their comprehension errors were caused by late-arising competition between the target object referent (e.g., the girl) and the competitor subject (e.g., the boy).

However, neither Sussman and Sedivy (2003) nor Dickey et al. (2007) explained their results in terms of retrieval interference. In contrast, Sheppard et al. (2015) specifically tested *the intervener hypothesis* in search for an explanation of comprehension failure in people with aphasia when they process two types of object *Wh*-questions (e.g., *Who* vs. *Which mailman did the fireman push yesterday afternoon?*). To ensure the felicity of the *which*-questions, the 4-referent display was replaced with an action picture in which one fireman and two mailmen were depicted in two simultaneous pushing events. The results suggested that the more people with aphasia looked at the incorrect mailman (i.e., the intervener) the more likely they were to answer the question, in particular, the *which*-question, incorrectly. A similar explanation was proposed by Clackson et al. (2011) in accounting for eye movements of adults and children in sentences with referentially ambiguous personal pronouns (e.g., *He [Peter] watched as Mr. Jones bought a huge box of popcorn for him.*.). Children were especially prone to look more at the gender-matched referent (e.g., Mr. Jones) in the position intervening between the pronoun (e.g., *him*) and its accessible antecedent (e.g., *Peter*) even though this intervener is ruled out by the Binding theory.

Our current application of the VWP provides a more direct way of testing retrieval interference in processing of sentences with syntactic dependencies. All of the previous studies required referent selection based on a forced choice between two referents explicitly named in the spoken materials, i.e., the target and competitor. In the 4-referent set-up employed by Sussman and Sedivy (2003), Dickey et al. (2007), and Clackson et al. (2011), the remaining 2 referents (i.e., a distractor and a location) attracted very few looks, thus, effectively restricting referential choice to two. In addition to the fact that all 4 distractor referents were explicitly named in the spoken context, the intervener was placed in the sentence between the filler and gap which increased their salience and availability during retrieval of the filler at the verb.

The case study described in this article employed the dual-task paradigm (Van Dyke and McElree, 2006), in which every one of the three distractor referents was a legitimate semantic intruder that was *outside* the spoken sentence. Hence, any interference from the distractors suggests that information contained within memory, but not part of the sentence itself, impacts successful retrieval of the actual target. This has important ramifications for the specification of the type of retrieval mechanism (i.e., one that matches to all contents of memory simultaneously, as in a global matching mechanism (e.g., Clark and Gronlund, 1996) or else a retrospective serial search that entertains each item in memory individually. The former predicts that all distractors





C

a. Interfering:

It was the button that the maid who returned from vacation *spotted* in the early morning.

b. Non-Interfering:

It was the button that the maid who returned from vacation *sewed* in the early morning.

FIGURE 1 | Sample experimental conditions. (A) Visual display for Picture Memory List Present, **(B)** Visual display for Picture Memory List Absent, **(C)** Two types of spoken sentences.

should receive increased looks when they match retrieval cues from the verb, while the latter predicts that only the target referent (which is the most recent) would receive looks from the verb. In addition, interference effects from extra-sentential distractors suggest that sentence processing utilizes the same memory capacity as that used for short-term memory, contrary to accounts that would give sentence processing a separate memory capacity (e.g., Caplan and Waters, 1999).

Using the VWP for studying retrieval interference in spoken language comprehension brings an additional advantage in that this method removes potential confounds related to reduced reading skill or difficulty comprehending complex task instructions, concerns which are paramount when investigating comprehension ability in linguistically diverse populations, such as children, bilingual and second language learners, and participants with language impairments. Instead, the VWP provides a naturalistic way to assess language processing while participants listen to verbal input and look at visual arrays. In addition, it could be employed in a passive listening mode that does not require verbal, gestural, or motor responses, making it amenable for use with older individuals or persons with aphasia (Hallowell et al., 2002; Ivanova and Hallowell, 2012).

"The Blank Screen" Paradigm in the VWP

The classic VWP experiments with spoken sentences found anticipatory looks toward an object when the verb precedes it (Kamide et al., 2003; Huettig and Altmann, 2005) demonstrating that the verb's selectional restrictions activate its argument structure. The latter, in its turn, drives looks to the referent that is named by the noun in post-verbal position. However, looks could be crucially dependent on the co-occurrence of linguistic input and the overt presence of the referent's picture. To counter this argument, Altmann (2004) demonstrated that the physical presence of the pictures was not necessary. Listeners still moved their eyes to the location of a previously displayed object even when the object was no longer present while they listened to the spoken sentence. This method received the name of the "blank screen" paradigm. Although the proportion of looks using this method was relatively low in absolute terms (16%; Altmann and Kamide, 2004, Figure 11.1), Altmann and Kamide interpreted these results as evidence that it is the mental representations of the objects held in memory that are activated by the verb's semantics. Therefore, eye movements in the VWP were shown to reflect the mental world, and not just visual attention in the form of iconic memory.

Because this method has particular theoretical significance in the VWP literature, we chose to implement the blank screen paradigm as a potential analog of the Memory-Load condition of Van Dyke and McElree. We hoped this would allow us to determine the extent of interference from visually presented distractors: If interference from *absent* distractors were observed, this would suggest that semantic interference from present distractors is not merely contingent on the current visual scene, but related to accessing all matching memory representations, whether currently active or not. As it turned out, firm conclusions on this point were frustrated by a methodological confound. Hence, although we present these results, our conclusions are drawn primarily from the Pictures Present conditions in our design.

In what follows below, we present a VWP implementation of the Van Dyke and McElree (2006) study, which examined how semantic properties can be used to guide retrieval of previously occurring constituents. Specifically, in (2), the grammatical encoding of the clefted NP makes it unambiguously identifiable as the object of a later occurring verb however, there is no prospective information about the semantic relationship between that object and the verb. Thus, any difference in looks to the target in the Interfering (e.g., *spotted*) vs. Non-Interfering (e.g., *sewed*) conditions has to occur only once the verb is heard (or after) and must be attributed to interference driven by the verb's semantic cues. The prediction of fewer looks to the correct target picture (*button*) in the Interfering conditions compared to the

Non-Interfering conditions is analogous to the finding in Van Dyke and McElree, where semantically similar distractors *outside* the sentence produced inflated reading times at the point of integrating the verb with its direct object.

A CASE STUDY: RETRIEVAL INTERFERENCE IN SPOKEN LANGUAGE COMPREHENSION

Participants

Twenty-four undergraduate students from the College of Staten Island participated in this study for credit as one of the requirements for an introductory psychology class. All participants (7 men, mean age = 21.4) identified themselves as native English speakers. This study was carried out in accordance with the ethical principles of psychologists and code of conduct of the American Psychological Association and was approved by the Institutional Review Board of the College of Staten Island. All participants gave written informed consent in accordance with the Declaration of Helsinki.

Materials

Each experimental item was realized as one of four conditions in a 2 × 2 (Interference × Picture) factorial design. The interference manipulation was identical to that in the original study by Van Dyke and McElree (2006), but the objects from the memory set were presented as pictures, and not as words: in the interfering condition, all pictured items could serve as the object of the main verb (e.g., *spotted*) in the sentence. For the corresponding non-interfering condition, the same pictures were presented but the main verb was changed in the spoken recording (e.g., *sewed*) so that only the clefted NP made sense as its object (See **Figure 1**). Each picture occupied one of the four quadrants on the stimuli computer monitor, and the clefted NP picture was evenly rotated through each quadrant. For the picture manipulation, pictures remained on the screen while the sentence played (Present) or were removed (Absent, the blank screen paradigm) after the participant named them.

As in Van Dyke and McElree's (2006) Memory Load conditions, the picture memory list was always presented to participants first—prior to reading the spoken sentence, as in (2). The sentence was always followed by a *yes/no* comprehension question, and then, finally, they were asked to recall the four pictures from the memory list. The four steps of the procedure were each crucial to the implementation of the memory interference paradigm. The picture memory list established potential distractors in the comprehension context, the sentence presented the main language processing task, the comprehension questions ensured that participants would attend to the sentence (rather than ignore it in favor of focusing all their attention on the memory task), and the recall task ensured that they would work to keep the pictures from the memory list within their active memory. Participants were explicitly told to do their best on each of the individual tasks.

An important dimension of exploring retrieval interference in the spoken modality is the possible effect that prosodic cues may play in mediating retrieval difficulty. It is currently not known whether or not these cues are considered by the retrieval mechanism, and what priority they may receive vis-à-vis other cues (e.g., semantic and syntactic). Because of this, we decided to employ neutral prosody so as to establish a baseline for whether the expected effects would manifest in eye-movement patterns. Although clefted constructions such as (2) often occur with a stress contour, there is no information about whether individual readers assign such a contour when they read them silently. This is significant because the original study by Van Dyke and McElree (2006) employed self-paced reading, which may have discouraged the natural assignment of implicit prosody. Thus, we considered the use of neutral prosody to be the best approximation to the reading conditions in the original study.

The 28 sets of experimental items were selected from the original 36 object cleft sentences of Van Dyke and McElree's (2006) self-paced reading experiment based on how well the items in the memory lists could be depicted. There were also 56 filler items of two types: eighteen subject cleft sentences (e.g., *It was the son who was wild that smashed the lego tower that nearly reached the ceiling.*—Picture Memory List: ROSE, POMEGRANATE, SICKLE, VIOLIN), and 38 non-clefted sentences (e.g., *The sailors knew that the treasure enticed the pirate on the hijacked ship*—Picture Memory List: HOUSE, STAR, ROBE, FAIRY). Pictures for the filler sentences were selected randomly; one half was presented with pictures, and the other half was paired with a blank screen. There were also five practice items with feedback. Four lists were constructed using the Latin Square design consisting of five practice, 28 experimental (7 items per condition) and 56 filler items in such a way that each experimental item was both preceded and followed by one of the fillers. Thus, all experimental items were separated by two fillers. Six participants were randomly assigned to each of the four lists, containing 89 trials in total.

The 356 pictures (89 trials × 4 pictures) were selected from the electronic database of object and action pictures created in the Neurolinguistics Laboratory (head: Dr. Olga V. Dragoy) at the Higher School of Economics (Moscow, Russia). The database is available online free of charge (http://stimdb.ru/) and contains black-and-white pictures normed on many dimensions (i.e., naming agreement, visual complexity, age of acquisition, frequency, and familiarity; Akinina et al., 2015).

All spoken sentences (experimental and filler) were recorded by a female native speaker of American English at a sample rate of 22,050 Hz. Every effort was made to pronounce them with neutral prosodic intonation to eliminate the contribution of special prosodic cues associated with cleft sentences in English, i.e., a fall-rise pitch accent on the clefted NP (Hedberg, 2013) and a prosodic break after the clefted NP indicating phrasal boundary, during retrieval. However, after data collection we discovered that this goal was not met: experimental sentences were recorded in two different sessions, which resulted in subtle perceptual and prosodic differences between the interfering and non-interfering conditions. We discuss this methodological error later. Speaking rate was slightly slower than is heard in everyday casual speech, due to efforts to enunciate each word; see Appendix B in Supplementary Material for example recordings.

The comprehension questions were designed following the method of Van Dyke and McElree (2006). Two thirds of the questions for the experimental items (19 out of 28) were about the subordinate clause (e.g., Example 4: *It was the cigarette that the criminal who robbed the electronics store smoked/sought in the dark alley*. Question: *Did the criminal rob a liquor store?*) and one third (9 items) were about the main clause with the clefted NP (e.g., for Example 2 the question was, *Was it the maid who was on vacation?*).

The pictures, spoken sentences, and comprehension questions for all 28 experimental items and a sample of eight representative fillers, as well as the two auditory versions of example (2) in both the interfering and non-interfering conditions, are provided in Audios 1 and 2 in the Supplementary Material.

Procedure

The experiment was controlled by DMDX software (Forster and Forster, 2003), with the game pad serving as the interface device. Participants were seated in front of a 17-inch Dell laptop (resolution of 1024 × 768 pixels) at a viewing distance of ∼60 cm. On each trial, participants first saw the four-picture memory list (**Figure 1A**), with each picture centered in one of the four 350 × 350-pixel quadrants of the display. Each of the four images subtended about 11 degrees of visual angle. Participants were asked to label the pictures in any order using just one word and then press the "Yes" button on the game pad to listen to the auditory sentence (**Figure 1C**, a-b). Specific picture labels were not sought in this experiment, hence no feedback was given in this phase. In the pictures present condition, participants continued to look at the pictures while listening to the sentence (**Figure 1A**); in the pictures absent conditions, they looked at the blank screen (**Figure 1B**). An auditory comprehension question automatically followed the sentence (e.g., *Was it the maid who was on vacation?*) and was answered by pressing either the "Yes" or "No" button on the game pad. As soon as the response was provided, DMDX presented a written reminder for the participants to recall the four pictures from the memory list (i.e., *Now recall the four pictures*), and their voice responses were recorded with the help of a microphone connected to a digital SONY DSR-30 video tape-recorder. Participants were asked to recall all of the pictures in any order, but were encouraged not to belabor the recall if they couldn't remember them.

The video tape-recorder was connected to the ISCAN ETL-500 remote eye-tracking system that collected participants' eye movements. Eye movements were sampled at a rate of 30 times per second. Prior to the experiment each participant underwent a short calibration procedure. The experiment was conducted in one session and lasted ∼1 h.

Statistical Analysis

Mixed-effects logistic regression was used to examine three measures: picture recall accuracy, comprehension question accuracy, and eye movement data. Mixed-effects modeling allows us to account for the clustered nature of the data, with responses nested within participants and items; furthermore, it makes it possible to examine variability within and between participants and items and is flexible in handling missing data (Raudenbush and Bryk, 2002). All models included crossed random intercepts for participants and items (Baayen et al., 2008). Random slopes for the-within-subjects independent variables were examined but not retained in any of the analyses, either because of convergence failure or because the random slopes did not improve the model fit.

Between-subjects outliers were trimmed following a 2-stage procedure: first, for each experimental condition we excluded subjects with average proportion of fixations more than 2.5 SD below or above the grand mean. Second, for each model, we examined the level-2 residuals and we re-fitted the models without observations with absolute standardized residuals greater than 2.5. This 2-stage procedure never led to the exclusion of more than 3% of the data.

Missing values due to equipment malfunctioning and track loss constituted 0.4 and 4.6% of the data, respectively. Data were analyzed with R version 3.1.2 (R Core Team, 2014) using the glmer function from the *lme4* package, version 1.1-7 (Bates et al., 2014).

RESULTS AND DISCUSSION

Recall of Pictures

In the beginning of the trial, participants were asked to label each picture in the 4-item memory list using one word, but they were free to choose any appropriate word. For example, they could choose to label a picture depicting a rug as "*carpet*." No feedback or corrections were provided except in the practice trials. Accuracy of recall of pictures was scored based on the actual number of pictures recalled for each trial and ranged from zero (no pictures recalled) to 100% (all four pictures recalled). Any order of recall was allowed as long as the pictures were labeled the way the participant labeled them in the beginning of the trial (e.g., saying *rug* when the picture was labeled as *carpet* was counted as an error). The top row of **Table 1** shows the mean correct recall of the pictures as a factor of Interference and Picture.

Mixed-effects logistic regression analysis was used to examine the effect of Interference and Picture on accuracy of picture recall. Results showed a significant effect of Picture, such that the recall of the pictures was significantly better in the Pictures Present than in the Pictures Absent conditions, 91.6 vs. 85.8% (cf. **Table 2**, left panel). There was no effect of Interference and no interaction. Interestingly, the recall of the pictures in our experiment was higher than that for written word memory lists in Van Dyke and McElree's (2006) experiment (non-interfering condition: 80%, interfering condition: 78% in that study) and this was true even in the Pictures Absent condition. We interpret this as evidence that visually presented items have increased salience in memory as compared to verbally encoded memory words. This could possibly be explained by the difference in encoding modality: an auditorily presented sentence interferes less with memory for visually encoded stimuli. It is also possible that recall was increased because participants had both a visual and verbal encoding of the stimuli (Nelson and Brooks, 1973; Snodgrass and McClure, 1975; Paivio, 1986).

TABLE 1 | Accuracy of recall of pictures and comprehension questions, % (SD).

	Interfering, Pictures Present	Interfering, Pictures Absent	Non-Interfering, Pictures Present	Non-Interfering, Pictures Absent
Picture recall	91.4 (8.17)	86.3 (12.8)	91.7 (7.19)	85.3 (12.0)
Comprehension question	32.1 (20.3)	29.2 (25.8)	35.1 (17.4)	35.1 (17.4)

TABLE 2 | Accuracy of recall of pictures and comprehension questions: Summary of mixed-effects logistic regression analyses (fixed effects only).

	Picture recall				Comprehension questions			
Variable	**Estimate**	**SE**	***z***	***p***	**Estimate**	**SE**	***z***	***p***
(Intercept)	2.108	0.284	7.417	<0.001	−0.775	0.265	−2.922	0.003
Interference	0.527	0.393	1.342	0.180	−0.283	0.260	−1.087	0.277
Picture	**0.969**	**0.424**	**2.286**	**0.022***	0.048	0.254	0.189	0.850
Interference × Picture	−0.754	0.613	−1.230	0.219	0.018	0.366	0.050	0.960

*$p < 0.05$.

Comprehension Question Accuracy

Accuracy of responses to the comprehension questions as a factor of Interference and Picture was low overall, 32.9% (see **Table 1**, bottom row.) Results of mixed-effects logistic regression analysis of accuracy showed no significant effects (see **Table 2**, right panel). This is consistent with the results in Van Dyke and McElree (2006), however despite no significant effects the participants in that study had much higher accuracy levels (87% in the Non-interfering condition vs. 83% in the Interfering condition, a statistically significant difference). We note that this low accuracy was not due to our participants' overall level of performance in the experiment—their overall high picture recall (88.7%) confirms that they did pay attention. One possible explanation for the difference between the current results and the Van Dyke and McElree results is that the latter used the self-paced reading method which allows participants to read at their own pace. This self-controlled, and likely slower, presentation rate affords participants additional time for encoding and/or deciphering the meaning of the sentence, which in turn positions them to do better on the comprehension questions. In contrast, the spoken sentence passes quickly in the listening paradigm used here, and together with memorizing the pictures, this may have made the task more difficult. This is consistent with other findings showing less accurate comprehension in the auditory modality compared with comprehension of the same sentences in the written modality (Johns et al., 2015). Another possibility, suggested by our comparatively higher recall accuracy, is that participants traded off attention during sentence reading with attention to the recall task. We discuss this further below.

Eye Movements

The spoken sentences were divided into four regions for purposes of statistical analysis of eye movements: three sentence regions illustrated in (4) and one second of silence following the end of the sentence. The actual durations of each ROI in individual sentences varied because of differences in lexical items that constituted the experimental items. Each ROI was constructed around the specific onsets and offsets of individual items, but in the time course figures (**Figure 2**), the vertical dashed lines are aligned with the average onsets of the 4 ROIs.

(4)	Region 1 Clefted NP-that-Subject-Relative Clause (RC)	Region 2 Verb	Region 3 PP	Region 4 Silence
	It was the button that the maid who returned from vacation	sewed/ spotted/	in the early morning	

Eye movements were coded from the launch of a saccade to one of the 4 referent pictures present in the visual display and included a fixation that followed, as long as their combined duration was at least 100 ms. Looks in between the referents were coded as *else*, and looks off the screen were considered track loss and were removed from statistical analysis. Descriptive statistics and a graphical representation of the time course of the proportions of fixations to the target picture over all trials and all regions are reported in **Table 3** and **Figure 2**, respectively.

Region 1: Clefted NP-that-Subject-RC

Results of mixed-effects logistic regression analysis (**Table 4**; **Figure 3A**) showed a to-be-expected significant effect of Picture, such that the proportion of looks to the quadrant of the target picture was greater in the Pictures Present condition than in the Pictures Absent condition, where the eyes may be more apt to roam around the blank screen. Unexpectedly, we observed significant effects of Interference and an Interference × Picture interaction in this region, such that the proportion of looks to the target picture was greater in the Interfering than in the Non-Interfering condition when the pictures were present, and smaller in the Interfering than in the Non-Interfering condition when the pictures were absent. As both the linguistic and picture contexts were identical for all conditions prior to the verb, we trace this effect to the prosodic differences in the sentence recordings. *Post-hoc* analyses revealed that the average durations of the two

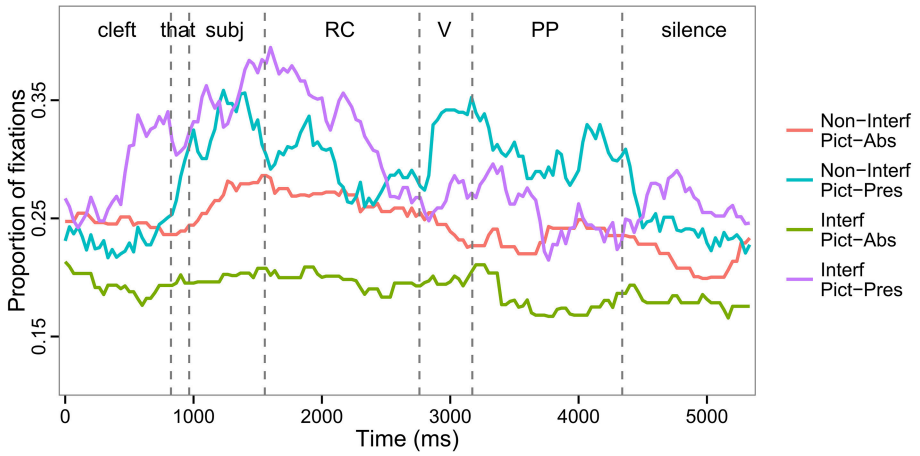


FIGURE 2 | Time course of fixation to the target picture as a function of Interference, Picture Memory List, and Region (all trials).

TABLE 3 | Proportions of looks to the target picture as a function of Interference, Picture, and Region, mean (SD).

Condition	Region 1	Region 2: V	Region 3: PP	Region 4: silence
Interfering, Pictures Absent	0.193 (0.131)	0.214 (0.166)	0.186 (0.130)	0.179 (0.144)
Interfering, Pictures Present	0.323 (0.118)	0.269 (0.162)	0.259 (0.115)	0.252 (0.151)
Non-Interfering, Pictures Absent	0.259 (0.127)	0.259 (0.170)	0.236 (0.132)	0.212 (0.139)
Non-Interfering, Pictures Present	0.281 (0.113)	0.294 (0.205)	0.306 (0.150)	0.243 (0.113)

components of the clefted NP—the cleft part (e.g., *it was the…*) and the target noun (e.g., *button*)—were consistently shorter in the Interfering than in the Non-Interfering condition, 309 ms vs. 343 ms [$t_{(26)} = 4.5496$, $p < 0.001$], and 323 ms and 346 ms, respectively [$t_{(26)} = 4.0341$, $p < 0.001$]. In addition, twice as many sentences in the Interfering condition (i.e., 15) than in the Non-Interfering condition (i.e., 8) had an extra prosodic break after the clefted NP (4) (// indicates a prosodic break). A representative pair of the actual recordings of the sentence types (e.g., Audio 1 and Audio 2) are available in the Supplementary Material.

(4) a. Interfering: It was the button // that the maid // who returned from vacation spotted…
b. Non-Interfering: It was the button // that the maid who returned from vacation sewed…

We speculate that despite the fact that we avoided pitch contours in an effort to keep prosody neutral, these differences created unintended prosodic cues that served to direct looks to the target noun in the Interfering, Pictures Present condition. The fewer looks to the target in the Interfering, Pictures Absent condition may also have resulted from the increased saliency of the target item, so that looking to the now-empty location of the target referent was not as necessary as it was for remembering the other, less salient referents[1]. Whether or not this account is correct, we emphasize that with respect to the Pictures Present condition, whatever bias drove these results went in the *opposite* direction to that predicted for the critical region (Region 2), where we expected looks to the target to *decrease* in the Interfering condition as compared to looks to competitors, which should increase in response to retrieval interference. Moreover, we conducted additional *post-hoc* analyses of the region after the target noun and found no additional prosodic differences between conditions. Hence, we are confident that results in Regions 2-3 are interpretable despite this methodological error. As for the Pictures Absent condition, this is a true confound. In order to better assess the presence of the Interference effect in relation to the Picture manipulation, we report the results of pair-wise comparisons for all future analyses.

Regions 2–3: Verb-PP

These two regions—Region 2 (Verb) and Region 3 (PP)—revealed the predicted pattern of results for the Interference manipulation. Region 2 is the critical region containing the verb that determines whether the pictured items are distractors or not (**Figure 3B**). We found a significant main effect of Interference, such that there were fewer looks to the target in the Interfering condition, where all of the pictured items could serve as the object of the main verb (e.g., they are all fixable in the example

[1]An alternative account is that the increased salience may reduce the need to look back during retrieval, perhaps because the target was already in a state of increased activation. Existing eye-movement evidence is not consistent with this interpretation however, as pre-activation of a target (greater looks to the target before hearing it) has not been related to reduced looks to the target upon hearing it or later (Altmann and Kamide, 1999; Coco et al., 2015). For example, Kukona et al. (2014) manipulated initial activation through making the target more or less predictable based on its relation to the verb. In the High Predictable condition ("eat cake") looks to the target were never less than looks to the target in the Low Predicable conditions ("move cake").

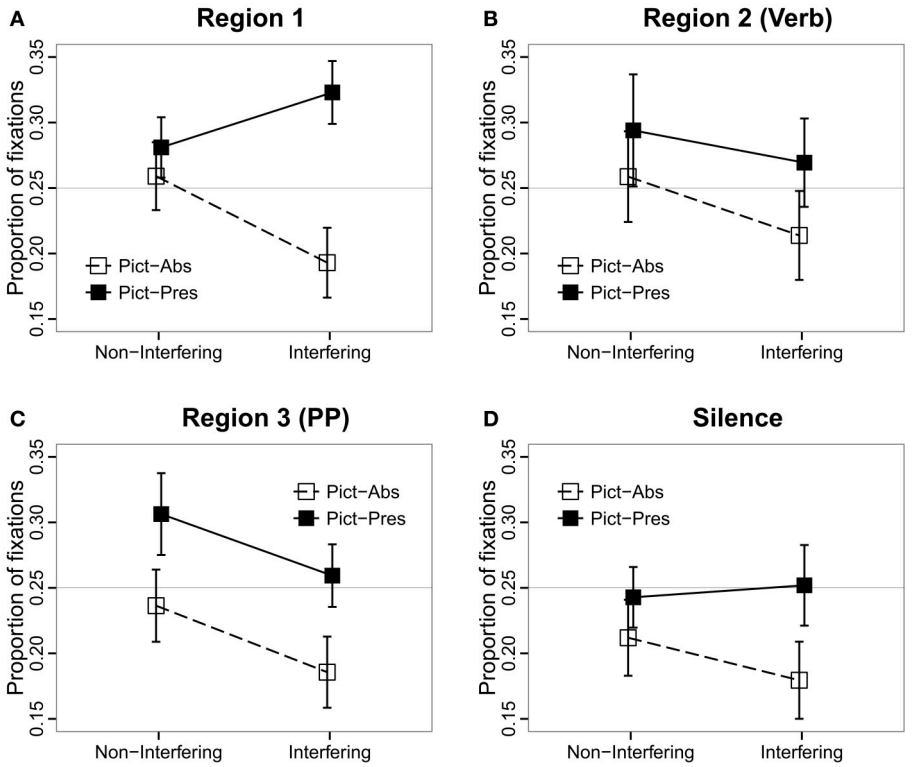


FIGURE 3 | Proportions of fixations to the target picture as a function of Interference, Picture Memory List, and Region. (A) Region 1 (Clefted NP-that-Subject-RC), **(B)** Region 2 Verb, **(C)** Region 3 (PP), **(D)** Region 4 Silence.

in **Figure 1**). We performed pair-wise comparisons to make sure that the main effect of Interference was not driven by the Pictures Absent condition. We found that the effect was significant in both the Pictures Present conditions (Tukey test: $z = -2.34, p < 0.05$) and in the Pictures Absent conditions (Tukey test: $z = -12.93$, $p < 0.001$). We also observed a significant effect of Picture, with a greater proportion of looks to the target picture in the Pictures Present condition as compared to the Pictures Absent condition. The interaction was not significant.

In Region 3, which contained the prepositional phrase, (**Figure 3C**) we observed the same pattern of results as in Region 2 (see **Table 4**). The proportion of looks to the target was greater in the Non-Interfering condition than in the Interfering condition. This effect obtained in the pairwise comparisons in both the Pictures Present conditions (Tukey test: $z = -6.88$, $p < 0.001$) and in the Pictures Absent conditions (Tukey test: $z = -6.36$, $p < 0.001$). Inspection of eye-movements in this time window (see **Figure 2**) suggests that this result was driven by looks to the target at the end of the sentence, and may reflect end-of-sentence wrap-up effects in which the participant is verifying his/her interpretation of the subject-verb dependency. As in the previous region, a significant effect of Picture was also observed, with more looks to the target in the Pictures Present condition.

Region 4: Silence

In the 1-s interval of silence following the end of the sentence (**Figure 3D**) the effect of Interference interacted with Picture, such that the proportion of looks to the target picture was comparable in the Interfering and Non-Interfering conditions when the pictures were present (Tukey test: $z = 0.63$, $p = 0.78$), and smaller in the Interfering than in the Non-Interfering condition when the pictures were absent (Tukey test: $z = -4.2$, $p < 0.001$). Visual inspection of these effects (**Figure 2**) suggests that the absence of an Interference effect in the Pictures condition, as compared to the significant Interference effect detected in the previous sentence regions, could be attributed to a proportional increase in looks to the target picture toward the end of the sentence for the Interfering conditions. We suggest that this effect can be associated with a repair process invoked when listeners realize they have constructed an incorrect interpretation due to interference from distractors. Similar late effects of semantic interference vis-à-vis retrieval cues have been observed in reading times (Van Dyke, 2007) and in BOLD signal during fMRI (Glaser et al., 2013).

Correct vs. Incorrect Trials

We performed a secondary analysis in which we separated the trials for which the comprehension questions were answered correctly from the ones with the incorrectly answered comprehension questions to assess the role of low accuracy on our results. **Figure 4** presents the time course of fixations for both subsets of trials; **Table 5** presents results of mixed-effect modeling. We observed a total of 219 correct trials, resulting in 33,356 total fixations; there was an average of 2.3 items per

TABLE 4 | Proportions of looks to the target picture: Summary of mixed-effects logistic regression analyses by Region (fixed effects only).

Region	Variable	Estimate	SE	*z*	*p*
Region 1 (clefted NP+that+ Subject+RC)	(Intercept)	−1.167	0.152	−7.699	<0.001
	Interference	**−0.405**	**0.031**	**−12.92**	**<0.001*** **
	Picture	**0.122**	**0.030**	**4.14**	**<0.001*** **
	Interference × Picture	**0.611**	**0.042**	**14.454**	**<0.001*** **
Region 2 (Verb)	(Intercept)	−1.148	0.138	−8.299	<0.001
	Interference	**−0.250**	**0.078**	**−3.206**	**0.001*** **
	Picture	**0.263**	**0.078**	**3.391**	**<0.001*** **
	Interference × Picture	0.072	0.109	0.659	0.51
Region 3 (PP)	(Intercept)	−1.310	0.153	−8.54	<0.001
	Interference	**−0.315**	**0.049**	**−6.359**	**<0.001*** **
	Picture	**0.384**	**0.046**	**8.391**	**<0.001*** **
	Interference × Picture	0.008	0.067	0.114	0.909
Silence	(Intercept)	−1.457	0.172	−8.465	<0.001
	Interference	**−0.229**	**0.054**	**−4.211**	**<0.001*** **
	Picture	**0.183**	**0.051**	**3.614**	**<0.001*** **
	Interference × Picture	**0.260**	**0.073**	**3.56**	**<0.001*** **

****p < 0.001.*

condition for each participant. We observed a total of 447 inaccurate trials, with a total of 69,374 fixations and 4.7 items per condition per participant. Inspection of the pattern of eye-movements in the two item subsets reveals two important observations (see **Table 5** for modeling results). First, the effect of the bias toward the target in Region 1, which was created by the unintentional prosodic cues in the Interference trials, was more pronounced in accurate trials. This is apparent from the larger beta estimates in accurate trials vs. inaccurate trials (see **Table 5** for main effect estimates). *Post-hoc* contrasts of the effect in the Pictures Present condition revealed a larger effect when pictures were present in accurate trials (Tukey test: $\beta = 0.45$, $z = 9.00$, $p < 0.001$) vs. inaccurate trials (Tukey test: $\beta = 0.12$, $z = 3.62$, $p < 0.005$). In particular, in the Non-Interfering, Pictures Present condition, there were more looks to the target in inaccurate ($M = 0.29$; $SD = 0.13$) than accurate trials ($M = 0.23$; $SD = 0.17$) in all 4 ROIs. As discussed in the analysis of overall results, the direction of the effect was reversed in the Pictures Absent condition, but the magnitude of beta was still larger in the accurate trials (Tukey test: $\beta = -0.35$, $z = -5.78$, $p < 0.001$) than for inaccurate trials ($\beta = -0.29$, $z = -7.72$, $p < 0.001$). This is consistent with the idea that prosodic cues in the Interfering condition served to distinguish the target, which enabled participants to more accurately comprehend the sentences. However, given that only 33% of trials were correctly answered, it appears that these prosodic cues were often not helpful for participants.

Secondly, and more importantly, the data suggest an interference effect regardless of trial accuracy, but with different time-course manifestations. For incorrectly answered trials (top panel, **Figure 4**), looks to the target in the Interfering condition are reduced compared to the Non-Interfering condition beginning at the critical Region 2 (Verb), and continuing on, until the end of the sentence. This main effect was significant in all regions (see **Table 5**); pairwise comparisons verify the finding for both Pictures Present contrasts (Region 2, Tukey test: $z = -2.29$, $p < 0.05$; Region 3, Tukey test: $z = -0.20$, $p < 0.001$; Region 4, Tukey-test: $z = -3.04$, $p < 0.005$) and Pictures Absent contrasts (Region 2, Tukey test: $z = -4.58$, $p < 0.001$; Region 3, Tukey test: $z = -3.59$, $p < 0.001$; Region 4, Tukey test: $z = -7.29$, $p < 0.001$. In contrast, for the correctly answered trials, the Interference effect seems not to arise until the later Region 3 (PP), where we observed more looks to the target in the Non-Interfering condition for both the Pictures Present and Pictures Absent conditions[2]. The reason for this later time-course seems likely related to the bias in the Interfering conditions created by prosodic cues, which encouraged more looks to the target just prior to the critical verb. For Pictures Present trials, **Figure 4** shows an immediate increase in looks to the target in Region 2 for the Non-Interfering conditions, however given the already inflated baseline for looks in the Interfering condition, the difference between the two took longer to manifest. It is especially notable that even with the bias toward looks to the target in the Interfering condition, a reduction in looks to the target in that condition compared to looks in the Non-Interfering condition was anyway observed. Moreover, in the Pictures Absent conditions, a substantial increase in looks to the target in the Non-Interfering condition compared both to the previous baseline for that condition as well as the Interfering, Pictures Absent condition is also apparent. This effect did reach statistical significance (Tukey test: $z = -6.22$, $p < 0.001$). Although these data patterns are not all confirmed statistically,

[2] Pairwise contrasts for accurate trials are inconclusive due to the low number of observations and corresponding high variability per condition. We discuss here only the apparent pattern of looks displayed in **Figure 4**. When pairwise effects do reach significance, they are noted in the text.

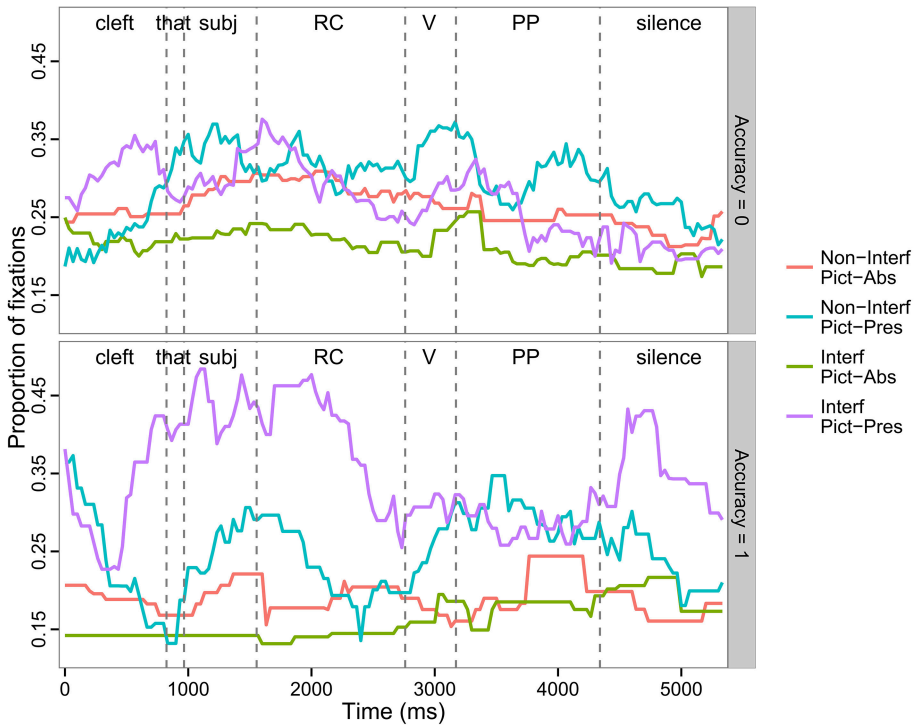


FIGURE 4 | Time course of fixation to the target picture as a function of Interference, Picture Memory List, and Region, separated by accuracy on comprehension questions. Top panel: Incorrectly answered trials (Acc = 0), Bottom panel: Correctly answered trials (Acc = 1).

they are consistent with the expected effect of the non-interfering verb as providing unambiguous cues for identifying the correct filler for the post-verbal gap.

Also of note in the accurate conditions, we observed a strong "correction" to the Interference effect in Region 4 (Silence), characterized by a steep increase in looks to the target in the Interference conditions. This effect was significant for both the Pictures Present contrast (Tukey test: $z = 4.47$, $p < 0.001$) and the Pictures Absent contrast (Tukey test: $z = 3.60$, $p < 0.001$). This is the same effect referred to in the overall analysis as a "wrap-up" or repair process. We conclude that this secondary analysis supports the repair interpretation of the Region 4 effect discussed above, as it was only the correctly answered trials that drove that late effect.

GENERAL DISCUSSION

The goal of the present experiment was to test whether the Visual World eye-tracking Paradigm can be extended to study retrieval interference in spoken language comprehension. We sought to determine whether the VWP could enable direct observation of online interference effects, through measuring overt looks to pictures of distractor referents held in memory, rather than needing to infer interference effects from reading times. The current study provides initial evidence—despite a methodological flaw—that indeed, retrieval interference effects do occur in the spoken modality, and the VWP provides a robust means of examining them. The key finding is of increased looks to extra-sentential competitors in the interference condition, which produced a concomitant *decrease* in looks to the target in this condition. This is consistent with the suggestion of Van Dyke and colleagues, that a cue-driven retrieval mechanism uses cues to query *all* of the contents of memory[3] . When the semantic cues from the verb also match the competitors, as in the Interference conditions, then this type of global matching will cause the competitors to affect processing (either by increasing reading times or engendering more looks to themselves), even though they are not in the sentence, or strongly related to each other or any other words in the sentence. The benefit of VWP paradigm is that we can directly observe the looks to the extra-sentential competitors, whereas in the original reading time studies a "No-Load" contrast condition was necessary to support the inference that the increased reading time at the verb in the interference condition was not due to a more difficult integration between the clefted NP and the verb. In what follows, we discuss our results further in relation to the original Van Dyke and McElree (2006) study.

Despite modality and methodological differences, the two studies are consistent in demonstrating effects of extra-sentential distractors on processes of argument integration. Although the dependent measures were different, i.e., eye-movement patterns

[3]We note that any effect of extra-sentential distractors is contrary to accounts that would give sentence processing a separate memory capacity (e.g., Caplan and Waters, 1999).

TABLE 5 | Proportions of looks to the target picture: Summary of mixed-effects logistic regression analyses by question accuracy (fixed effects only).

Region	Variable	Estimate	SE	*z*	*P*
INCORRECT TRIALS					
Region 1 (clefted NP+that+ Subject+RC)	(Intercept)	−1.071	.0639	−16.742	<0.001
	Interference	**−0.289**	**0.038**	**−7.714**	**<0.001*** **
	Picture	**0.096**	**0.036**	**2.671**	**<0.01** **
	Interference × Picture	**0.414**	**0.051**	**8.161**	**<0.001*** **
Region 2 (Verb)	(Intercept)	−0.947	0.137	−6.928	<0.001
	Interference	**−0.424**	**0.093**	**−4.579**	**0.001*** **
	Picture	0.046	0.093	0.495	0.621
	Interference × Picture	0.213	0.130	1.642	0.101
Region 3 (PP)	(Intercept)	−1.091	0.094	−11.64	<0.001
	Interference	**−0.202**	**0.057**	**−3.583**	**<0.001*** **
	Picture	**0.280**	**0.054**	**5.207**	**<0.001*** **
	Interference × Picture	**−0.181**	**0.077**	**−2.342**	**<0.05 ***
Silence	(Intercept)	−1.242	0.104	−11.941	<0.001
	Interference	**−0.215**	**0.064**	**−3.368**	**<0.001*** **
	Picture	−0.005	0.061	−0.075	0.940
	Interference × Picture	0.152	0.088	1.739	0.082
CORRECT TRIALS					
Region 1 (clefted NP+that+ Subject+RC)	(Intercept)	−1.458	0.163	−8.941	<0.001
	Interference	**−0.353**	**0.061**	**−5.782**	**<0.001*** **
	Picture	**0.291**	**0.054**	**5.447**	**<0.001** **
	Interference × Picture	**0.806**	**0.078**	**10.341**	**<0.001*** **
Region 2 (Verb)	(Intercept)	−1.592	0.245	−6.495	<0.001
	Interference	0.099	0.156	0.637	0.524
	Picture	**0.414**	**0.150**	**2.769**	**<0.01** **
	Interference × Picture	−0.121	0.207	−0.581	0.561
Region 3 (PP)	(Intercept)	−1.594	0.217	−7.361	<0.001
	Interference	**−0.658**	**0.106**	**−6.218**	**<0.001*** **
	Picture	**0.552**	**0.088**	**6.271**	**<0.001*** **
	Interference × Picture	**0.525**	**0.134**	**3.926**	**<0.001*** **
Silence	(Intercept)	−2.171	0.271	−8.013	<0.001
	Interference	**0.397**	**0.110**	**3.597**	**<0.001*** **
	Picture	**0.740**	**0.096**	**7.695**	**<0.001*** **
	Interference × Picture	−0.002	0.137	−0.021	0.983

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

over pictures vs. reading times in self-paced reading, the locus of the effect was the same across paradigms—at and after the manipulated verb, which provided either discriminating or ambiguous retrieval cues for identifying the target direct object. In the current experiment, participants looked significantly less to the target picture in the Interfering conditions than in the Non-Interfering conditions beginning at the critical verb, while in the written modality they read this verb more slowly. In both cases, we hypothesize that these effects are due to the presence of the distracting referents, be they pictures or words, which matched the retrieval cues of the verb (e.g., *spotted*) in the Interfering condition, but not in the Non-Interfering condition (e.g., *sewed*).

Moreover, the VWP proved sensitive to dynamic processes associated with recovering from incorrect retrievals, as evidenced by the marked increase in looks to the target for Interfering conditions in the silence region for correct trials. This is similar to the sentence-final effect of semantic interference from distractors *within* a sentence observed by Van Dyke (2007), however the VWP has the added benefit of providing direct evidence that the increased reading times are associated with additional processing of the target in the Interfering conditions but not in the

Non-Interfering conditions. In both cases, we take this increased late effort to reflect repair processes, invoked when listeners realize they have constructed an incorrect interpretation.

Despite the weakness in the current study related to unintended prosodic cues, which may have created increased encoding opportunities for the target in the Interfering condition, the interference effect was clearly observed in the Pictures Present conditions. It attests to the robustness of both the VWP method for indexing integrative processes (e.g., Tanenhaus et al., 1995; Huettig et al., 2011) and the retrieval interference effect itself. One might have expected that the more salient target would have promoted correct integration of the clefted NP, however the eye-movement patterns suggest interference effects in both correct and incorrect trials (although low power yielded non-significant results in the latter category). This demonstrates that salience alone is not sufficient to override the immediate effects of ambiguous retrieval cues on argument integration.

We attempted to further validate this conclusion using the blank screen paradigm, where we expected the same pattern of results as in the pictures present condition. This would have replicated the Altmann and Kamide (1999) results and leant further support to the hypothesis that the looks to the target in the pictures present condition are not a mere epiphenomenon due to visual cues, but instead reflect integrative processing driven by a cue based retrieval mechanism. Unfortunately, as described above, results from the pictures absent condition were difficult to interpret. Nevertheless, there remains a significant body of research that has established that looks to target objects during sentence processing cannot be entirely attributed to visual cues, but instead reflect activation of mental representations at least partially guided by the parser (Spivey and Geng, 2001; Altmann, 2004; Altmann and Kamide, 2004; Johansson et al., 2006). Moreover, the current study replicates Van Dyke and McElree (2006) which used the exact same sentences to demonstrate interference effects in relation to retrieval of previously stored distractors. Based on these considerations we are confident in concluding that our findings in the pictures present condition reflect the memory retrieval mechanisms at work during sentence comprehension. However, we do acknowledge the need for future work to demonstrate the validity of this approach to examining interference effects in the spoken modality more generally.

A further unexpected outcome was the extremely low accuracy to comprehension questions. We believe the primary reason for low accuracy is that the dual process task is quite difficult. High scores in the picture recall suggests that participants traded off attention to that task, for attention to the sentence task, which impacted their ability to correctly answer questions. It is highly possible that answering offline comprehension questions, which require a meta-analysis of what was heard, may be difficult for these participants for reasons that are entirely unrelated to our manipulation (e.g., poor meta-analysis skills or difficulty querying the situation model). In addition, the dissociation between accuracy scores for picture recall (high) and comprehension questions (low), together with the significant effects of Interference observed in the Pictures Present conditions, suggests that performance on comprehension questions is a poor index of whether participants experienced online effects of interference. Even when the eye movement record shows evidence of interference effects, there is no guarantee that participants were able to *accurately resolve* the interference, leading to correct performance on the comprehension questions. Thus, we take the accuracy scores to be orthogonal to the main conclusion to be drawn from these data; namely, that the VWP can reliably index retrieval interference effects during spoken language comprehension. We interpret our observation of these effects in eye movements, despite low comprehension, as an even stronger indicator that the VWP is a sensitive method for these effects.

Finally, we note an additional contribution of the current study, which is to further the goal of determining which cues guide retrieval and how they are combined (Van Dyke and McElree, 2011). This study provides an initial indication that retrieval interference effects occur independently of prosodic cues. This will be an important area for future research, some of which is already occurring in our laboratories. This paper demonstrates that the VWP is a useful method for investigating these effects. In addition, the sensitivity of VWP to indexing effects of retrieval interference opens up new possibilities for evaluating predictions of the Cue-Based Retrieval Theory in non-reading populations, such as people with aphasia, children, and auditory second language learners.

AUTHOR CONTRIBUTIONS

IS: Design, data collection, data coding; LC: Statistical analysis; JV: Design, materials, theory development; IS, LC, and JV equally contributed to writing of the article.

ACKNOWLEDGMENTS

The article was prepared within the framework of the Basic Research Program at the National Research University Higher School of Economics (HSE) and supported within the framework of a subsidy granted to the HSE by the Government of the Russian Federation for the implementation of the Global Competitiveness Program. It was also supported by PSC-CUNY grant # 66048-00 44 awarded to Irina Sekerina and by NIH grant HD 073288 (National Institute of Child Health and Human Development) to Haskins Laboratories: (JV, PI). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We thank Jason Bishop for prosodic analysis of the experimental sentences, Nina Gumkowski for recording the spoken materials, and Namseok Yong for his assistance in running the experiment and data coding. We are also very grateful to the two reviewers who have tirelessly pushed us to make this article the best it can be.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpsyg.2016.00873

REFERENCES

Akinina, Y., Malyutina, S., Ivanova, M., Iskra, E., Mannova, E., and Dragoy, O. (2015). Russian normative data for 375 action pictures and verbs. *Behav. Res. Methods* 47, 691–707. doi: 10.3758/s13428-014-0492-9

Altmann, G. T. M. (2004). Language-mediated eye movements in the absence of a visual world: the "blank screen paradigm." *Cognition* 93, B79–B87. doi: 10.1016/j.cognition.2004.02.005

Altmann, G. T. M., and Kamide, Y. (1999). Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition* 73, 247–264. doi: 10.1016/S0010-0277(99)00059-1

Altmann, G. T. M., and Kamide, Y. (2004). "Now you see it, now you don't: mediating the mapping between language and the visual world," in *The Interface of Language, Vision, and Action: Eye Movements and the Visual World*, eds J. M. Henderson and F. Ferreira (New York, NY: Psychology Press), 347–386.

Baayen, R. H., Davidson, D. J., and Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *J. Mem. Lang.* 59, 390–412. doi: 10.1016/j.jml.2007.12.005

Baddeley, A. D. (1966). Short-term memory for word sequences as a function of acoustic, semantic and formal similarity. *Q. J. Exp. Psychol.* 18, 362–365. doi: 10.1080/14640746608400055

Baddeley, A. D. (2012). Working memory: theories, models, and controversies. *Annu. Rev. Psychol.* 63, 1–29. doi: 10.1146/annurev-psych-120710-100422

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2014). *lme4: Linear Mixed-Effects Models Using Eigen and S4*. Available onlien at: http://CRAN.R-project.org/package=lme4

Caplan, D., DeDe, G., Waters, G., Michaud, J., and Tripodis, Y. (2011). Effects of age, speed of processing, and working memory on comprehension of sentences with relative clauses. *Psychol. Aging* 26, 439–450. doi: 10.1037/a0021837

Caplan, D., and Waters, G. S. (1999). Verbal working memory and sentence comprehension. *Behav. Brain Sci.* 22, 77–94.

Chee, M. W. L., O'Craven, K. M., Bergida, R., Rosen, B. R., and Savoy, R. L. (1999). Auditory and visual word processing studied with fMRI. *Hum. Brain Mapp.* 7, 15–28.

Clackson, K., Felser, C., and Clahsen, H. (2011). Children's processing of pronouns and reflexives in English: evidence from eye-movements during listening. *J. Mem. Lang.* 65, 128–144. doi: 10.1016/j.jml.2011.04.007

Clark, S. E., and Gronlund, S. D. (1996). Global matching models of recognition memory: how the models match the data. *Psychonom. B. Rev.* 3, 37–60. doi: 10.3758/BF03210740

Coco, M. I., Keller, F., and Malcolm, G. L. (2015). Anticipation in real-world scenes: the role of visual context and visual memory. *Cognitive Sci.* 39, 1–15. doi: 10.1111/cogs.12313

DeDe, G. (2013). Reading and listening in people with aphasia: effects of syntactic complexity. *Am. J. Speech Lang. Pathol.* 22, 579–590. doi: 10.1044/1058-0360(2013/12-0111)

Dickey, M. W., Choy, J. J., and Thompson, C. K. (2007). Real-time comprehension of wh- movement in aphasia: evidence from eyetracking while listening. *Brain Lang.* 100, 1–22. doi: 10.1016/j.bandl.2006.06.004

Forster, K. I., and Forster, J. C. (2003). DMDX: a windows display program with millisecond accuracy. *Behav. Res. Methods Instrum. Comput.* 35, 116–124. doi: 10.3758/BF03195503

Glaser, Y. G., Martin, R. C., Van Dyke, J. A., Hamilton, A. C., and Tan, Y. (2013). Neural basis of semantic and syntactic interference in sentence comprehension. *Brain Lang.* 126, 314–326. doi: 10.1016/j.bandl.2013.06.006

Gordon, P. C., Hendrick, R., and Levine, W. H. (2002). Memory-load interference in syntactic processing. *Psychol. Sci.* 13, 425–430. doi: 10.1111/1467-9280.00475

Hallowell, B., Wertz, R. T., and Kruse, H. (2002). Using eye movement responses to index auditory comprehension: an adaptation of the Revised Token Test. *Aphasiology* 16, 587–594. doi: 10.1080/02687030244000121

Hedberg, N. (2013). "Multiple focus and cleft sentences," in Cleft *Structures*, eds K. Hartmann and T. Veenstra (Amsterdam: John Benjamins), 227–250.

Huettig, F., and Altmann, G. T. M. (2005). Word meaning and the control of eye fixation: semantic competitor effects and the visual world paradigm. *Cognition* 96, B23–B32. doi: 10.1016/j.cognition.2004.10.003

Huettig, F., Rommers, J., and Meyer, A. S. (2011). Using the visual world paradigm to study language processing: a review and critical evaluation. *Acta Psychol.* 137, 151–171. doi: 10.1016/j.actpsy.2010.11.003

Ivanova, M. V., and Hallowell, B. (2012). Validity of an eye-tracking method to index working memory in people with and without aphasia. *Aphasiology* 26, 556–578. doi: 10.1080/02687038.2011.618219

Johansson, R., Holsanova, J., and Holmqvist, K. (2006). Pictures and spoken descriptions elicit similar eye movements during mental imagery, both in light and in complete darkness. *Cogn. Sci.* 30, 1053–1079. doi: 10.1207/s15516709cog0000_86

Johns, C. L., Matsuki, K., and Van Dyke, J. A. (2015). Poor readers' retrieval mechanism: efficient access is not dependent on reading skill. *Lang. Sci.* 6:1552. doi: 10.3389/fpsyg.2015.01552

Kamide, Y., Altmann, G. T. M., and Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: evidence from anticipatory eye movements. *J. Mem. Lang.* 49, 133–156. doi: 10.1016/S0749-596X(03)00023-8

Kukona, A., Cho, P. W., Magnuson, J. S., and Tabor, W. (2014). Lexical interference effects in sentence processing: Evidence from the visual world paradigm and self-organizing models. *J. Exp Psychol. Learn.* 40, 326–347. doi: 10.1037/a0034903

Levy, R., Fedorenko, E., and Gibson, E. (2013). The syntactic complexity of Russian relative clauses. *J. Mem. Lang.* 69, 461–495. doi: 10.1016/j.jml.2012.10.005

Lewis, R. L., Vasishth, S., and Van Dyke, J. A. (2006). Computational principles of working memory in sentence comprehension. *Trends Cogn. Sci.* 10, 447–454. doi: 10.1016/j.tics.2006.08.007

Liberman, A. M., Mattingly, I. G., and Turvey, M. T. (1972). "Language codes and memory codes," in *Coding Processes in Human Memory*, eds A. W. Melton and E. Martin (Washington, DC: Winston & Sons), 307–334.

Martin, A. E., and McElree, B. (2009). Memory operations that support language comprehension: evidence from verb-phrase ellipsis. *J. Exp. Psychol. Learn. Mem. Cogn.* 35, 1231–1239. doi: 10.1037/a0016271

Martin, A. E., Nieuwland, M. S., and Carreiras, M. (2012). Event-related brain potentials index cue-based retrieval interference during sentence comprehension. *Neuroimage* 59, 1859–1869. doi: 10.1016/j.neuroimage.2011.08.057

McElree, B. (2006). "Accessing recent events," in The *Psychology* of *Learning* and *Motivation*, ed B. Ross (San Diego, CA: Academic Press), 155–200.

McElree, B., Foraker, S., and Dyer, L. (2003). Memory structures that subserve sentence comprehension. *J. Mem. Lang.* 48, 67–91. doi: 10.1016/S0749-596X(02)00515-6

Michael, E. B., Keller, T. A., Carpenter, P. A., and Just, M. A. (2001). fMRI investigation of sentence comprehension by eye and by ear: modality fingerprints on cognitive processes. *Hum. Brain Mapp.* 13, 239–252. doi: 10.1002/hbm.1036

Nelson, D. L., and Brooks, D. H. (1973). Functional independence of pictures and their verbal memory codes. *J. Exp. Psychol.* 98, 44–48. doi: 10.1037/h0034299

Paivio, A. (1986). *Mental Representations: A Dual Coding Approach*. New York, NY: Oxford University Press.

Raudenbush, S. W., and Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods, 2nd Edn*. Thousand Oaks, CA: Sage Publications.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Available online at: http://www.R-project.org/

Rüschemeyer, S.-A., Zysset, S., and Friederici, A. D. (2006). Native and non-native reading of sentences: an fMRI experiment. *Neuroimage* 31, 354–365. doi: 10.1016/j.neuroimage.2005.11.047

Sheppard, S. M., Walenski, M., Love, T., and Shapiro, L. P. (2015). The auditory comprehension of Wh-Questions in aphasia: pport for the intervener hypothesis. *J. Speech Lang. Hear. R.* 58, 781–797. doi: 10.1044/2015_JSLHR-L-14-0099

Snodgrass, J. G., and McClure, P. (1975). Storage and retrieval properties of dual codes for pictures and words in recognition memory. *J. Exp. Psychol. Hum. Learn.* 1, 521–529. doi: 10.1037/0278-7393.1.5.521

Spivey, M. J., and Geng, J. J. (2001). Oculomotor mechanisms activated by imagery and memory: eye movements to absent objects. *Psychol. Res.* 65, 235–241. doi: 10.1007/s004260100059

Sussman, R. S., and Sedivy, J. (2003). The time-course of processing syntactic dependencies: evidence from eye movements. *Lang. Cognitive Proc.* 18, 143–163. doi: 10.1080/01690960143000498

Tanenhaus, M., Spivey-Knowlton, M., Eberhard, K., and Sedivy, J. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science* 268, 1632–1634. doi: 10.1126/science.7777863

Trueswell, J. C., and Tanenhaus, M. K. (eds.). (2005). *Approaches to Studying World-Situated Language Use: Bridging the Language-As-Product and Language-As-Action Traditions*. Cambridge, MA: MIT Press.

Van Dyke, J. A. (2007). Interference effects from grammatically unavailable constituents during sentence processing. *J. Exp. Psychol. Learn. Mem. Cogn.* 33, 407–430. doi: 10.1037/0278-7393.33.2.407

Van Dyke, J. A., and Johns, C. L. (2012). Memory interference as a determinant of language comprehension. *Lang. Linguist. Compass* 6, 193–211. doi: 10.1002/lnc3.330

Van Dyke, J. A., Johns, C. L., and Kukona, A. (2014). Low working memory capacity is only spuriously related to poor reading comprehension. *Cognition* 131, 373–403. doi: 10.1016/j.cognition.2014.01.007

Van Dyke, J. A., and Lewis, R. L. (2003). Distinguishing effects of structure and decay on attachment and repair: a cue-based parsing account of recovery from misanalyzed ambiguities. *J. Mem. Lang.* 49, 285–316. doi: 10.1016/S0749-596X(03)00081-0

Van Dyke, J. A., and McElree, B. (2006). Retrieval interference in sentence comprehension. *J. Mem. Lang.* 55, 157–166. doi: 10.1016/j.jml.2006.03.007

Van Dyke, J. A., and McElree, B. (2011). Cue-dependent interference in comprehension. *J. Mem. Lang.* 65, 247–263. doi: 10.1016/j.jml.2011.05.002

Vasishth, S., Brussow, S., Lewis, R. L., and Drenhaus, H. (2008). Processing polarity: how the ungrammatical intrudes on the grammatical. *Cogn. Sci. Multidiscip. J.* 32, 685–712. doi: 10.1080/03640210802066865

Waters, G. S., and Caplan, D. (2005). The relationship between age, processing speed, working memory capacity, and language comprehension. *Memory* 13, 403–413. doi: 10.1080/09658210344000459

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2016 Sekerina, Campanelli and Van Dyke. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 24 March 2015
doi: 10.3389/fpsyg.2015.00331

The processing of raising and nominal control: an eye-tracking study

Patrick Sturt[1]* *and Nayoung Kwon*[2]

[1] Psychology, School of Philosophy, Psychology and Language Sciences, University of Edinburgh, Edinburgh, UK,
[2] Department of English Language, Konkuk University, Seoul, South Korea

According to some views of sentence processing, the memory retrieval processes involved in dependency formation may differ as a function of the type of dependency involved. For example, using closely matched materials in a single experiment, Dillon et al. (2013) found evidence for retrieval interference in subject-verb agreement, but not in reflexive-antecedent agreement. We report four eye-tracking experiments that examine examine reflexive-antecedent dependencies, combined with raising (e.g., "John seemed to Tom to be kind to himself…"), or nominal control (e.g., "John's agreement with Tom to be kind to himself…"). We hypothesized that dependencies involving raising would (a) be processed more quickly, and (b) be less subject to retrieval interference, relative to those involving nominal control. This is due to the fact that the interpretation of raising is structurally constrained, while the interpretation of nominal control depends crucially on lexical properties of the control nominal. The results showed evidence of interference when the reflexive-antecedent dependency was mediated by raising or nominal control, but very little evidence that could be interpreted in terms of interference for direct reflexive-antecedent dependencies that did not involve raising or control. However, there was no evidence either for greater interference, or for quicker dependency formation, for raising than for nominal control.

Keywords: parsing, memory retrieval, eye-tracking, dependency formation, binding, raising, control

OPEN ACCESS

Edited by:
Colin Phillips,
University of Maryland, USA

Reviewed by:
Dan Parker,
College of William and Mary, USA
Jeffrey Thomas Runner,
University of Rochester, USA

***Correspondence:**
Patrick Sturt,
Psychology, School of Philosophy, Psychology and Language Sciences, University of Edinburgh, 7 George Square, Edinburgh EH9 9JZ, UK
patrick.sturt@ed.ac.uk

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: *22 December 2014*
Accepted: *07 March 2015*
Published: *24 March 2015*

Citation:
Sturt P and Kwon N (2015) The processing of raising and nominal control: an eye-tracking study. Front. Psychol. 6:331. doi: 10.3389/fpsyg.2015.00331

1. Introduction

Successful language comprehension requires the computation of grammatical dependencies between linguistic elements in each sentence. For example, the interpretation of (1) requires a dependency between the reflexive *himself* and its antecedent *John*:

1. Bill thought that John was kind to himself.

However, although a great deal of research has been directed at the factors that affect processing difficulty during sentence comprehension, it is only recently that researchers have begun to turn their attention to the actual mechanisms that are used in on-line dependency formation.

One important aspect of dependency computation that has recently been examined in a number of studies is *memory retrieval*. Given that linguistic input is sequential, the two end-points of a dependency (e.g., *John* and *himself* in 1) are necessarily separated in time. In cases like (1), this means that memory access is required to solve the dependency—in order to interpret *himself*, the antecedent *John* needs to be retrieved from working memory. Recent work in human

sentence processing has sought to examine the types of memory retrieval processes that best characterise linguistic dependency formation. According to a well-known view (e.g., McElree et al., 2003; Lewis and Vasishth, 2005; Lewis et al., 2006; Van Dyke and McElree, 2006), memory retrieval in sentence processing is a content-addressable process, in which potential targets in memory are activated in response to retrieval cues. For example, in (1), when *himself* is processed, the retrieval cues might include gender (the antecedent has to be masculine), as well as relevant structural information (the antecedent has to be in an appropriate local position relative to *himself*). According to such models, the activation of dependency targets is a parallel process, where multiple potential targets may be activated simultaneously through partial cue matching. This means that the retrieval of a grammatically licit retrieval target may be affected by the presence of other (grammatically illicit) items that partially match the retrieval cues, a phenomenon known as *interference*. For example, in (1), during the retrieval of the grammatically correct antecedent *John*, the grammatically illicit antecedent *Bill* may become partially activated, as it matches the *male* feature required by *himself*. This may affect the time taken to retrieve the correct antecedent *John*.

Computational models that make predictions concerning retrieval speed (e.g., Lewis and Vasishth, 2005; Lewis et al., 2006) predict that that interference can be either *facilitatory* (speeding up retrieval) or *inhibitory* (slowing down retrieval), depending on the contents of working memory at the point where retrieval takes place, and on the retrieval cues. These models assume a monotonic relation between retrieval times predicted by the model and reading times at the relevant point of the sentence where retrieval is assumed to occur. Below, we briefly describe two patterns of interference that have been reported in the literature. In this paper, we will refer to these as *facilitatory interference* and *inhibitory interference* respectively.

Facilitatory interference can be illustrated using the subject-verb agreement examples given in (2a,b), taken from the self-paced reading study reported by Wagers et al. (2009):

2a. The musician who the reviewer praise so highly will probably win a grammy.
2b. The musicians who the reviewer praise so highly will probably win a grammy.

Both (2a) and (2b) are ungrammatical, due to the number mismatch between the plural verb *praise* and the singular relative clause subject *reviewer*. However, Wagers et al. (2009) found that the reading time penalty was significantly reduced in (2b), which includes a plural distractor *the musicians*, relative to (2a), which does not. In this paper, we will use the term *facilitatory interference* specifically to refer to the reduction of processing difficulty (and thus faster retrieval) for a mismatching dependency, due to the presence of a partially matching distractor (see also Vasishth et al. (2008) and Xiang et al. (2009) for examples of facilitatory interference in other types of dependencies).

In the computational model proposed by Lewis and Vasishth (2005) and Lewis et al. (2006), facilitatory interference is explained in terms of mis-retrieval of the illicit retrieval target. For example, in (2a), the retrieval cues of the verb lead to activation of all potential targets in parallel, including both a licit and an illicit antecedent. However, the mismatching number feature on the licit antecedent, *the reviewer* in (2a,b) means that its retrieval takes a relatively long time. Similarly, in (2a), there is relatively little feature overlap between the distractor, *the musician* and the retrieval cues, leading to a lower activation of the distractor, and thus lower probability of misretrieval. On the other hand, in (2b) the distractor, *the musicians* partially matches the retrieval cues of the verb, sometimes leading to mis-retrieval of *the musicians* as the subject of *praise*. This "illusionary licensing" effect could lead to faster processing in (2b) relative to in (2a).

The second phenomenon that has been argued to follow from a content-addressable memory system is *inhibitory interference*. To illustrate this phenomenon, consider (3a,b), from Badecker and Straub (2002):

3a. John thought that Bill owed him another chance to solve the problem.
3b. John thought that Beth owed him another chance to solve the problem.

In both (3a) and (3b), the only grammatically licit antecedent for the pronoun *him* is *John*. However, in (3a), there is also a gender-matching (grammatically illicit) distractor (*Bill*), while (3b) contains a mismatching distractor *Beth*. Badecker and Straub (2002) found that the two words following the pronoun were read more slowly in (3a) than in (3b). In this paper, we use the term *inhibitory (retrieval) interference* to refer to processing difficulty (and thus slow retrieval) that occurs when the intended dependency target completely matches the retrieval cues (e.g., *John* in 3a), but where there is also a partial match with the distractor—for example, *Bill* in 3a is a distractor, as it is not a grammatically possible antecedent for *him*, but it partially matches the retrieval cue, as it bears the required *male* feature[1].

In the computational model of Lewis and Vasishth (2005) and Lewis et al. (2006), inhibitory interference effects can be explained in terms of the parallel activation of all partially matching retrieval targets; in (3a), the distractor *Bill* has a relatively high activation level during the retrieval of *John*, due to the fact that it partially matches the features of the retrieval cue (i.e., it is masculine), and this leads to competition, slowing down the retrieval of the intended referent *John*. In contrast, in (3b), the distractor *Beth* overlaps to a lesser degree with the retrieval cue, leading to a relatively low activation, and thus less competition.

To summarize, in this paper, we use the term *facilitatory interference* to refer specifically to facilitation in the retrieval of a *feature mismatching* retrieval target; and we use *inhibitory interference* to refer speficically to inhibition in the retrieval of a *feature matching* retrieval target. In both cases, this is due to a the presence of a distractor that partially matches the retrieval cue. The previous literature on interference effects in dependency formation has yielded a mixed picture—although both facilitatory and inhibitory interference effects have been found, neither of these have been found consistently across different dependency

[1]See Van Dyke (2007) and Gordon et al. (2006) for examples of inhibitory interference involving other types of dependencies. Gordon et al. (2006) argue that inhibitory similarity-based interference reflects feature-overlap or feature overwriting in the encoding stage, rather than multiple cue-overlap in retrieval.

types. For example, while Badecker and Straub (2002) found inhibitory interference for both pronoun-antecedent dependencies and reflexive-antecedent dependencies, these results have seldom been replicated. In fact, subsequent studies have failed to replicate inhibitory interference for both pronouns (Chow et al., 2014) and reflexives (e.g., Dillon et al., 2013, *inter alia)*.

Facilitatory interference is reliably found for subject-verb agreement (e.g., Wagers et al., 2009) and negative polarity licensing (Vasishth et al., 2008; Xiang et al., 2009) but has not been consistently found for reflexive-antecedent agreement. Indeed, one recent study (Dillon et al., 2013) has directly compared these two dependency types in a single experiment, and found facilitatory interference effects only for subject-verb agreement, but no evidence for interference for reflexive-antecedent agreement.

The correct explanation for this variability in interference effects is not currently known. Concentrating on the variability of facilitatory interference across dependency types, Phillips et al. (2011) suggest that the parser may make use of either a structure-sensitive search process, or a content-addressable retrieval process, depending on the certain features of the dependency that is being computed: Specifically, Phillips et al. (2011) suggest that the type of memory access that is used may depend on how quickly structural information becomes available relative to other information.

Another possibility, argued by Dillon et al. (2013), is that all dependency types involve a content addressable retrieval process, but that the cues used for retrieval differ between different dependency types. This idea predicts that different types of dependency may lead to different interference profiles, even though they may target the same item in memory, for example, the subject of the local clause. For example, Dillon et al. (2013) contrasted reflexive-antecedent dependencies with subject-verb dependencies, both of which involve the local subject as a retrieval target. Based on their finding of facilitatory interference only in subject-verb dependencies, they argued that, while both dependencies make use of structural cues targeting the local subject, only subject-verb agreement uses the featural cue of number. The use of number as a retrieval cue in the subject-verb dependencies predicts that number-matching distractors become activated during retrieval, leading to interference, as observed by both Dillon et al. (2013) and others such as Wagers et al. (2009). In contrast, Dillon et al. (2013) argue that reflexive-antecedent dependencies only make use of structural retrieval cues, but not featural cues such as number and gender. If number is not used as a retrieval cue for reflexive-antecedent dependencies, this predicts that number-matching distractors are not activated during retrieval, which in turn predicts a lack of interference effects, in contrast with subject-verb agreement[2].

The idea that different types of dependencies could involve different retrieval cues or processes, however, has not yet been tested using a wide range of dependencies. In particular, few studies have examined the retrieval processes of lexically-based dependencies, or compared them systematically with more structurally-based dependencies. Accordingly, it is not clear how retrieval processes would differ between these two types of dependencies. Thus, in this paper, we compared the retrieval processes of raising and nominal control constructions, which are illustrated in (4a,b) below.

4a. **Raising:**
It was surprising that John seemed Ø to be kind to himself.
4b. **Nominal control:**
I was surprised at John's agreement Ø to be kind to himself.

In (4a,b), the phonologically unexpressed subject of the infinitive (marked Ø in the above examples) participates in a dependency with its antecedent *John*. In (4a), the dependency is formed through raising, while in (4b), it is formed through nominal control. Raising and nominal control differ in many ways that could be relevant for processing. One important difference lies in the way that a dependency is motivated. That is, while the interpretation of raising is structurally constrained, the interpretation of nominal control depends crucially on lexical properties of the control nominal. For example, compare (5a) and (5b) below:

5a. John's agreement with Mary Ø to be kind to himself.
5b. John's order to Mary Ø to be kind to herself.

In (5a), the control nominal *agreement* is an instance of *giver control* (see Culicover and Jackendoff, 2001, for an overview of nominal control), meaning that Ø is interpreted as co-referential with the *giver* of the agreement (i.e., *John*)[3]. This leads to an interpretation in which *John* is kind to himself. In (5b), in contrast, *order* is an instance of *recipient control*, meaning that Ø is interpreted as co-referential with the *recipient* of the order (i.e., *Mary*). The interpretation is that *Mary* is kind to herself. In contrast with nominal control dependencies, raising dependencies, such as (4a) above do not exhibit lexically specific variability in the range of potential interpretations: if a raising verb (e.g., *seemed*) has a referential subject (e.g., *John* in 4a), then this must always be co-referential with the subject of the infinitive complement (i.e., Ø in 4a; cf. Hornstein, 1999). Thus, in (6) below, even though an "experiencer" distractor argument (i.e., *Mary*) intervenes, the raising construction still requires co-reference with *John*:

6. John seems to Mary Ø to be kind to himself.

These differences arguably have an analog in a representational distinction that syntacticians typically draw between raising and control. For example, in the Principles and Parameters framework (e.g., Chomsky, 1986) the empty subject in the raising example (6) is assumed to be an instance *NP-trace*, which participates in a strictly local and structurally constrained dependency with its antecedent. In contrast, the empty subject in all varieties of control, including nominal control (5), is assumed to be PRO, a pronominal element that is much less constrained, and whose choice of antecedent will depend on many factors, including the

[2]In addition to facilitatory interference and inhibitory interference, one other interference profile has been reported in the literature: in a study on the resolution of Spanish dependencies involving *otro* (similar to English "another"), Martin et al. (2012) reported processing disruption for grammatical dependencies where a distractor mismatched the retrieval cue.

[3]We used nominal control rather than verbal control because nominal control gives a wide range of control predicates that can be used in the *giver*-control condition (e.g., *vow oath*, *promise*). For verbal control, in contrast, there are very few control verbs that can be used in an analogous way.

type of control relation. In other syntactic frameworks, the representational difference between (5) and (6) is even more marked—for example, in certain varieties of Lexicalized Tree Adjoining grammar (LTAG), the raising example (6) would not be assumed to include an empty infinitival subject at all[4], while the control example (5) would include PRO, as in the Principles and Parameters framework (see the X-tag grammar of English, XTAG Research Group, 1998, for a framework that takes this approach). For the purposes of the present paper, we will continue to assume that both raising and nominal control involve an empty subject, but we will return to consider the predictions of the LTAG proposal where relevant below.

What types of cues might be used in retrieving the antecedent of the empty subject Ø in (5) and (6)? We assume that the empty subject is initially recognized around the point where *to be kind* is reached in the input, and we also assume that a retrieval process is launched around this point, to find the antecedent of Ø. Given the discussion above, it would make sense to assume that the retrieval cue for the raising dependency (6) would be structural in nature, (for example, targeting the subject of the next-highest finite clause). For nominal control (6), the retrieval cue would need to be represented in a more complex way, as it would need to refer to the control predicate (e.g., *agreement* or *order*), and locate the required target based on that predicate's control properties and argument structure.

In the studies reported in this paper, we examine the processing of sentences that are similar to (5) and (6), in that they combine a control or raising dependency with a reflexive-antecedent dependency. In both (5a) and (6), the dependency between the reflexive *himself* and its ultimate antecedent *John* is indirect—there is one (anaphor-antecedent) dependency between the reflexive and the empty subject, and another (raising or control) dependency between the empty subject and its antecedent. In other words, the dependency between the reflexive and its antecedent is mediated by nominal control (5) or raising (6). We therefore assume that the process of retrieval of the reflexive's antecedent is also mediated by nominal control or raising in cases like these (5) and (6). As a consequence, there are (at least) two retrieval events that involve raising or control in each of these sentences—the initial retrieval of the empty subject's antecedent around the infinitival verb, and a second retrieval, triggered by the reflexive. This second retrieval event, which is the focus of the experiments reported in this paper, has a wider range of cues that could potentially be relevant, because the reflexive provides gender and number information that is not available at the point where the empty infinitival subject is initially recognized—in the case of (5a) and (6), the reflexive requires its antecedent Ø to be male and singular, so Ø in turn must also require its antecedent to be male and singular. Whether each of these dependencies actually uses gender or number as retrieval cues is an empirical question. However, given that the nominal control dependency involves the element PRO, which is a species of pronoun, whose resolution is influenced by a wide range of factors, we believe that this dependency is more likely to use gender and number as a retrieval cue than the purely structural raising dependency.

In this paper, we test the hypothesis that nominal control dependencies would be (a) more prone to interference, and (b) processed more slowly, than raising dependencies. There are several reasons why nominal control dependencies might be expected to be more susceptible to interference than raising dependencies. One reason is that, as discussed above, the resolution of nominal control dependencies requires the use of complex constraints involving lexical information, while raising dependencies can be resolved through purely structural means. This might lead to more indeterminacy in the retrieval process for nominal control, leading to more interference, or it might mean that the two dependency types use qualitatively different retrieval mechanisms, for example, an interference-prone content-addressable mechanism for nominal control but a direct structure-based search for raising. A second possible reason is that, even if both dependency types use a content-addressable mechanism, nominal control dependencies may use a wider array of retrieval cues than raising dependencies, allowing more opportunity for a partial match with a distractor. In the present paper, we are particularly concerned with gender as a retrieval cue, as we use an experimental paradigm that manipulates gender agreement via reflexive-antecedent dependencies, allowing for the possibility of interference via a gender-matching distractor (see below for details). Under these circumstances, control dependencies would be expected to be susceptible to interference if they can use gender as a retrieval cue, while raising dependencies would be expected to be less susceptible, if their retrieval cues are purely structural. Finally, if nominal control and raising dependencies involve very different syntactic representations (e.g., if nominal control uses an empty infinitival subject, while raising does not, as suggested by the LTAG analysis, XTAG Research Group, 1998), then this could lead to different retrieval profiles for the two dependencies. We will postpone further discussion of this last point until the introduction to Experiment 4 below.

Our second hypothesis was that nominal control dependencies would be processed more slowly than raising dependencies. In order to examine this question, as well as retrieval interference, we used a *gender mismatch paradigm* (Sturt, 2003), combining raising or control dependencies with reflexive-antecedent gender agreement, as mentioned above. In this type of experiment, the matching between the reflexive and its antecedent is manipulated. For example, compare example (6) above, with the mis-matching variant in (7):

7. John seems to Mary Ø to be kind to herself.

In (7), the gender of the reflexive *herself* mismatches with the structurally appropriate antecedent *John*. Previous work, using eyetracking during reading, has shown that readers fixate for longer on a reflexive the when its gender mismatches that of its structurally licit antecedent (relative to matching controls) (see for example Sturt, 2003, inter alia).

In this paper, we refer to such processing difficulty as the *mismatch cost*, and we are particularly interested in the *onset* of the mismatch cost in the eye-movement record, in relation to the onset of the first fixation on the reflexive, as a measure of how

[4]Technically, *John* is substituted into the subject position of the elementary tree headed by *kind*, and the elementary tree headed by *seems* is adjoined (i.e., inserted) into the elementary tree headed by *kind*.

quickly the grammatically appropriate antecedent is identified. In previous studies using eye-tracking, the mismatch cost for reflexive-antecedent dependencies has been observed very early in the eye-movement record. For example, Sturt (2003) reported that the first fixation on a reflexive with a gender mismatching antecedent was reliably shorter than when the antecedent matched in gender. Since the average fixation duration in reading is around 250 ms, this implies that the structurally appropriate antecedent must have been recognized within 250 ms after the reader first started fixating the reflexive.

In fact, there is some evidence to suggest that the *onset* of the mismatch cost may differ depending on the structure of the sentence that contains the reflexive. For example, in a series of eye-tracking experiments, Cunnings and Sturt (2014) used the gender mismatch paradigm to examine the resolution of reflexive pronouns sentences like (8a,b):

8a. He heard that the soldier had positioned himself/herself in the middle of the mess hall.
8b. He heard that the soldier had a picture of himself/herself in the middle of the mess hall.

The design included reflexives that either matched (*himself*) or mismatched (*herself*) the stereotypical gender of the antecedent (*the soldier*). In separate experiments, Cunnings and Sturt (2014) found evidence of a mismatch cost for both (8a) and (8b)—in both cases, readers began to slow down after they had initially fixated a mismatching reflexive (relative to a matching one). However, the onset of this mismatch cost appeared earlier in (8a) (where the reflexive and its antecedent *the soldier* are co-arguments of the same verb *positioned*), relative to (8b) (where the reflexive is embedded in a picture noun phrase, and is thus not a direct co-argument of the antecedent)[5]. This difference in the onset of the mismatch cost may indicate that the speed of dependency formation for reflexives is affected by the structure of the sentence—for example, it may be that initial retrieval processes consider co-arguments as potential antecedents, leading to an earlier formation of the dependency in (8a), and thus an earlier appearance of the mismatch cost.

The present research aims to follow up on these results by examining whether the onset of the mismatch cost for a reflexive is also affected by whether its antecedent is accessed via a raising or a nominal control dependency. There are several reasons why this may be the case. As mentioned above, the raising dependency can be resolved using purely structural information, while the control dependency requires a more complex evaluation of the control nominal's argument structure[6]. A second reason is related to the possibility that nominal control and raising may involve different syntactic representations. For example, if raising does not involve an empty infinitival subject, as suggested by the LTAG view XTAG Research Group (1998), then the dependency between a reflexive and its antecedent in an example like (6) is direct. This would contrast with nominal control, where the dependency would be assumed to be mediated by an empty subject. It may therefore be plausible to assume that a direct dependency might be processed more quickly than an indirect one, leading to an earlier onset of the mismatch cost for raising, relative to control.

In the remainder of this paper, we report four experiments that were designed to examine the formation of raising and nominal control dependencies. Experiment 1 establishes a baseline by examining reflexive-antecedent dependencies that are not mediated by raising or control. Experiment 2 directly compares raising and nominal control dependencies, without distractors, thus allowing us to test for differences in the onset of the mismatch cost. Then, in Experiments 3 and 4, we include distractors, focusing on nominal control (Experiment 3) and finally raising (Experiment 4).

We believe that it is important to consider a wide range of dependency types in our search to understand memory access and dependency formation in sentence comprehension. Raising and control dependencies offer a potentially interesting domain of enquiry, because they differ in theoretically relevant ways, while sharing considerable surface similarity. We also believe that it is important to consider not only simple direct dependencies between overt linguistic elements within a sentence, but also indirect dependencies, such as the reflexive-antecedent dependencies that are mediated by raising or control, which we examine here. We hope that the four experiments that we report below add new data points that will increase our understanding of the factors that affect retrieval interference, and will also provide a first step toward gaining a picture of retrieval in indirect dependencies.

2. Experiment 1

In Experiment 1, we establish a baseline by examining the processing of a direct dependency between a reflexive and its antecedent, without incorporating raising or control dependencies. In all other respects, the sentences are very similar to those used in the other experiments.

2.1. Materials and Methods

2.1.1. Participants

Thirty-two participants from the University of Edinburgh community were paid to participate in the experiment. All were native speakers of English, with normal or corrected-to-normal vision, and none reported any reading disability. All of the participants in the four experiments reported in this paper gave informed consent to take part. The research protocol was approved by the Psychology Research Ethics Committee, of the University of Edinburgh.

2.1.2. Stimuli

The stimuli of Experiment 1 were similar to (9)[7]:

9a. **Accessible-match Inaccessible match:**
John didn't trust Tom but was kind to himself appropriately and very sincerely.

[5]Co-argumenthood is an important notion in certain theoretical treatments of anaphoric binding (see Reinhart and Reuland, 1993, for a well-known example of such a theory).

[6]However, we acknowledge that the extra complexity of the control nominal dependency might not necessarily result in slower access. As pointed out by a reviewer, it is possible that the richer lexical information would in fact make access faster.

[7]The stimuli for all experiments are available in the Supplemental Material.

9b. **Accessible-match Inaccessible mismatch:**
John didn't trust Amy but was kind to himself appropriately and very sincerely.

9c. **Accessible-mismatch Inaccessible match:**
Mary didn't trust Tom but was kind to himself appropriately and very sincerely.

9d. **Accessible-mismatch Inaccessible mismatch:**
Mary didn't trust Amy but was kind to himself appropriately and very sincerely.

Given this design, the main effect of *accessible antecedent* matching can be used to gauge the time at which the parser first becomes sensitive to the gender matching between the reflexive and its grammatically correct antecedent, and can thus, given the assumptions above, be used as a measure of how quickly the structurally appropriate antecedent is identified. For example, if this effect is initially found in first fixation duration it would suggest that the antecedent is identified very early (see the Section 2.2 below for details of the eye-movement measures). Moreover, the effect of *inaccessible antecedent* (or its interaction with *accessible antecedent*) is informative about any effect of interference. For example, if the mismatch cost for the accessible antecedent is reduced where the inaccessible antecedent matches (11d) relative to when it does not (11c), this could be indicative of a *facilitatory interference* effect. Alternatively, if we find evidence for extra processing difficulty when both potential antecedents match the reflexive (11a) relative to when only the accessible antecedent matches (11b), then this could be interpreted as *inhibitory interference*. Given the experimental design, either of these two patterns, or their combination, would result in an interaction between the two experimental factors. Specifically, *facilitatory interference*, on its own, would result in a difference between the two accessible mismatch conditions (i.e., a penalty for inaccessible mismatch relative to inaccessible match), with no difference among the accessible match conditions. *inhibitory interference*, on its own, would result in a difference between the two accessible match conditions (i.e., a penalty for inaccessible match relative to inaccessible mismatch), with no difference among the accessible mismatch conditions. Finally, a combination of these two interference profiles would result in a cross-over pattern of means.

2.1.3. Procedure

The experiment was carried out using an SR Research Eyelink 1000 eye-tracker, with a sampling rate of 1000 Hz. The tracker was used in tower mode. Only the right eye was tracked, although viewing was binocular. The eye-tracker was calibrated at the start of each participant's session, with recalibration being carried out as necessary through the experiment. At the start of each trial, a black box appeared at the left of the screen, in the position of the first character of text. When a stable fixation was detected in this position, the box disappeared, and the text appeared. The stimuli were presented in black on a white background, using Times Roman 16 point. The stimuli were presented in either one or two lines of text. In all cases, the critical reflexive was always placed at least two words before the end of a line.

The stimuli were combined with 102 filler sentences of varied sentence types. Thirty-six of the fillers were from an unrelated experiment on the processing of emotion words. A comprehension question followed around two thirds of all stimuli, including all of the experimental items (as an example, the question for (9) was "Was the kindness appropriate?"). The participant had to answer the question by pressing a button to select one out of two displayed answers. The stimuli were distributed into four lists, using latin square counterbalancing.

2.2. Data Analysis

The sentences were divided into regions for the purpose of analysis. Here, we will report data for the following regions:

- **pre-critical region:** kind to
- **critical region (reflexive):** himself
- **spillover:** appropriately and
- **final:** very sincerely

Eye-fixation data were screened and manually corrected for vertical drift. Fixations of less than 80 ms were incorporated into larger fixations within a distance of one character, and then we deleted any remaining fixations of less than 80 ms, as well as any over 1200 ms.

We will report data for five eye-movement measures. First fixation is the duration of the first fixation in a region, from the time the region is first entered from the left, until a subsequent fixation is made. First pass reading time is the sum of fixation durations within the region, from the time the region is first entered from the left, until the region is exited, either to the left or right. Go-past is the sum of fixation durations from the time the region is first entered from the left until it is exited to the right (including any fixations made to the left of the region). Total time is the summed duration of all fixations on the region. In the above measures, for any given trial, if the measure returned no data (e.g., if there were no fixations on the region), the trial was treated as a missing value in the analysis. Finally, Second Pass reading time is the summed duration of all re-fixations on the region, after it has already been fixated for the first time. As is customary, for Second Pass reading time, trials that do not include a relevant fixation are included in the analysis as zero millisecond data points. Note that the first fixation measure is most meaningfully applied to single-word regions, which can be assumed to be processable within a single fixation. Thus, we report first fixation durations only for the critical reflexive region.

The results for all eye-movement measures were submitted to 2×2 Analyses of variance, aggregating by subject (F_1) and by item (F_2). The factors in the analysis were Accessible antecedent matching (match vs. mismatch) and Inaccessible antecedent matching (match vs. mismatch), both of which were within item and within participant.

2.3. Results

Two items were excluded from analysis because of typographical errors. Therefore, the item analysis is based on 38 items, with a corresponding reduction in the degrees of freedom for the F_2 analysis. Means for Experiment 1 are presented in **Table 1**, and statistical results are presented in **Table 2**.

TABLE 1 | Means (and Standard Errors) for Experiment 1.

Region		Pre-critical	Critical	Spillover	Final
Accessible	**Inaccessible**				
FIRST FIXATION					
Match	Match	–	239 (9)	–	–
Match	Mismatch	–	245 (9)	–	–
Mismatch	Match	–	251 (8)	–	–
Mismatch	Mismatch	–	262 (10)	–	–
FIRST PASS					
Match	Match	308 (16)	256 (12)	517 (35)	397 (24)
Match	Mismatch	329 (18)	271 (10)	500 (29)	465 (29)
Mismatch	Match	337 (19)	286 (14)	488 (32)	412 (27)
Mismatch	Mismatch	334 (21)	291 (13)	508 (31)	425 (30)
GO-PAST					
Match	Match	396 (26)	343 (22)	722 (64)	1482 (119)
Match	Mismatch	388 (21)	357 (19)	707 (54)	1463 (123)
Mismatch	Match	438 (42)	367 (18)	816 (61)	1667 (159)
Mismatch	Mismatch	403 (25)	378 (26)	914 (66)	1628 (167)
SECOND PASS					
Match	Match	224 (29)	153 (21)	298 (32)	126 (17)
Match	Mismatch	190 (24)	143 (19)	292 (30)	133 (20)
Mismatch	Match	244 (34)	176 (19)	360 (42)	139 (22)
Mismatch	Mismatch	223 (22)	179 (19)	392 (40)	159 (23)
TOTAL TIME					
Match	Match	525 (34)	404 (25)	813 (51)	533 (33)
Match	Mismatch	514 (32)	406 (23)	795 (47)	609 (38)
Mismatch	Match	565 (45)	447 (27)	845 (62)	560 (38)
Mismatch	Mismatch	548 (39)	457 (24)	898 (59)	600 (47)

As in previous work (e.g., Sturt, 2003), there was very early evidence for a mismatch cost for the accessible antecedent; the effect appeared in the first fixation duration on the critical reflexive (the earliest measurable point, given the eye-tracking methodology), and this was mirrored in first pass times in the same region. However, this early effect did not interact with the matching of the inaccessible antecedent. The inaccessible antecedent had a marginal effect on fixation times in the final region in Total Time and First Pass. The pattern was for the inaccessible mismatch condition to lead to longer reading times than the corresponding match condition. In First Pass, this effect in the final region interacted with the accessible antecedent, but only in the analysis by item—the reading time penalty for the inaccessible mismatch condition (relative to inaccessible match) was greater when the reflexive matched the accessible antecedent (465 vs. 397 ms; a relative cost of 68 ms; both F's > 6, both p's < 0.02) than when it did not (425 vs. 412 ms; a relative cost of 13 ms; both F's < 1).

2.4. Discussion

This experiment sets a baseline using direct reflexive-antecedent dependencies, for the following experiments, where the reflexive-antecedent dependency is mediated by raising and control. We find that an early main effect of accessible antecedent on the critical reflexive, indicating an early onset of the mismatch cost. There was little evidence of either inhibitory interference or facilitatory interference, at least in the early measures. Later effects suggest a difficulty for *mismatching*, relative to *matching* inaccessible antecedents. This pattern may possibly be interpreted in terms of facilitatory interference. However, this interpretation is not straightforward, as the effect of inaccessible antecedent appeared as a main effect rather than the interaction predicted by current memory models. In fact, the marginal interaction in First Pass in the final region shows, if anything, that the facilitatory effect was larger for the *grammatical* sentences than the ungrammatical sentences, which is not the pattern that is expected for facilitatory interference. In addition, we note that First Pass reading times are often hard to interpret in the final region of a sentence, due to the possibility of relatively short initial fixations preceding regressions out of the region (see Sturt, 2007; Sturt et al., 2010).

3. Experiment 2

Experiment 2 used reflexive-antecedent dependencies that are mediated by raising or nominal control, depending on condition, in simple sentences that do not contain distractor noun phrases. This allows us to determine whether there are any baseline differences in the time-course of processing of raising-mediated and control-mediated dependencies, over and above those that may be explained in terms of interference effects. If the dependencies are formed more quickly when they are mediated by raising than when they are mediated by control, then we would expect the onset of the mismatch cost to appear earlier in the eye-movement record in raising than in control.

4. Materials and Methods

4.1. Participants

Thirty-two participants from the University of Edinburgh community were paid to participate in the experiment. All were native speakers of English, with normal or corrected-to-normal vision, and none reported any reading disability.

4.2. Stimuli

There were 40 stimuli, which were similar to those given in (10):

10a. **Control Match:**
I was surprised at John's agreement to be kind to himself appropriately and very sincerely.

10b. **Control Mismatch:**
I was surprised at John's agreement to be kind to herself appropriately and verysincerely.

10c. **Raising Match:**
It was surprising that John seemed to be kind to himself appropriately and very sincerely.

10d. **Raising Mismatch:**
It was surprising that John seemed to be kind to herself appropriately and very sincerely.

The design manipulated sentence type (Raising vs. Control), and gender matching (Match vs. Mismatch).

TABLE 2 | Anova results for Experiment 1 ($^{+}p < 0.1$; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$).

Region	Pre-critical		Critical		Spillover		Final	
	$F_{1(1,31)}$	$F_{2(1,37)}$	$F_{1(1,31)}$	$F_{2(1,37)}$	$F_{1(1,31)}$	$F_{2(1,37)}$	$F_{1(1,31)}$	$F_{2(1,37)}$
FIRST FIXATION								
Accessible	–	–	5.51*	5.54*	–	–	–	–
Inaccessible	–	–	2.21	1.57	–	–	–	–
Acc * Inacc	–	–	<1	<1	–	–	–	–
FIRST PASS								
Accessible	4.37*	1.34	6.98*	7.20*	<1	<1	<1	<1
Inaccessible	<1	<1	1.85	2.72	<1	<1	6.18*	3.84+
Acc * Inacc	1.32	1.39	<1	<1	1.23	<1	1.46	5.81*
GO-PAST								
Accessible	1.87	2.20	1.45	2.84	17.98***	10.36**	3.18+	6.77*
Inaccessible	1.47	1.45	<1	<1	1.50	<1	<1	<1
Acc*Inacc	<1	<1	<1	<1	2.10	2.26	<1	<1
SECOND PASS								
Accessible	1.26	3.74+	5.15*	7.25*	12.84**	10.65**	<1	1.82
Inaccessible	3.04+	1.77	<1	<1	<1	<1	<1	1.19
Acc * Inacc	<1	<1	<1	<1	<1	<1	<1	<1
TOTAL TIME								
Accessible	2.06	4.08+	10.40**	14.88***	6.16*	8.43**	<1	<1
Inaccessible	<1	<1	<1	<1	<1	<1	7.12*	4.04+
Acc * Inacc	<1	<1	<1	<1	2.21	2.30	<1	2.07

As we mentioned in the introduction, we assume that the raising and nominal control dependencies are initially formed around the point where *to be kind* is received in the input, and that there is a second retrieval event that is triggered by the reflexive, which is also mediated by control (10a,b) or raising (10c,d). It is this second retrieval event that we are measuring in this experiment, using the gender-mismatch paradigm. It is important to recognize that this second retrieval event involves two dependencies, (a) a reflexive-antecedent dependency (between *himself* and its direct antecedent, the empty subject of the infinitival clause), and (b) a raising or control dependency (between the empty subject and *John*). The logic of the design is that, as the relevant aspects of the reflexive-antecedent dependency are essentially identical between the raising and control conditions, any differences that we might find in the onset of the mismatch cost must be due to processing differences related to raising or control.

4.3. Procedure

The sentences were divided into regions for the purpose of analysis as shown below.

- **pre-critical region:** kind to
- **critical region:** himself
- **spillover:** appropriately and
- **final:** very sincerely

The pre-critical region consisted of the two words immediately preceding the critical reflexive. The spillover region consisted of the two words immediately following the reflexive. The final region consisted of the last two words of the sentence.

4.4. Results

The means are given in **Table 3**, and statistical results in **Table 4**.

As in Experiment 1, there was an early effect of matching, indicating a cost for the gender mis-matching items. This effect is present in all eye-movement measures on the critical reflexive, and persisted into the spill-over region. As this includes measures indicative of early processing, such as first-pass reading time, and first fixation, this suggests that the antecedent was identified equally quickly, whether the dependency was mediated by raising or control. In fact, the timing was in line with the co-argument reflexive-antecedent dependencies examined in Experiment 1. This early mismatch cost did not interact with structure. In addition, a main effect of structure type suggested that the control sentences were harder to read than the raising sentences (see Total Time, pre-critical region, and First Pass, spill-over region). However, this overall difference is not the focus of the current investigation.

4.5. Discussion

In this experiment, we investigated sentences where raising and control dependencies were combined with reflexive-antecedent dependencies. The main effect of matching appeared in both first fixation and first pass on the critical reflexive. This is the earliest detectable point given the eye-tracking methodology, and is in line with the timing of the accessible mismatch effect in the co-argument reflexive-antecedent dependencies examined in

TABLE 3 | Means (and Standard Errors) for Experiment 2.

Region		Pre-critical	Critical	Spillover	Final
FIRST FIXATION					
Control	Match	–	251 (8)	–	–
Control	Mismatch	–	265 (9)	–	–
Raising	Match	–	245 (9)	–	–
Raising	Mismatch	–	262 (9)	–	–
FIRST PASS					
Control	Match	343 (19)	274 (10)	567 (43)	412 (31)
Control	Mismatch	348 (19)	296 (14)	547 (43)	454 (42)
Raising	Match	339 (19)	274 (12)	516 (39)	419 (29)
Raising	Mismatch	323 (16)	300 (12)	515 (40)	429 (27)
GO-PAST					
Control	Match	449 (26)	366 (26)	819 (73)	1914 (294)
Control	Mismatch	507 (35)	454 (44)	928 (79)	2006 (286)
Raising	Match	450 (32)	366 (23)	727 (64)	1767 (231)
Raising	Mismatch	447 (27)	462 (42)	1030 (85)	1848 (247)
SECOND PASS					
Control	Match	293 (50)	189 (30)	438 (78)	186 (28)
Control	Mismatch	315 (49)	245 (35)	496 (70)	198 (33)
Raising	Match	239 (33)	181 (30)	367 (44)	185 (32)
Raising	Mismatch	304 (45)	259 (37)	509 (55)	196 (30)
TOTAL TIME					
Control	Match	636 (60)	456 (33)	1006 (101)	602 (49)
Control	Mismatch	656 (54)	530 (39)	1043 (86)	659 (63)
Raising	Match	575 (47)	448 (32)	880 (67)	613 (50)
Raising	Mismatch	621 (52)	545 (41)	1023 (75)	633 (46)

Experiment 1. As the effect was not modulated by sentence structure, there is no indication of any difference in the time-course of antecedent identification, whether the dependency was mediated by raising or nominal control dependencies. However, the study was carried out as a baseline, and did not include distractor noun phrases. Thus, although the study suggests no clear difference in time-course between dependency types, it does not rule out that the two dependency types may be differentially susceptible to interference. In Experiments 3 and 4, we address this issue, by including distractor antecedents in Nominal control (Experiment 3) and Raising (Experiment 4) sentences.

5. Experiment 3

Experiment 3 was designed to test the susceptibility of nominal control dependencies to interference.

5.1. Materials and Methods

5.1.1. Participants

Thirty-two new participants from the University of Edinburgh community were paid to participate in the experiment. All were native speakers of English, with normal or corrected-to-normal vision, and none reported any reading disability.

5.1.2. Stimuli

There were forty experimental items similar to those in (11)[8]:

11a. **Accessible-match Inaccessible match:**
John's agreement with Tom to be kind to himself was surprising to everyone.

11b. **Accessible-match Inaccessible mismatch:**
John's agreement with Amy to be kind to himself was surprising to everyone.

11c. **Accessible-mismatch Inaccessible match:**
Mary's agreement with Tom to be kind to himself was surprising to everyone.

11d. **Accessible-mismatch Inaccessible mismatch:**
Mary's agreement with Amy to be kind to himself was surprising to everyone.

The items all used giver control nominals (exemplified by *agreement* in 11a–d; Culicover and Jackendoff, 2001), with the result that the accessible antecedent for the reflexive was always the genitive subject of the control nominal (e.g., *John's* in 11). The design orthogonally manipulated the gender matching of the reflexive with the accessible antecedent (e.g., *Mary* vs. *John*) and with the inaccessible antecedent (*Tom* vs. *Amy*).

5.1.3. Procedure

All relevant aspects of the procedure were identical to Experiment 1.

We will report analyses based on the following regions:

- **pre-critical region:** kind to
- **critical region (reflexive):** himself
- **spillover:** was surprising
- **final:** to everyone.

5.2. Results

The means are given in **Table 5**, and statistical results in **Table 6**.

The results show evidence of a mismatch cost for the accessible antecedent in go-past, total time and second pass in the critical reflexive region. Go-past and Total times on this region were not modulated by any interactions with inaccessible antecedent matching. There was some marginal evidence that reading was affected by the inaccessible antecedent, in measures of later processing. The effect of inaccessible matching was significant (in the subjects analysis only) in second pass in both the critical and spillover regions; as in Experiment 1, the tendency was for inaccessible mismatch conditions to be read more slowly than inaccessible match conditions.

There was a marginal interaction between accessible and inaccessible gender matching in go-past and first-pass reading time in the spill-over region. This interaction was examined using pairwise comparisons, to test the simple effect of inaccessible antecedent, within (a) the accessible match conditions, and (b) the accessible mismatch conditions. For first-pass reading times,

[8]The position of the control nominal in the sentence is different in Experiment 3 from Experiment 2. This is because Experiment 2 needed to use a sentence frame that allowed a comparison of nominal control with raising, while Experiment 3 only used nominal control, so could use a more naturally suited sentence frame.

TABLE 4 | Anova results for Experiment 2 ($^{+}p < 0.1$; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$).

Region	Pre-critical		Critical		Spillover		Final	
	$F_{1(1,31)}$	$F_{2(1,39)}$	$F_{1(1,31)}$	$F_{2(1,39)}$	$F_{1(1,31)}$	$F_{2(1,39)}$	$F_{1(1,31)}$	$F_{2(1,39)}$
FIRST FIXATION								
Structure	–	–	1.61	<1	–	–	–	–
Matching	–	–	8.29**	8.30**	–	–	–	–
Struc * Match	–	–	<1	<1	–	–	–	–
FIRST PASS								
Structure	1.97	2.29	<1	<1	4.47*	4.02^{+}	<1	<1
Matching	<1	<1	13.39**	10.39**	<1	<1	2.32	2.20
Struc * Match	<1	<1	<1	<1	<1	<1	<1	<1
GO-PAST								
Structure	1.87	1.97	<1	<1	<1	<1	2.10	<1
Matching	1.79	1.36	8.65**	16.73***	13.93**	20.26***	1.44	1.36
Struc * Match	2.69	1.83	<1	<1	2.77	3.60^{+}	<1	<1
SECOND PASS								
Structure	2.84	2.27	<1	<1	<1	<1	<1	<1
Matching	4.79*	8.75**	13.58**	7.68**	11.49**	14.21**	<1	<1
Struc * Match	<1	1.43	<1	<1	3.18^{+}	1.69	<1	<1
TOTAL TIME								
Structure	6.04*	5.10*	<1	<1	2.63	4.51*	<1	<1
Matching	1.85	3.09^{+}	18.16***	9.72**	8.36**	6.37*	2.85	1.37
Struc * Match	<1	<1	<1	<1	2.91^{+}	2.29	<1	<1

neither of these pairwise comparisons was reliable (all p's > 0.1). However, for Go-Past time, pairwise comparison (b) (i.e., within the accessible mismatch conditions) showed significantly faster reading times for the inaccessible match (797 ms) relative to inaccessible mismatch (948 ms) [$F_{1(1,31)} = 5.72$, $p < 0.05$; $F_{2(1,39)} = 4.11$, $p < 0.05$], while comparison (a) (i.e., within the accessible match conditions) showed a non-significant difference in the opposite direction (794 ms vs. 768 ms) [both p's < 1].

5.3. Discussion

The first appearance of the mismatch cost for the accessible antecedent was in the Go-past measure on the critical reflexive. This shows that the ungrammatical dependency in the accessible mismatch conditions disrupted processing fairly quickly—soon after the participants initially fixated the reflexive, and before they moved on to fixate subsequent words.

The experiment did not show strong interference effects, but we believe that the results for Go-past in the spill-over region are highly suggestive, at the very least. Despite the fact that the interaction was marginal, the results of the pairwise comparisons are as predicted for facilitatory interference, since the cost for the accessible mismatch was significantly reduced when the inaccessible antecedent matched the gender of the reflexive, relative to when it did not.

6. Experiment 4

Experiment 1 showed no evidence that could be straightforwardly interpreted in terms of interference, for direct reflexive-antecedent dependencies that were not mediated by raising or control. Experiment 3 showed some marginal evidence for facilitatory interference, in the resolution of reflexive-antecedent dependencies that were mediated by nominal control. In Experiment 4, we examine reflexive-antecedent dependencies that are mediated by raising, using a design that is analogous to that of Experiment 3.

6.1. Materials and Methods

6.1.1. Participants

Thirty-two new participants from the University of Edinburgh community were paid to participate in the experiment. All were native speakers of English, with normal or corrected-to-normal vision, and none reported any reading disability.

6.1.2. Stimuli

There were 40 stimuli, which were similar to those in (12):

12a. **Accessible-match Inaccessible match:**
John seemed to Tom to be kind to himself appropriately and very sincerely.

TABLE 5 | Means (and Standard Errors) for Experiment 3.

Region		Pre-critical	Critical	Spillover	Final
Accessible	**Inaccessible**				
FIRST FIXATION					
Match	Match	–	238 (8)	–	–
Match	Mismatch	–	230 (8)	–	–
Mismatch	Match	–	245 (10)	–	–
Mismatch	Mismatch	–	239 (9)	–	–
FIRST PASS					
Match	Match	313 (19)	254 (9)	487 (26)	358 (40)
Match	Mismatch	322 (16)	257 (11)	462 (21)	357 (29)
Mismatch	Match	317 (17)	266 (12)	440 (22)	334 (29)
Mismatch	Mismatch	331 (20)	260 (13)	465 (26)	331 (31)
GO-PAST					
Match	Match	415 (29)	307 (14)	794 (76)	1138 (112)
Match	Mismatch	406 (22)	340 (25)	768 (58)	1165 (107)
Mismatch	Match	399 (20)	386 (36)	797 (83)	1198 (147)
Mismatch	Mismatch	414 (27)	388 (39)	948 (71)	1305 (158)
SECOND PASS					
Match	Match	190 (26)	97 (15)	180 (25)	60 (12)
Match	Mismatch	202 (27)	114 (18)	217 (26)	64 (12)
Mismatch	Match	205 (31)	134 (23)	233 (37)	58 (15)
Mismatch	Mismatch	239 (34)	174 (23)	272 (31)	61 (11)
TOTAL TIME					
Match	Match	500 (31)	343 (19)	664 (43)	470 (55)
Match	Mismatch	514 (32)	364 (21)	679 (36)	443 (40)
Mismatch	Match	517 (34)	394 (27)	673 (50)	407 (34)
Mismatch	Mismatch	565 (39)	434 (29)	733 (45)	414 (42)

12b. **Accessible-match Inaccessiblex mismatch:**
John seemed to Amy to be kind to himself appropriately and very sincerely.

12c. **Accessible-mismatch Inaccessible match:**
Mary seemed to Tom to be kind to himself appropriately and very sincerely.

12d. **Accessible-mismatch Inaccessible mismatch:**
Mary seemed to Amy to be kind to himself appropriately and very sincerely.

The items used a raising construction incorporating an experiencer argument (e.g., *to Amy*). The accessible antecedent for the reflexive was always the subject of the main clause (e.g., *Mary*), while the experiencer argument was always an inaccessible antecedent. The design orthogonally manipulated the gender matching of accessible and inaccessible antecedents.

Recall from the introduction of this paper that we expected raising-mediated dependencies to be less susceptible to interference than the control-mediated dependencies that we examined in Experiment 3. The introduction lists some reasons for this expectation, such as potential differences in access mechanisms, retrieval cues, or syntactic representation. Here, we will briefly elaborate on how differences in syntactic representation may lead to different retrieval profiles, using Lexicalized Tree Adjoining Grammar (LTAG) as an example grammatical framework. In LTAG, the matrix subject in (12) (e.g., *John*), would be assumed to occupy the subject position of a predicative elementary tree[9], projected by *kind*, without this relationship being mediated by an empty subject position in the infinitival clause (see XTAG Research Group, 1998, p.106–107). In contrast, in the nominal control stimuli (see 11 in Experiment 3), *John's* would be assumed to occupy the specifier position of *agreement*, while the infinitival clause would have an empty subject, occupied by the empty element PRO (see XTAG Research Group, 1998, p.97–101, for an analysis of verbal control)[10]. Thus, according to the LTAG proposal, *John* is effectively a co-argument of *himself* in the raising sentences, but is not a direct co-argument of *himself* in the nominal control sentences. Accordingly, this approach would predict that the interference profile for raising-mediated dependencies would pattern like the co-argument dependencies examined in Experiment 1, rather than like the control-mediated dependencies examined in Experiment 2.

6.1.3. Procedure

As the experiment was based on Experiment 3, the regions were defined identically:

- **pre-critical region:** kind to
- **critical region (reflexive):** himself
- **spillover:** appropriately and
- **final:** very sincerely

6.2. Results

The means are given in **Table 7**, and statistical results in **Table 8**.

As with Experiment 3, the first evidence of a mismatch cost for the accessible antecedent is in the go-past measure on the critical reflexive region, with a main effect of accessible matching. This main effect persists until the final region, and is found (in the critical and spill-over regions) also in the Total Time and Second Pass measures.

Second pass reading time shows a significant interaction between accessible and inaccessible matching in both the critical and spill-over regions. Pair-wise comparisons on both of these regions show a pattern consistent with facilitatory interference: there was a reliable difference between the two accessible mismatch conditions, with longer second pass times when the inaccessible antecedent also mismatches the reflexive than when it does not {critical region: 266 ms vs. 215 ms [$F_{1(1,\,31)} = 4.10$, $p = 0.052$; $F_{2(1,\,39)} = 5.66$, $p < 0.05$]; spill-over region: 519 ms vs. 398 ms [$F_{1(1,\,31)} = 6.92$, $p < 0.05$; $F_{2(1,\,39)} = 7.68$, $p < 0.01$]}. In contrast, the difference between the two accessible match conditions was in the other direction, but much smaller, and non-significant (critical region: 149 ms vs. 153 ms; spill-over region: 337 ms vs. 359 ms; all F's < 1).

On the final region, there were marginal interactions in both Second pass and Go-past (significant only by subjects for Go-past, and only by items for Second-pass). Pairwise comparisons revealed patterns of significance that were suggestive of

[9] An elementary tree can be thought of as a lexically-stored extended projection of a head word.

[10] The XTAG grammar does not treat nominal control, but we assume that the analysis would be analogous to that of verbal control.

TABLE 6 | Anova results for Experiment 3 ($^{+}p < 0.1$; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$).

Region	Pre-critical		Critical		Spillover		Final	
	$F_{1(1,31)}$	$F_{2(1,39)}$	$F_{1(1,31)}$	$F_{2(1,39)}$	$F_{1(1,31)}$	$F_{2(1,39)}$	$F_{1(1,31)}$	$F_{2(1,39)}$
FIRST FIXATION								
Accessible	–	–	1.61	3.21^{+}	–	–	–	–
Inaccessible	–	–	1.95	<1	–	–	–	–
Acc * Inacc	–	–	<1	<1	–	–	–	–
FIRST PASS								
Accessible	<1	<1	<1	1.18	5.01*	1.94	1.67	1.00
Inaccessible	1.57	2.51	<1	<1	<1	<1	<1	<1
Acc * Inacc	<1	<1	<1	<1	3.19^{+}	3.57^{+}	<1	<1
GO-PAST								
Accessible	<1	<1	5.38*	7.18*	3.43^{+}	4.16*	1.17	1.21
Inaccessible	<1	<1	<1	1.88	3.01^{+}	1.29	<1	<1
Acc * Inacc	<1	<1	<1	<1	3.17^{+}	3.65^{+}	<1	<1
SECOND PASS								
Accessible	2.53	3.03^{+}	14.81**	16.63***	11.20**	7.42*	<1	<1
Inaccessible	1.44	1.07	5.38*	2.64	7.43*	1.85	<1	<1
Acc * Inacc	<1	<1	1.02	<1	<1	<1	<1	<1
TOTAL TIME								
Accessible	4.14^{+}	4.25*	17.78***	27.39***	3.13^{+}	1.44	2.76	<1
Inaccessible	3.29^{+}	1.40	3.87^{+}	2.15	3.91^{+}	<1	<1	<1
Acc * Inacc	<1	<1	<1	<1	<1	<1	<1	<1

inhibitory interference. Among the accessible match conditions, reading times were longer when the inaccessible antecedent also matched the reflexive, than when it did not {Go-past: 1741 ms vs. 1518 ms; [$F_{1(1,31)} = 4.63$, $p < 0.05$; $F_2(1, 39) = 1.60$, $p = 0.21$]; Second pass: 206 ms vs. 150 ms; [$F_{1(1,31)} = 4.32$, $p < 0.05$; $F_2(1, 39) = 5.54$, $p < 0.05$]}. Among the accessible mismatch conditions, the difference was in the opposite direction, but was not reliable (Go-past: 1832 ms vs. 1955 ms; both F's < 1.2, both p's > 0.3; Second pass: 208 ms vs. 221 ms; both F's < 1).

6.3. Discussion

As in Experiment 3, the first indication of a mismatch cost was the main effect of accessible matching in the critical reflexive region. There was also some clear evidence of facilitatory interference, in second-pass reading times in the critical and spill-over regions. Thus, the interference profile for this raising-mediated dependency resembled that of the control-mediated dependencies in Experiment 3, and differed from the the co-argument dependencies examined in Experiment 1, where no strong evidence of interference was found. Thus there is no evidence for the hypothesis that raising-mediated dependencies should show reduced interference effects relative to control-mediated dependencies, based on differences in the access mechanism, retrieval cues, or syntactic representation.

Unlike any of the previous experiments, there was also some evidence of inhibitory interference. However, this result is hard to interpret, as it comes from second-pass and go-past measures on the final region, and could thus be contaminated by wrap-up effects, or preparations for the comprehension question. Here, second pass time is based on the fixations that are made when the final region is re-fixated, following any initial regressions back to earlier points in the sentence, and before the button is pressed to indicate the end of the trial. Go-past time on this region also includes these fixations. Thus, if inhibitory interference is indeed present, it occurred very late in the trial, probably during sentence-final wrap-up.

7. General Discussion

The above experiments were designed to examine the interference profile, and speed of dependency formation, for raising and nominal control dependencies. We began with the hypothesis that nominal control dependencies would be more subject to interference, and processed more slowly, than raising dependencies. This prediction was not confirmed overall. In the following, we will discuss the issues of time-course and interference in turn.

Experiment 1 established a baseline using reflexive-antecedent dependencies without the involvement of raising or control, and it replicated previous work in showing that gender mismatching between a reflexive and its accessible antecedent can slow down processing as early as the first fixation on the reflexive. Experiment 2 further established that, in the absence of inaccessible distractor antecedents, dependencies that were mediated by raising and nominal control elicited an equally early onset of the gender mismatch difficulty. Experiments 3 (control) and 4 (raising) included inaccessible distractor antecedents. These experiments showed the accessible mismatching cost on the critical reflexive

TABLE 7 | Means (and Standard Errors) for Experiment 4.

Region		Pre-critical	Critical	Spillover	Final
Accessible	**Inaccessible**				
FIRST FIXATION					
Match	Match	–	246 (8)	–	–
Match	Mismatch	–	255 (8)	–	–
Mismatch	Match	–	257 (8)	–	–
Mismatch	Mismatch	–	263 (8)	–	–
FIRST PASS					
Match	Match	313 (18)	277 (9)	510 (29)	442 (28)
Match	Mismatch	310 (15)	289 (9)	518 (38)	458 (32)
Mismatch	Match	318 (19)	293 (11)	534 (31)	422 (24)
Mismatch	Mismatch	313 (14)	298 (11)	498 (26)	412 (24)
GO-PAST					
Match	Match	407 (25)	333 (15)	775 (61)	1741 (177)
Match	Mismatch	404 (21)	379 (24)	782 (60)	1518 (152)
Mismatch	Match	400 (23)	427 (49)	882 (57)	1832 (211)
Mismatch	Mismatch	442 (34)	461 (35)	960 (67)	1955 (212)
SECOND PASS					
Match	Match	220 (27)	153 (23)	359 (37)	206 (29)
Match	Mismatch	225 (26)	149 (17)	337 (36)	150 (21)
Mismatch	Match	294 (37)	215 (21)	398 (44)	208 (33)
Mismatch	Mismatch	323 (41)	266 (31)	519 (62)	221 (28)
TOTAL TIME					
Match	Match	533 (32)	426 (28)	871 (52)	672 (47)
Match	Mismatch	533 (28)	431 (18)	857 (57)	618 (44)
Mismatch	Match	607 (41)	499 (25)	932 (55)	645 (41)
Mismatch	Mismatch	626 (45)	541 (31)	1016 (73)	647 (40)

in go-past, as well as in Total Time and Second pass, but, unlike in Experiments 1 and 2, not in first-fixation or first-pass.

Although we need to be cautious in interpreting between-experiment differences among first-pass measures, the control-mediated dependencies did not show an earlier onset for the mismatch cost than raising-mediated dependencies. Instead, the overall pattern of results is consistent with a slightly delayed onset of the mismatch cost for both the raising and control dependencies in Experiments 3 and 4 (go-past on the critical reflexive), relative to the co-argument reflexive-coargument dependencies tested in Experiment 1 (first-fixation and first pass on the critical reflexive). This delayed onset does not appear to be due to the involvement of raising or control dependencies *per se*, as Experiment 2, which used these dependencies (but without distractor phrases), showed an onset of mismatch difficulty in first-fixation, as early as that of Experiment 1. Rather, if anything, the delayed onset appears to be due to the presence of potentially interfering distractor phrases (whatever their gender marking), in conjunction with the use of raising and control dependencies. This should be interpreted as a preliminary finding, pending further investigation using more complex within-participant designs that have sufficient power to allow the statistical detection of potentially small differences in the onset of the mismatch cost. Such studies could also be supplemented by studies that allow a more direct measure of processing speed (e.g., Speed Accuracy Tradeoff; McElree et al., 2003).

Turning now to the discussion of interference, the results did not support the idea that dependencies mediated by nominal control would be more susceptible to interference than raising dependencies. On the one hand, assuming that the marginal interaction effect for Experiment 3 (control) reflects genuine interference, it may be the case that interference occurs earlier where the dependency is mediated by nominal control, compared with when it is mediated by raising. This is because the interference effect for Experiment 3 occurred shortly after readers had progressed forwards from the critical reflexive (i.e., in Go-past on the spill-over region), while in Experiment 4 (raising), the same region showed the effect only in second-pass. On the other hand, the interference effect seems to be stronger in Experiment 4 (raising) than in Experiment 3 (control). That is, in Experiment 3, the interaction between accessible and inaccessible gender matching was (marginally) significant only in go-past and first-pass reading time in the spill-over region, while in Experiment 4 it was fully significant in second-pass on the critical and spill-over regions (and marginal in two measures on the final region of the sentence). Thus, overall patterns of results do not support the hypothesis that the involvement of lexically-driven dependencies (control) leads to more interference than that of structurally-driven dependencies (raising), or that the access mechanism differs due to different retrieval cues or syntactic representations.

Both Experiment 3 (control) and Experiment 4 (raising) showed the profile expected for facilitatory interference. The pattern was such that when the reflexive did not match the gender of its structurally licit antecedent, the processing cost was reduced if there was an intervening distractor that matched the reflexive, relative to when the distractor did not match. The fact that interference was facilitatory, rather than inhibitory, accords with previous studies on subject-verb agreement (e.g., Wagers et al., 2009) and negative polarity licensing (e.g., Vasishth et al., 2008; Xiang et al., 2009), where interference was found only among ungrammatical (or otherwise degraded) conditions. Thus, like those earlier studies, our results do not tell us whether interference also affects grammatical, non-degraded dependencies. Moreover, our interference effect was found in measures that reflect fixation behavior after the reader has already progressed forwards from the critical reflexive, and thus, after the point where the mismatching of the accessible antecedent had started to cause a slow-down in reading. Because of this, we believe that the retrieval interference for these dependencies occurred, not during the initial retrieval of the antecedent, but during the repair process, possibly reflecting a *re-retrieval*, while readers searched for an acceptable interpretation of the ungrammatical sentences in the accessible mismatch conditions. In fact, the pattern of results can be summarized by saying that, while the *onset* of the accessible mismatch cost was unaffected by the gender of the distractor, the *duration* of this processing difficulty was affected by the distractor—i.e., the duration was shorter when the distractor matched the reflexive's gender.

Recall that the stimuli of Experiments 3 and 4 used dependencies that involved both reflexive-antecedent dependencies

TABLE 8 | Anova results for Experiment 4 ($^{+}p < 0.1$; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$).

Region	Pre-critical		Critical		Spillover		Final	
	$F_{1(1,31)}$	$F_{2(1,39)}$	$F_{1(1,31)}$	$F_{2(1,39)}$	$F_{1(1,31)}$	$F_{2(1,39)}$	$F_{1(1,31)}$	$F_{2(1,39)}$
FIRST FIXATION								
Accessible	–	–	2.37	1.18	–	–	–	–
Inaccessible	–	–	2.55	<1	–	–	–	–
Acc * Inacc	–	–	<1	<1	–	–	–	–
FIRST PASS								
Accessible	<1	<1	1.98	1.66	<1	<1	3.84^{+}	5.69*
Inaccessible	<1	<1	1.30	<1	<1	<1	<1	<1
Acc * Inacc	<1	<1	<1	<1	<1	<1	<1	<1
GO-PAST								
Accessible	<1	<1	7.43*	11.03**	15.90***	6.28*	7.09*	18.75***
Inaccessible	1.42	<1	2.49	2.79	1.41	<1	<1	<1
Acc * Inacc	1.75	1.28	<1	<1	<1	<1	7.42*	2.35
SECOND PASS								
Accessible	17.76***	26.27***	37.35***	31.87***	12.81**	14.47***	2.89^{+}	5.22*
Inaccessible	<1	1.03	1.61	2.59	3.79^{+}	3.18^{+}	1.83	1.03
Acc * Inacc	<1	<1	5.63*	4.93*	6.17*	5.94*	3.00^{+}	4.98*
TOTAL Time								
Accessible	8.82**	19.42***	25.03***	25.92***	12.08**	10.26**	<1	<1
Inaccessible	<1	<1	1.37	1.34	1.59	1.84	<1	<1
Acc * Inacc	<1	<1	1.46	<1	1.86	1.72	1.09	1.07

and control (or raising) dependencies. However, Experiment 1 used co-argument reflexive-antecedent dependencies with superficially very similar materials, and it showed no reliable evidence that could be straightforwardly interpreted in terms of the facilitatory interference or inhibitory interference. We therefore interpret Experiments 3 and 4 as support for the claim that reflexive-antecedent dependencies that are mediated by raising or control are processed more slowly and are more susceptible to interference than the co-argument dependencies when there is a distractor. Below, we outline a possible sequence of events that, while admittedly speculative, might explain how our raising and control sentences are affected by interference. For expository reasons, we focus on the *accessible-mismatch inaccessible-match* condition for Nominal Control in Experiments 3, as exemplified in (13): but analogous remarks also apply to Experiment 4.

13 **Nominal Control: Accessible mismatch, Inaccessible Match**
Mary's agreement with Tom Ø to be kind to himself was surprising to everyone.

As discussed in the introduction of this paper, in (13), we assume that the control dependency is initially formed around the point where *to be kind* is read. Note that the retrieval is effectively triggered by a null element (i.e., Ø), so the retrieval cue cannot include gender information, so this retrieval is not expected to have been affected by gender-based interference. It is not possible to measure interference at this early point in the sentence with our design (and indeed, the experiment was not designed to detect this). In fact, our experiments investigated a second retrieval event, related to the processing of the reflexive, but so far, we have not discussed this second event in any detail. Accordingly, we now sketch a possible account, based on our experimental results.

In (13), we assume that, following the initial retrieval event at *to be kind*, the null element Ø is associated with information about its antecedent *Mary*, including the fact that the antecedent is female. At *himself*, the null element Øis retrieved, and the gender incompatibility with the reflexive is registered, causing processing difficulty, and triggering a repair process. During the repair process, a new retrieval process is launched for Øto find its antecedent. This now includes a male gender cue due to the fact that *himself* is male. It is at this point that *Tom* can be misretrieved as the antecedent of Ø, leading to processing facilitation. Note that, in order for this mis-retrieval to occur in the way that we have suggested, it would have to be possible for the reflexive to use gender as a retrieval cue (contra Dillon et al., 2013), as least during the repair process.

An alternative to the above account is that the interference that we observe in sentences like (13) is driven entirely by a repair process involving the reflexive-antecedent dependency in response to the gender mismatch, without a new control-related retrieval being launched. Thus, for example, the error at the reflexive might reduce confidence in the structural encoding, increasing sensitivity to other gender-matching items in the sentence. However, such an account would still need to explain the apparent lack of interference in the direct reflexive-antecedent dependencies examined in Experiment 1. In other words, if the reflexive triggered an error-based retrieval (without invoking control or

raising dependencies) in Experiments 3 and 4, then why did it not also trigger an analogous error-based retrieval in Experiment 1? While this may potentially be due to other differences between the stimuli of Experiment 1 and the other experiments, we believe that the most likely reason is the fact that the relation between the reflexive and its antecedent is direct in Experiment 1, but mediated by control (or raising) in experiments 3 and 4, and that the control (or raising) dependency plays a role in the observed interference.

A question for future research is whether indirect dependencies (such as the ones that we examined in Experiments 3 and 4) are in general more prone to interference than direct dependencies (such as the one that we examined in Experiment 1).

Author Contributions

PS Supervised the running of the experiments, conducted the statistical analyses, and drafted the article. NK Supervised the creation of stimuli, and participated in writing the article. Both authors contributed equally to the planning of the research, and to the theoretical interpretation of the results.

Funding

This work was supported by the National Research Foundation of Korea grant funded by the Korean Government (NRF-2014S1A2A2028232), and by a British Academy/Leverhulme Small Research Grant (SG120693).

Acknowledgments

We thank Oliver Stewart and Gloria Chamorro for assistance with data collection.

Supplementary Material

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fpsyg.2015.00331/abstract

References

Badecker, W., and Straub, K. (2002). The processing role of structural constraints on the interpretation of pronouns and anaphora. *J. Exp. Psychol. Learn. Mem. Cogn.* 28, 748–769. doi: 10.1037/0278-7393.28.4.748

Chomsky, N. (1986). *Barriers.* Cambridge, MA: MIT Press.

Chow, W. Y., Lewis, S., and Phillips, C. (2014). Immediate structural constraints on antecedent retrieval in pronoun resolution. *Front. Psychol.* 5:630. doi: 10.3389/fpsyg.2014.00630

Culicover, P. W., and Jackendoff, R. (2001). Control is not movement. *Linguist. Inq.* 22, 493–512. doi: 10.1162/002438901750372531

Cunnings, I., and Sturt, P. (2014). Coargumenthood and the processing of reflexives. *J. Mem. Lang.* 75, 117–139. doi: 10.1016/j.jml.2014.05.006

Dillon, B., Mishler, A., Sloggett, S., and Phillips, C. (2013). Contrasting interference profiles for agreement and anaphora: experimental and modeling evidence. *J. Mem. Lang.* 69, 85–103. doi: 10.1016/j.jml.2013.04.003

Gordon, P. C., Hendrick, R., Johnson, M., and Lee, Y. (2006). Similarity-based interference during language comprehension: evidence from eye tracking during reading. *J. Exp. Psychol. Learn. Mem. Cogn.* 32, 1304–1321. doi: 10.1037/0278-7393.32.6.1304

Hornstein, N. (1999). Movement and control. *Linguist. Inq.* 30, 69–96. doi: 10.1162/002438999553968

Lewis, R., and Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cogn. Sci.* 29, 1–45. doi: 10.1207/s15516709cog0000/25

Lewis, R., Vasishth, S., and Van Dyke, J. (2006). Computational principles of working memory in sentence comprehension. *Trends Cogn. Sci.* 10, 447–454. doi: 10.1016/j.tics.2006.08.007

Martin, A., Nieuwland, M., and Carreiras, M. (2012). Event-related brain potentials index cue-based retrieval interference during sentence comprehension. *Neuroimage* 59, 1859–1869. doi: 10.1016/j.neuroimage.2011.08.057

McElree, B., Foraker, S., and Dyer, L. (2003). Memory structures that subserve sentence comprehension. *J. Mem. Lang.* 48, 67–91. doi: 10.1016/S0749-596X(02)00515-6

Phillips, C., Wagers, M., and Lau, E. (2011). "Grammatical illusions and selective fallibility in real-time language comprehension," in *Experiments at the Interfaces, Syntax and Semantics*, Vol. 37, ed J. Runner (Bingley: Emerald Publications), 153–186.

Reinhart, T., and Reuland, E. (1993). Reflexivity. *Linguist. Inq.* 24, 657–720.

Sturt, P. (2003). The time-course of the application of binding constraints in reference resolution. *J. Mem. Lang.* 48, 542–562. doi: 10.1016/S0749-596X(02)00536-3

Sturt, P. (2007). Semantic re-interpretation and garden path recovery. *Cognition* 105, 477–488. doi: 10.1016/j.cognition.2006.10.009

Sturt, P., Keller, F., and Dubey, A. (2010). Syntactic priming in comprehension: parallelism effects with and without coordination. *J. Mem. Lang.* 62, 333–351. doi: 10.1016/j.jml.2010.01.001

Van Dyke, J. A. (2007). Interference effects from grammatically unavailable constituents during sentence processing. *J. Exp. Psychol. Learn. Mem. Cogn* 33, 407–430. doi: 10.1037/0278-7393.33.2.407

Van Dyke, J. A., and McElree, B. (2006). Retrieval interference in sentence processing. *J. Mem. Lang.* 55, 157–166. doi: 10.1016/j.jml.2006.03.007

Vasishth, S., Brüssow, S., Lewis, R. L., and Drenhaus, H. (2008). Processing polarity: how the ungrammatical intrudes on the grammatical. *Cogn. Sci.* 32, 685–712. doi: 10.1080/03640210802066865

Wagers, M., Lau, E., and Phillips, C. (2009). Agreement attraction in comprehension: representations and processes. *J. Mem. Lang.* 61, 206–237. doi: 10.1016/j.jml.2009.04.002

Xiang, M., Dillon, B., and Phillips, C. (2009). Illusory licensing effects across dependency types. *Brain Lang.* 108, 40–55. doi: 10.1016/j.bandl.2008.10.002

XTAG Research Group. (1998). *A Lexicalized Tree Adjoining Grammar for English*. Technical report, Institute for Research in Cognitive Science, University of Pennsylvania.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2015 Sturt and Kwon. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 18 December 2015
doi: 10.3389/fpsyg.2015.01839

Structure Modulates Similarity-Based Interference in Sluicing: An Eye Tracking study

Jesse A. Harris *

Department of Linguistics, University of California, Los Angeles, Los Angeles, CA, USA

In cue-based content-addressable approaches to memory, a target and its competitors are retrieved in parallel from memory via a fast, associative cue-matching procedure under a severely limited focus of attention. Such a parallel matching procedure could in principle ignore the serial order or hierarchical structure characteristic of linguistic relations. I present an eye tracking while reading experiment that investigates whether the sentential position of a potential antecedent modulates the strength of similarity-based interference, a well-studied effect in which increased similarity in features between a target and its competitors results in slower and less accurate retrieval overall. The manipulation trades on an independently established Locality bias in sluiced structures to associate a *wh*-remnant (*which ones*) in clausal ellipsis with the most local correlate (*some wines*), as in *The tourists enjoyed some wines, but I don't know which ones.* The findings generally support cue-based parsing models of sentence processing that are subject to similarity-based interference in retrieval, and provide additional support to the growing body of evidence that retrieval is sensitive to both the structural position of a target antecedent and its competitors, and the specificity or diagnosticity of retrieval cues.

Keywords: working memory, similarity-based interference, ellipsis, eye tracking, sentence processing

OPEN ACCESS

Edited by:
Colin Phillips,
University of Maryland, USA

Reviewed by:
Masaya Yoshida,
Northwestern University, USA
Andrea Eyleen Martin-Nieuwland,
University of Edinburgh, UK

***Correspondence:**
Jesse A. Harris
jharris@humnet.ucla.edu

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: *10 June 2015*
Accepted: *13 November 2015*
Published: *18 December 2015*

Citation:
Harris JA (2015) Structure Modulates Similarity-Based Interference in Sluicing: An Eye Tracking study. Front. Psychol. 6:1839. doi: 10.3389/fpsyg.2015.01839

INTRODUCTION

The rapid formation of non-adjacent syntactic dependencies during online sentence comprehension offers intuitive evidence for the importance of an efficient retrieval system. Two well-studied cases are argument-verb dependencies, in which an argument must be related to its verb no matter the amount of intervening material (1a), and anaphoric dependencies, in which a pronominal element, like *him* or *one*, is associated with another co-referring expression, possibly from among multiple possibilities (1b). The noun phrase *the barber* must be retrieved from memory in each case: either in the subject-trace (gap) position t_1 (1a) or else as a co-referring expression (1b). Only in the former case is the dependency unambiguously determined by structure; the second case is simply the most plausible given the topicality of *the barber* and real world knowledge, as illustrated by fact that likely co-reference possibilities depend on the predicate (1c).

(1) a. The barber$_1$ (that John$_2$ wanted his son$_3$ to visit) t$_{1/*2,*3}$ died yesterday.
b. He$_{1>2>3}$ was 95 years old.
c. He$_{2>3,\#1}$ was saddened to hear the news.

Findings from both types of cases support recent cue-based parsing models of sentence processing in which all possible antecedents are activated in parallel through a fast, domain-general associative cue-matching procedure (for review of evidence and models see, e.g., Lewis et al., 2006; Van Dyke and Johns, 2012; Caplan and Waters, 2013). In such models, a retrieval cue—e.g., the verb *died* in (1a) or the pronoun *he* in (1b), initiates direct access to all possible targets (possible antecedents or dependencies) from memory. However, not all possible targets must be equally activated within memory: some targets might receive greater activation by virtue of sharing syntactic or semantic features with the retrieval cue, while other targets might receive less activation as a function of temporal decay (Van Dyke and Johns, 2012, for review, though a reviewer points out that decay as the primary source of forgetting is not strongly supported in the general memory literature, as demonstrated by Keppel and Underwood, 1962 and Waugh and Norman, 1965). Further, the allocation of attentional resources in memory is severely constrained, possibly limited to a single item within the focus of attention (e.g., McElree, 2001). Thus, the memory architecture employed in sentence processing strongly resembles the architecture thought to underlie domain-general tasks (e.g., McElree, 2000; McElree et al., 2003; Van Dyke and Lewis, 2003; Lewis and Vasishth, 2005; Lewis et al., 2006).

A major research question within the cue-based parsing literature addresses the extent to which domain-specific knowledge influences the retrieval process; in particular, whether syntactic constraints affect antecedent retrieval, and, if so, at precisely what stage of retrieval. Much of the research conducted thus far has examined whether syntactically inaccessible targets are considered viable candidates for retrieval. The two major schools of thought addressing this issue in the cue-based parsing literature are (i) structure-based accounts, and (ii) unconstrained-cue accounts. In the former, grammatical constraints filter the set of possible antecedents retrieved from memory by constraining the search set to grammatically permissible positions (e.g., Nicol and Swinney, 1989; Sturt, 2003; Xiang et al., 2009; Chow et al., 2014). Thus, items in structurally inaccessible positions would be effectively ignored in the retrieval process, though grammatically illicit items may feed repair processes triggered by retrieval failure (Chow et al., 2014). In contrast, the latter, unconstrained-cue, approach proposes that grammatical constraints are but one of the many possible factors guiding retrieval (e.g., Badecker and Straub, 2002; Lewis and Vasishth, 2005; Chen et al., 2012; Jäger et al., 2015). Badecker and Straub (2002), for example, propose that the set of possible antecedents are restricted not by tree geometry, but a number of other factors, such as attention or discourse importance. Elements in the focus of attention are often identified with discourse topics or a center, though details regarding how to define topics vary considerably (see, e.g., Chafe, 1976; Gundel, 1985; Gundel et al., 1993; Grosz et al., 1995). Under this type of account, syntactically inaccessible items might well interfere with retrieving a target, particularly if such items are highly salient in the discourse, though, of course, structure may be one of the factors that determines discourse salience. Although it should be clear that both viewpoints agree that structural information is a factor in retrieval, they permit very different types of mechanisms by which structure is utilized.

Unfortunately, results that would clearly arbitrate between the two camps are somewhat mixed. For example, while filler-gap dependencies are sensitive to syntactic islands (for review see, e.g., Sprouse et al., 2012), retrieving pronouns and reflexives has sometimes been found to be susceptible to interference from structurally inaccessible items (e.g., Badecker and Straub, 2002; Jäger et al., 2015; though see Chow et al., 2014 for a failure to replicate Badecker and Straub's results). The general finding is that syntactic information plays an important role in retrieval for at least certain types of dependencies (Van Dyke and McElree, 2006, 2011; Van Dyke, 2007; Dillon et al., 2014; Cunnings et al., 2015), even if it is currently unclear whether that role is to filter possible antecedents or provide weighted or probabilistic cues regarding their likelihood of matching the probe (Badecker and Straub, 2002; Cunnings et al., 2015). Further, it is unclear at this stage whether syntactic information should be treated as value in a feature bundle (e.g., Lewis and Vasishth, 2005), and if configurational relations like c-command inherently preclude such a treatment (see Alcocer and Phillips, 2012; Kush, 2013, for discussion).

The present research seeks to widen the empirical coverage of the issue of structural import by leveraging an independently established bias guiding the resolution of correlate-remnant pairs in clausal ellipsis (sluiced) structures. In (2), for example, the remnant *which (ones)* might be paired with either a subject correlate (*a few linguists*) or an object correlate (*some silly examples*), spelled out as (2a) and (2b), respectively.

(2) [A few linguists] gave [some silly examples], but I don't remember which (ones).

a. …I don't remember which linguists. (Subject correlate)
b. …I don't remember which examples. (Object correlate)

Although sluicing ellipsis permits correlates in both subject and object position above, it shows a strong preference for the latter (Frazier and Clifton, 1998). As mentioned, much research in content-addressable retrieval systems in language processing addresses whether grammatical restrictions govern the resolution of various types of anaphoric dependencies, especially the availability of syntactically illicit antecedents (e.g., Sturt, 2003; Martin et al., 2012, 2014; Cunnings et al., 2015; see also Phillips et al., 2011, for review). In contrast, correlates to sluiced remnants are merely heavily biased, rather than constrained grammatically. Thus, sluiced structures provide a potentially revealing counterpoint to studies investigating syntactic barriers to accessibility: if structurally defined preferences, in addition to grammatical restrictions, influence the retrieval process, the retrieval system may use structural information to privilege some products of memory over others. In which case, the retrieval system might be said to avail itself of domain-specific, linguistically defined structural preferences, in addition to hard-coded grammatical principles. In the remainder of the paper, I review the aspects of cue-based parsing that are most relevant for this study, along with the basic assumptions regarding the ellipsis structures explored here. I

then present two experiments which collectively support the idea that retrieving antecedents in sluiced sentences is subject to interference effects (congruent with Martin and McElree, 2011, discussed in detail below) and that the strength of such interference effects depends on the sentential position of possible targets and strength of the cue provided at retrieval.

Cue-based Parsing and Similarity-based Interference

Many models of how linguistic representations are encoded and retrieved during real-time language comprehension have been proposed (for review, see Van Dyke and Johns, 2012; Caplan and Waters, 2013). In contrast to cue-based retrieval systems, traditional models employ an operation that searches through items in memory, typically in a serial fashion (Dosher and McElree, 1992; McElree and Dosher, 1993), as developed for short-term memory in general (Sternberg, 1966, 1975). In these models, the path of the search varies according to whether the search queue starts with the first item encountered, as in forward search, or the most recent item, as in backward search. The basic prediction of search models is that retrieval time should increase as a function of search space within the path: the more items that must be searched through, the greater the search time required to do so. This prediction, however, has not been supported by Speed Accuracy Tradeoff (SAT) experiments which report that speed of reaching a stable judgment or interpretation is unaffected by the size of the putative search space, even though accuracy is (McElree and Dosher, 1989; McElree, 2000, 2006; McElree et al., 2003; Foraker and McElree, 2007, 2011; Martin and McElree, 2008, 2009, 2011). For example, McElree (2001) manipulated the amount of intervening material (underlined) between a relative clause filler *the book* and its object position gap t_1, shown in (3). Subjects were trained to make an acceptability judgment in response to tones at various post-sentence intervals. Measuring the rate at which responses achieved a stable interpretation (the asymptote) as a function of time, he found that while accuracy decreased as more material intervened between the filler and the gap, the speed at which the asymptote was reached did not (see also McElree et al., 2003).

(3) a. This was [the book]$_1$ that the editor admired t_1.
    b. This was [the book]$_1$ that the editor who the receptionist married admired t_1.
    c. This was [the book]$_1$ that the editor who the receptionist who quit admired t_1 married.

That accuracy, not the rate of reaching the asymptote, is affected by increasing items in memory not only provides an important argument against search models, it also suggests that the quality of the representations recovered is susceptible to interference from distractors. A content-addressable retrieval system is able to account for this tradeoff by proposing that all items are compared against the target in parallel, not by a search operation, but by an automatic associative cue-matching procedure which compares partial representations of items in memory (possibly encoded as bundles of semantic and grammatical features; Clark and Gronlund, 1996) against cues from the target (as in ACT-R models, Anderson et al., 2004; Lewis et al., 2006). Thus, linear or structural distance is irrelevant for retrieval times, but additional competitors interfere with the cue-matching process by introducing partial matches, thereby degrading the quality of the retrieved item. In other words, the greater the similarity or overlap between a target and its competitors, the greater the effect of interference (Crowder, 1976; Anderson and Neely, 1996).

Similarity-based interference, i.e., the failure to successfully distinguish a target from similar competitors in retrieval, has now been well documented both in general memory manipulations (Nairne, 2002; Öztekin and McElree, 2007) and in language processing contexts (Lewis, 1996; Gordon et al., 2001, 2002, 2004, 2006; Van Dyke and Lewis, 2003; Van Dyke and McElree, 2006, 2011; Van Dyke, 2011; Autry and Levine, 2014, among others). Dual task paradigms combine the two, so that a subject attempts to retain a list of words in working memory while processing a separate sentence for comprehension. When items in the memory set are similar to critical words in the sentence, performance decreases on both reading speed and comprehension accuracy (Fedorenko et al., 2006), especially if those items overlap with retrieval cues for long distance dependencies (Van Dyke and McElree, 2006).

What is less clear is the extent to which structural information modulates the accessibility of a target. On the one hand, grammatically illicit cues appear to license a syntactically dependent element in case of illusory licensing (including negative polarity items, NPI, Vasishth et al., 2008; Xiang et al., 2013, as well as agreement attraction, Wagers et al., 2009; Dillon et al., 2013). To illustrate, comprehenders often accept configurations in which an NPI licensor like *no* simply precedes, rather than c-commands, the NPI *ever*, as in *The restaurants that no newspapers have recommended in their reviews have ever gone out of business*, raising the possibility that configurational information may sometimes be ignored. On the other hand, several recent studies show that distractors within the same syntactic position more greatly interfere in the formation of long distance dependencies (Van Dyke and McElree, 2006, 2011; Dillon et al., 2013, 2014). For example, Van Dyke and McElree (2011) observed that syntactic constraints limit the amount of interference exerted by semantically similar distractors, without eliminating interference completely (also Van Dyke, 2007; Van Dyke and Johns, 2012). They attribute these results to greater disturbance from *retroactive interference*, in which retrieval is hampered by a distractor D that separates a cue C from its target T (schematically: T-**D**-C), than from *proactive interference*, in which retrieval is impeded by distractors processed before the target (**D**-T-C).

Similarly, Dillon and colleagues find an increased interference effect for structurally distant antecedents of a reflexive *ziji* in Mandarin Chinese—i.e., when a distractor intervened between a target and the reflexive. They propose that structurally local domains restrict the initial area for dependency formation.

(4) **Local search hypothesis:** The parser uses positional syntactic information during the retrieval of syntactic dependents, and

positional cues serve to restrict retrieval to constituents in some local syntactic domain.

The Local Search Hypothesis contrasts with content-addressable retrieval models that would explain putative effects of Locality in terms of activation decay (see comments in Lewis et al., 2006). Lewis and Vasishth (2005), for example, sharply discriminate items within the focus of attention from those outside of it, which are subject to decay, unless they are reactivated during retrieval processes. A model of this type need not rely on structural information, or even serial order, for retrieval cues. To the extent that syntactic information is utilized, it is established through encoding of morphosyntactic features like [±Theme] or [±Object] in their feature bundles, which collectively identify appropriate structures for retrieval. Thus, such models are entirely domain-general in the sense that the memory operations active during retrieval of linguistic material are the same as those that are active during other types of retrieval. If correct, this uniformity would be a powerful virtue—why postulate specialized retrieval procedures for linguistic structures when the unique aspects of language parsing could be captured simply in terms of specialized features comprising the representations that form the products of memory? In other words, what would be distinctive about language processing would be not so much the *mechanisms* involved in retrieval as how objects over which such mechanisms operate would be *encoded* in memory.

Nevertheless, the following study lends further support to the general finding that the sentence position for an antecedent is relevant during the retrieval process as linguistic dependencies are resolved, though it cannot by itself resolve whether it is best to conceive of such information as structural in nature over sequential or temporal orderings. The study capitalizes on key properties of sluicing ellipsis, as introduced in the following section.

Sluicing and the Locality Bias

Sluicing describes focus-sensitive clausal ellipsis after a *wh*-question (Ross, 1967, 1969; Chung et al., 1995; Merchant, 2001, among others), such as (5a) below. Following Merchant (2001) and others, I assume an account of sluicing which derives the overt structure through movement of the *wh*-element who_1 from its base-generated, clause-internal position to a fronted position followed by optional ellipsis <~~he is meeting~~ t_1> of the remainder of the clause. Thus, the unelided sentence (5b) is the source for the sluice (5a).

(5) a. John is meeting {someone/a friend} for dinner, but I can't tell you who_1 <~~he is meeting~~ t_1>.
b. John is meeting {someone/a friend} for dinner, but I can't tell you who_1 he is meeting t_1.

Sluicing places restrictions on the types of nouns that can serve as correlate to the remnant, though these restrictions depend on the type of *wh*-element and its restrictor residing in the remnant (Chung et al., 1995; Romero, 1998). For example, proper names and definite nouns are often unacceptable correlates for a *who*-remnant unless it is followed by *else* (6a). In select cases, the *wh*-element may co-refer with an adjunct (6b) or argument (6c) correlate that did not appear in the antecedent clause overtly, in an operation that Chung et al. (1995) call "sprouting."

(6) a. John is meeting Mary/the president for dinner, but I can't tell you who *(else).
b. John is meeting Mary/the president for dinner, but I can't tell you where/why/with who.
c. John ate, but I can't tell you $what_1$.

Sluicing, along with other forms of ellipsis, has received much attention in recent processing literature (Frazier and Clifton, 1998, 2005, 2011; Carlson et al., 2009; Martin, 2010; Poirier et al., 2010; Dickey and Bunger, 2011; Martin and McElree, 2011, among others). Previous results from processing ellipsis support the expectations of content-addressable retrieval systems, in that retrieval of antecedent material at the ellipsis site appears not to be affected by the size or complexity of the recovered material, an effect explained as either a cost-free copying mechanism (Frazier and Clifton, 2001, 2005; Frazier, 2008) or as a direct pointer in memory (Martin and McElree, 2009). In addition, Martin and McElree (2011) find that increasing the distance to a correlate in sluiced sentences affects retrieval accuracy, not retrieval speed, as predicted by content-addressable systems in which retrieval speed is held constant (as various models of retrieval propose, e.g., McElree, 2000; McElree et al., 2003; Lewis and Vasishth, 2005; Lewis et al., 2006). More detailed comparison to previous studies on sluiced sentences and cue-based parsing is delayed until the General Discussion.

Another important general result is that sluices show a structural preference to associate the remnant with the most local correlate in the antecedent clause, a principle formalized as the Locality bias below (see also Harris and Carlson, 2015, for a similar preference with *let alone* ellipsis).

(7) **Locality bias:** Associate the remnant of clausal ellipsis with a correlate occupying the structurally most local position.

Initial evidence for the Locality bias came from Frazier and Clifton (1998), who manipulated whether a sluiced sentence contained one or more possible correlates (8). In a self-paced reading study, they found that cases with multiple possible antecedents (8b) were read faster than unambiguous structures (8a). The penalty for (8a) can be attributed, in effect, to a violating the preference for Local correlates[1].

(8) a. Somebody claimed that the president fired someone but nobody knows who.
b. Somebody claimed that the president fired Fred but nobody knows who.

[1]Fully ambiguous sentences also show a strong Locality bias in silent reading, as explored in the controls for a different study (Harris, 2015). Subjects saw four ambiguous sluiced sentences like "An editor called a journalist, but I can't say which one it was any more" and were asked how they interpreted the sentences. There was an 86% bias toward the most local, object-correlate interpretation. In addition, the vast majority of the subjects ($N = 48$) displayed a complete or majority Locality bias for the items: 3 subjects were at chance, and only 1 subject had a consistent preference for non-local correlates.

Carlson et al. (2009) provide additional support for the Locality bias for sluiced sentences in an auditory questionnaire. They observed that unless the subject was focus-marked by a pitch accent or an *it*-cleft, sentences with object correlates were rated higher than alternatives.

A similar Locality bias has been observed for sluices with sprouted antecedents. Frazier and Clifton (2005) report a naturalness rating and reading time advantage when a verb with an implicit object (*studied*) appeared in second conjunct position (*slept and studied*) as compared to first conjunct position (*studied and slept*), as in *Michael (slept and studied/studied and slept), but he didn't tell me what*[2]. The basic result coheres with the expectations of the Locality bias in that near antecedents confer a processing advantage over far antecedents (see also Martin and McElree, 2011), though it does complicate the prediction that antecedent distance or complexity is not relevant to retrieval. As an aside, the Locality bias is not unique to sluiced sentences. It has been observed in other move-and-delete types of ellipsis, such replacives (Carlson, 2013) and *let alone* ellipsis (Harris and Carlson, 2015), as well as ambiguous gapping structures (Carlson et al., 2005).

The locus of the Locality bias is open to multiple possibilities. One such possibility derives from the assumption that the processor must ultimately recover elided material for interpretation, presumably by employing default biases and cues from information structure. Another, perhaps not mutually exclusive, possibility suggested by Frazier and Clifton (1998) and Carlson et al. (2005, 2009) is that the most likely correlate is determined by default-focus marking on the most embedded constituent (Selkirk, 1984; Cinque, 1993). In canonical English SVO sentences, the most deeply embedded constituent happens to be the object. For whatever reason, the preferences guiding remnant resolution in sluicing ellipsis appear to diverge from the first-mention bias established for third person pronouns, in which a pronoun is preferentially associated with the subject of a preceding clause (Arnold, 1998, among others).

In any event, there seems to be good evidence that sluiced sentences prefer the most local correlate as the antecedent for the remnant. We now turn to how the expectation for Local correlates might affect content-addressable retrieval systems, as outlined above.

The Current Study

An important advantage of using sluicing ellipsis to address the questions above is that the retrieval cues may be explicitly manipulated by modifying the inner restrictor of the *wh*-element. In example (2), for instance, the correlate-remnant pair can be disambiguated simply by repeating the nominal phrase directly, as in *A few linguists gave some silly examples, but I don't remember which linguists*. Such cases determine which noun functions as the correlate to the remnant by completely specifying the relationship: in such cases, there is *total overlap* between remnant and correlate. The eye tracking experiment below exploits this possibility by manipulating whether the restrictor is completely specified by a nominal like *which tourists/wines* (cue-rich probe) or partially specified by a pronoun like *which ones* (cue-poor probe), along with whether the indefinite (assumed to be the preferred correlate) appears in the preferred object location (9a) or not (9b). In addition, it manipulated whether a definite noun distractor appeared in the plural, thereby providing partial cue overlap with the indefinite.

(9) a. The tourists sampled some wine(s), but I don't know which wines/ones.
b. Some tourist(s) sampled the wines, but I don't know which tourists/ones.

Sluiced sentences like (9) conceivably involve two instances of retrieval: first, the recovery of the elided IP after the remnant, and, second, the pairing between the remnant and the correlate. Regarding the recovery of the elided IP, I adopt an approach in which a syntactic representation is recovered through some sort of cost-free mechanism, such as syntactic copying or recycling (Frazier and Clifton, 2001, 2005) or a pointer in memory (Martin and McElree, 2009), such that the size and complexity of the antecedent clause is essentially irrelevant for retrieval speed (Martin and McElree, 2009, 2011).

Regarding the pairing of the remnant with the correlate, there are several theoretical options to consider, especially with respect to the different types and strengths (diagnosticity) of cues in the remnant. First, we might imagine that the parser forms a dependency between the remnant and correlate selectively, that is, only when the remnant contains a pronoun, as in *ones*, but not when its inner restrictor is fully specified, as in *tourists* or *wines*. In this case, a fully specified restrictor could be interpreted via straightforward composition, without retrieving a correlate. However, this approach is unlikely given results from sprouting in sluicing ellipsis, which show a penalty when there is no overt correlate in antecedent clause (Frazier and Clifton, 1998; Dickey and Bunger, 2011).

The two remaining options would require that a dependency be formed between all types of correlates and remnants, but differ in what type of mechanism establishes it. One option to consider is one in which the nominal in the restrictor obviates the associative cue-matching procedure by forming a direct link to the previous instance of the noun, trivially avoiding cue overload effects altogether. Another option is that establishing a dependency between correlate and remnant evokes an associative cue-matching procedure, as proposed for anaphoric dependencies in general, but mitigates cue overload effects by virtue of the total overlap in cues between the remnant probe and the target correlate. In either case, we would expect the strength of the cue at the remnant to modulate the retrieval process. For concreteness, I adopt the latter approach, acknowledging that the experiments below do not depend on or arbitrate between these two possibilities.

As observed by a reviewer, it may be important that the two types of dependencies are not independent: if a comprehender resolves the remnant to an object-position correlate in (9), she is also committed to a particular syntax for the IP ellipsis, e.g., *[which wines]$_1$ they sampled t_1/[which tourists]$_1$ t_1 sampled them.*

[2]Frazier and Clifton (2005) argued that this asymmetry indicates the presence of syntactic structure within the ellipsis site, as the restriction on accessibility follows from their "conjunction domain hypothesis," an independently motivated syntactic constraint on extraction. See Martin and McElree (2011) for commentary.

While this dependency should be explored in depth, we will not do so here[3]. I will simply assume that however recovery of the IP ellipsis impacts retrieval of the remnant, the effects will be comparable across conditions.

An important conceptual issue for cue-based parsing models in general is what types of information constitute cues for retrieval. In such models, it is conceivable that any information coded as a feature value in a feature bundle is qualified to serve as a cue for retrieval, though some types of information, especially relational information, might be less amenable to representation by features than others (Kush, 2013). Following recent literature (e.g., Lewis and Vasishth, 2005), I assume that retrieval is driven, at least in part, by features from lexical (gender and number) and morphosyntactic (grammatical roles and case) information derived from context, and that what is retrieved are partial representations of constituents. Further, retrieval occurs whenever an item has to be associated with another item in memory for complete interpretation, including better-studied cases of anaphora resolution, verbal agreement, NPI licensing, and variable binding, although different kinds of dependencies might attend to distinct types of information. I remain largely agnostic about the internal organization of retrieval with respect to other interpretive processes, e.g., whether retrieval is discrete, continuous, or cascaded, as the study was not designed to address these issues, and the results below are consistent with any number of possibilities.

Assuming an associative cue-matching procedure and a preference for local antecedents, the reading experiment below was designed to test the following two basic predictions:

P1. Locality: The most local antecedent, in this case the object noun, is favored for retrieval.

P2. Nominal Advantage: Nominal restrictors (*which tourists*/*which wines*) include a rich set of cues specifying retrieval, and thus facilitate retrieval over cue-poor probes like *which ones*.

The most important prediction, however, is one in which distractors outside the local (object) domain are subject to varying degrees of interference; a strong effect of interference is predicted only in case of partial overlap, as fully specifying the inner restrictor with a nominal should eliminate the effects of cue overload, either by delivering the appropriate correlate directly, or via total overlap between the remnant and the target correlate.

P3. Structure-Dependent Interference: A retrieval penalty for violating Locality

i. arises when a distractor in the preferred (object) position shares features with the remnant, and
ii. increases if retrieval is initiated by cue-poor probes (*ones*).

Prior research investigating the effect of structural constraints on retrieval has often used the gender feature in a feature mismatch paradigm (e.g., Clifton et al., 1999; Badecker and Straub, 2002; Sturt, 2003; Chow et al., 2014). Manipulating gender agreement between the remnant and the antecedents was not possible here, given that English does not encode gender for impersonal pronouns like *ones*. Therefore, we must first show that plural definite nouns are viable correlate competitors for unambiguous *which* remnants, like *which tourists* or *which wines*, the central task of the following experiment. An affirmative finding will support the assumption that the plurality feature sufficiently induces similarity-based interference effects in the formation of correlate-remnant pairs with pronouns like *ones* in the next experiment. In addition, it will address the assumption regarding whether the indefinite determiner *some* marks the preferred correlate for sluices with *which*-remnants as opposed to the definite determiner.

As a final terminological note, the present use of "interference" diverges somewhat from a common use in the literature, in which the distractor is not a grammatical antecedent, or otherwise inaccessible (e.g., Van Dyke, 2007; Phillips et al., 2011). If both nouns in the matrix are acceptable as antecedents, the manipulation might be best cast in terms of a "fan effect" in which multiple non-referents interfere with dependency resolution (Anderson, 1974; Anderson and Reder, 1999; Autry and Levine, 2014). However, as the effects in either case would ideally be driven by the same underlying types of retrieval mechanisms, I retain the use of interference here, in hopes of expanding the empirical range of strongly biased, though not strictly speaking ungrammatical, structural preferences.

EXPERIMENT 1

A forced-choice completion test was first conducted over the Internet in order to determine the extent to which a plural definite noun competes with a plural indefinite as a correlate. I take such cases to be indicative of similarity-based interference effects, although they may differ in kind from other types of interference. A further question is whether the extent to which plurality makes a definite noun an appealing correlate is affected by its structural position.

Method

Participants

Twenty-nine subjects were recruited using Amazon Mechanical Turk, an Internet-based service where individuals complete

[3]The dependency between retrieval of a correlate and the syntax of the ellipsis site raises a number of pertinent issues regarding the relationship between the Locality Bias and parallelism between the matrix and the second conjunct. The interpretation of sluicing is known to be sensitive to parallelism (Carlson, 2002; Dickey and Bunger, 2011), as is conjunction generally (Frazier et al., 1984). If the processor takes the remnant phrase to be an object, perhaps on the basis Minimal Attachment (Frazier, 1978), the processor might prefer an object correlate to create structural parallelism between clauses. In an unpublished auditory forced-choice completion study ($N = 48$), Katy Carlson and I manipulated the surface position of the remnant to appear either as a subject (*A waiter talked to a guest, but . . . which waiter/which guest isn't clear*) or as an object (*A waiter talked to a guest, but . . . it's not clear which waiter / which guest*), along with the location of a pitch accent (subject or object). In addition to a general 66% bias for local object-position correlates, we found that subject position remnants (*which guest / which waiter isn't clear*) failed to elicit more subject responses than object position remnants did, and that pitch accent placement was the primary determinant of continuation choice. Further, accenting the object resulted in 8% more object continuations with a surface subject remnant, suggesting that linguistic focus (manifested here in the form of pitch accent), not syntactic parallelism, is the driving force behind the bias for local correlates in standard sluicing constructions.

short tasks online for payment. One subject self-identified as a non-native English speaker, and was removed from analysis. A pretest evaluated subjects' competency with three difficult to interpret questions. Three subjects were removed for answering one or more of these questions incorrectly. Four catch items were included to identify inattentive subjects, but no subject was removed on this basis. However, an additional subject was removed for counterbalancing purposes, leaving 24 subjects in the final data set. All subjects were compensated $4 for their participation, regardless of native language or performance. This experiment, along with the following, were carried out with prior Internal Review Board approval from Pomona College. All subjects gave written informed consent before starting the experiment, and were permitted to remove themselves at any time from the procedure without penalty.

Materials

The 2×2 experimental design crossed Indefinite Location (Object indefinite, Subject indefinite) with Definite Number (Plural, Singular). The levels of the Indefinite Location condition were determined by its syntactic position in the matrix clause. In other words, there were two sequences of determiner in the matrix clause: either (i) a definite subject (singular or plural) followed by a plural indefinite object, or (ii) a plural indefinite subject followed by a definite object (singular or plural). The Plural condition was created from the Singular condition simply by adding the plural marker to the definite noun phrase, e.g., *tourist* ∼ *tourists* or *wine* ∼ *wines.* All critical nouns except one (*fireman* ∼ *firemen*) were regular plurals. Twenty-four quartet fragments like (10) were constructed below. Items are reported in Appendix A of Supplementary Material.

(10) Object indefinite

a. *Plural definite:* The tourists sampled some wines, but I've forgotten...
b. *Singular definite:* The tourist sampled some wines, but I've forgotten...

Subject indefinite

c. *Plural definite:* Some tourists sampled the wines, but I've forgotten...
d. *Singular definite:* Some tourists sampled the wine, but I've forgotten...

Two forced-choice completions (11) were provided under the fragments in (10). The response options always agreed in plurality with the preceding sentence fragment, e.g., *tourists*/*wines* in (10a,c), *tourist*/*wines* in (10b), and *tourists*/*wine* in (10d). Answers were presented in a different random sequence for each subject.

(11) Forced-choice options

i. Subject correlate response: which tourist(s).
ii. Object correlate response: which wine(s).

After a short guided practice consisting of three sample sentences, subjects were presented with an additional 52 items from unrelated experiments with various structures, 12 non-experimental fillers, in which both responses were acceptable, and four catch items permitting only a single correct answer, for a total of 92 items.

Procedure

Items were presented in an individually randomized and fully counterbalanced order, so that subjects saw one and only one sentence fragment from each quartet. Subjects were instructed to rely on their intuitions to select whichever response would make "the resulting sentence sound the most natural." Subjects were given an hour to complete the task, but all subjects finished within 40 min, with an average of 25 min per subject. In addition, encrypted versions of IP addresses were recorded to identify subjects who may have taken the experiment more than once. No such cases were observed.

Results

One item (item 1 in Appendix A of Supplementary Material) contained a typo and was removed from analysis. Data analysis was conducted in R version 3.1.2 (R Development Core Team, 2014). Mean percent and standard deviations for *subject completion* responses by condition are presented in **Table 1**.

Conditions were given sum (deviation) coding so that the hypothetically simplest condition, the Object Correlate—Singular condition, was treated as the statistical baseline. The response data was modeled as a logistic linear mixed effects regression model using the lme4 package (Bates and Maechler, 2009) with by-subjects and by-items random slopes and intercepts, shown in **Table 2**.

As expected, the choice between Subject and Object correlate response closely corresponded to the location of the indefinite *some*: subject position indefinites garnered greater overall Subject correlate responses ($M = 71\%$, $SE = 3$) than object position indefinites ($M = 24\%$, $SE = 3$), $z = 6.77$, $p < 0.001$. The result confirms the intuition that language users prefer indefinites as correlates to remnants in sluiced structures, though

TABLE 1 | Experiment 1: percent *subject* response selected.

Indefinite location	Definite number		Correlate mean	Attraction effect
	Plural	Singular		
Object	28% (4)	19% (3)	24% (3)	−9%
Subject	67% (4)	75% (4)	71% (3)	8%

Standard errors are in parentheses.

TABLE 2 | Experiment 1: results of linear mixed effects regression model.

	Estimate	Std. error	Wald *Z*	*p*-estimate
(Intercept)	−0.179	0.362	−0.494	0.621
Correlate location	**1.661**	**0.285**	**5.834**	**< 0.001**
Definite number	−0.038	0.208	−0.183	0.855
Correlate × Definite number	**−0.473**	**0.199**	**−2.376**	**< 0.05**

Significant effects are printed in bold.

we should note that the preference is not absolute; see also the Discussion Section of Experiment 2, which acknowledges several complications.

While there was no main effect of Definite Number in this model, there was an interaction between Indefinite Location and Definite Number. In the case of a subject indefinite, more subject completions were observed when the object noun was singular (10d) than plural (10c); in the case of an object indefinite, more subject completions were elicited when the subject noun was plural than when it was singular (10b) than plural (10a), $z = 3.67$, $p < 0.001$. This reversal is to be expected if the indefinite provides the preferred candidate for the correlate, but a plural distractor interferes with the distinctiveness of the indefinite target.

These patterns are consistent with the theoretically-motivated assumption that *which*-remnants in sluicing prefer the antecedent with the most accessible set of individuals. In this case, the indefinite description *some* makes a set of alternatives salient in the discourse, as opposed to a definite description, which arguably introduces a plural sum of individuals that can be interpreted as a single entity (Link, 1983). Accordingly, responses were transformed to reflect the pairing in which the remnant forms a contrast with the indefinite, as depicted in **Table 3**. The transformed response data was modeled as a logistic linear mixed effects regression model using the lme4 package (Bates and Maechler, 2009) with by-subjects and by-items random slopes and intercepts and deviation coding as before. The result is shown in **Table 4**.

The model supports a sole effect of Definite Number, in that a plural definite noun (*the wines/the tourists*) resulted in fewer responses that co-referred with the indefinite target ($M = 69\%$, $SE = 3$) than singular definite distractors ($M = 78\%$, $SE = 3$), $t = -2.38$, $p < 0.05$, although the indefinite is still generally preferred.

TABLE 3 | Experiment 1: responses by the percentage of cases in which the indefinite was selected as the correlate.

Correlate choice	Plural interference		Correlate mean	Interference difference
	Interference	No interference		
Object	72% (4)	81% (3)	76% (3)	9%
Subject	67% (4)	75% (4)	71% (3)	8%
Interference mean	69% (3)	78% (3)		

Standard errors in parentheses.

TABLE 4 | Experiment 1: results of linear mixed effects regression model on the proportion of transformed responses.

	Estimate	Std. error	Wald *Z*	*p*-estimate
(Intercept)	1.661	0.285	5.835	< 0.001
Correlate location	−0.177	0.361	−0.49	0.624
Definite number	**−0.474**	**0.199**	**−2.38**	**< 0.05**
Correlate × Definite number	−0.038	0.208	−0.181	0.857

Significant effects are printed in bold.

Discussion

The results suggest that the structures are not fully ambiguous: there is a strong preference to associate the remnant with an indefinite correlate. The transformed results also show clear support for general similarity-based interference, in that plural definite distractors, which shared the plurality feature with an indefinite correlate, attracted more remnant resolutions than singular definite distractors.

EXPERIMENT 2

A second experiment was conducted to test the central predictions outlined above. First, by *Locality*, subject position indefinite nouns should elicit an online processing penalty over their more local, object position counterparts. Second, by *Nominal Advantage*, *wh*-restrictors with a fully specified nominal should facilitate the retrieval of their correlates compared to pronominal restrictors like *ones* by virtue of providing a richer feature set for cue-matching. Lastly, by *Structure-Dependent Interference*, plural definite nouns should exert a greater interference effect on retrieval when occupying object position, and, further, that such effects should manifest predominantly when the cues for retrieval are poor.

Method

Participants

Fifty-six native English speaking college students with normal or corrected-to-normal vision were recruited for the experiment, and were compensated $10 for participation. Nine students were excluded due to excessive blinks leading to extreme data loss, as detailed below, resulting in a final dataset of 47 subjects.

Materials

Twenty-four sextets were constructed from the items in Experiment 1, modified so that there were three animate subject correlate (12a–c) and three inanimate object correlate (12d–f) conditions. In both cases, there was a condition with a definite plural distractor and a fully-specified nominal in the *wh*-restrictor, e.g., *tourists* or *wines* (12a,d), a definite plural distractor and the plural pronoun *ones* (12b,e), and a definite singular distractor and the plural pronoun *ones* (12c,f). Note that in (12a–c) the indefinite noun appears in the object, and so by hypothesis the local and preferred, position. The pipe symbol "|" indicates how materials were later divided into seven regions for analysis. All conditions were identical after the remnant region

(12) a. |The tourists |sampled |some wines, |but I've forgotten |which wines, …
b. |The tourists |sampled |some wines, |but I've forgotten |which ones, …
c. |The tourist |sampled |some wines, |but I've forgotten |which ones,
d. |Some tourists |sampled |the wines, |but I've forgotten |which tourists, …
e. |Some tourists |sampled |the wines, |but I've forgotten |which ones, …
f. |Some tourists |sampled |the wine, |but I've forgotten |which ones, …
|since they all |seem the same to me.

Lexical level characteristics of length and frequency were computed for nouns in subject and object position. Subject (M = 6.83; SE = 0.27) and object nouns (M = 6.96, SE = 0.41) did not differ on number of characters, $t_{(23)} = -0.24$, $p = 0.82$. Two measures of frequency were obtained from the English Lexicon Project (Balota et al., 2007). Subject (M = 8.12; SE = 0.32) and object nouns (M = 8.86, SE = 0.42) did not differ on log HAL frequency, $t_{(23)} = -1.56$, $p = 0.13$. Further, subject (M = 2.26; SE = 0.15) and object nouns (M = 2.50, SE = 0.14) match on frequency calculated from SUBTLEX, $t_{(23)} = -1.22$, $p = 0.23$. In addition, the length of the remnant region was always included as a predictor in models of that region.

A reviewer notes that the spillover regions may not have been informative with respect to the intended interpretation. However, items were intentionally designed so that properties of the inner restrictor of the remnant provided the only disambiguating information. Further, spillover regions were consistent within an item across all conditions, and thus are unlikely to explain any effects. However, as noted above, it may be fruitful to explore the influence of the structure in unelided counterparts, as in *I don't know which ones (they sampled/were sampled*). Not all interesting contrasts could be presented in a single experiment, for fear of reducing statistical power or saturating readers with too many similar constructions. Another concern was that the example above contains the ambiguous pronoun *they* after the remnant. However, the item above is unique in that respect. As shown in the Appendix of Supplementary Material, no other item contains a pronoun of any sort.

Procedure

The experiment was presented using EyeTrack, the UMass Amherst presentation software (http://www.psych.umass.edu/eyelab/). Materials were presented in a sound isolated room on a 32-bit Dell Optiplex tower, running Windows 7, with peripheral programs and the Internet connection turned off. Text was presented as a single line in black 11pt monospaced font against a white background. The monitor was situated such that approximately three characters subtended 1° of visual angle. Eye movements were recorded on an SR Research Eyelink 1000 eye tracker, mounted on the table approximately 50 cm away from a 19" Mitsubishi Diamond Pro 900u flat-screen CRT monitor running at 170 Hz. Sampling rate was set to 1000 Hz. Drift correct was performed between each trial. Subjects were instructed to read naturally and for comprehension, and were encouraged to take breaks as often as needed.

All items were followed by comprehension questions probing the subject's interpretation (13). Questions were presented in CAPS to clearly differentiate comprehension questions from experimental materials.

(13) WHAT DID I FORGET?

i. Subject response: WHICH TOURIST(S)
ii. *Object response:* WHICH WINE(S)

Subjects selected the answer from among two possible choices on a Microsoft USB Sidewinder gamepad. Question responses were not considered in the analysis below. Experiments lasted approximately 40 min on average.

Results

Individual trials were removed if the participant blinked once or more during the first pass on the remnant region. No trials were removed if blinks occurred in another region or during re-reading of the remnant. Individual trials were also removed if excessive blinking led to significant track loss, or if track loss occurred for some other reason during the experiment (<4% of total trials).

Additionally, short (under 80 ms) and long (over 1200 ms) fixation times were removed from the data, as were trials with blinks on the remnant and track losses using the program EyeDoctor (http://www.psych.umass.edu/eyelab). Several standard eye tracking measures were used in the analysis (Rayner, 1998), computed with the DOS version of EyeDry analysis software: first *pass durations* (also known as gaze duration), the sum of all fixation durations within a region before leaving that region in any direction, *go past time*, the time spent after first entering a region to first moving past the region to the right, *percentage of regressions out* of and *percentage of regressions into* a region, second *pass time*, the time spent rereading a region once the region has been exited to the right including zero times indicating failure to re-read, and *total time*, the sum of all fixation times in a region during any point in reading (see, e.g., Staub and Rayner, 2007, for a concise review of these measures). Means and standard errors for these measures are presented in **Table 5** below.

Linear mixed effects regression models were used for all statistical analyses. Fixation and reading time measures (first pass, go past, second pass, and total times) were analyzed with linear regression models, whereas proportion data (regressions in and out of a region) were analyzed with logistic regression models. As models with maximal random effects error structures (as recommended by Barr et al., 2013) typically failed to converge, all models reported here were specified with by-subject and by-items random intercepts, but not with random slopes. Fixed effect predictor contrasts were assigned deviation coding that best cohered with the conceptual aims of the study. To assess the presence of a Locality bias, object noun correlates were coded as the baseline for the Correlate position predictor. To evaluate the effect of Interference, cue-poor probes (*which ones*) without plural interference were treated as the baseline for the Interference predictor, so that the model tests for similarity-based interference effects for nominal and pronominal cues over a no interference condition with a pronominal cue.

Instead of reporting the statistical results of each measure individually, the results are discussed in terms of the predictions of interest, noting when other effects were present. All significant effects for the measures of interest are reported.

Locality

Evidence for the Locality bias was observed in multiple measures of the eye movement record. The earliest evidence was found in first pass times immediately at the remnant (Region 5). indefinite subject correlates (M = 274 ms, SE = 5) elicited longer first pass times than indefinite object correlates (M = 249 ms, SE = 5), t = 3.88. No other effects were observed in first pass times. Additional evidence for the Locality bias appears in later eye movement measures, as well. Indefinite subject correlates elicited

TABLE 5 | Experiment 2: means and standard deviations for all eye tracking measures.

		Subject	Verb	Object	But I don't know	Which X	Spill over	Final region
					FIRST PASS			
Interference nominal	Object	263 (15)	319 (13)	384 (16)	368 (14)	281 (9)	325 (13)	391 (18)
	Subject	292 (16)	320 (18)	327 (14)	375 (16)	300 (10)	335 (14)	405 (23)
Interference pronoun	Object	282 (16)	317 (14)	372 (15)	373 (15)	230 (8)	336 (14)	394 (18)
	Subject	298 (16)	292 (14)	318 (15)	352 (14)	257 (7)	338 (15)	393 (20)
No interference pronoun	Object	282 (21)	336 (17)	366 (14)	368 (15)	234 (8)	333 (15)	408 (20)
	Subject	295 (16)	321 (15)	298 (15)	370 (15)	264 (9)	339 (14)	418 (22)
					GO PAST			
Interference nominal	Object	263 (15)	414 (18)	502 (23)	385 (15)	338 (15)	387 (29)	818 (66)
	Subject	326 (29)	380 (21)	438 (25)	389 (17)	363 (15)	350 (19)	834 (54)
Interference pronoun	Object	282 (16)	381 (18)	527 (29)	392 (17)	267 (14)	352 (17)	854 (66)
	Subject	298 (16)	402 (29)	472 (30)	374 (17)	303 (18)	352 (16)	1071 (94)
No interference pronoun	Object	282 (21)	464 (25)	559 (34)	405 (21)	279 (16)	377 (24)	908 (64)
	Subject	295 (16)	378 (20)	448 (29)	394 (18)	317 (20)	400 (25)	975 (70)
					SECOND PASS			
Interference nominal	Object	90 (12)	101 (16)	39 (9)	99 (18)	55 (11)	129 (17)	NA
	Subject	84 (12)	104 (17)	29 (7)	79 (13)	42 (11)	140 (16)	NA
Interference pronoun	Object	79 (12)	116 (15)	64 (13)	56 (12)	28 (7)	122 (18)	NA
	Subject	149 (21)	173 (21)	84 (16)	88 (16)	61 (16)	151 (19)	NA
No interference pronoun	Object	128 (14)	154 (23)	73 (14)	68 (12)	37 (8)	123 (15)	NA
	Subject	110 (16)	137 (17)	87 (15)	87 (16)	45 (10)	148 (16)	NA
					TOTAL TIMES			
Interference nominal	Object	249 (17)	439 (23)	446 (19)	458 (23)	344 (16)	431 (20)	473 (23)
	Subject	323 (27)	408 (25)	365 (17)	449 (20)	353 (16)	436 (22)	504 (26)
Interference pronoun	Object	270 (18)	429 (23)	472 (23)	412 (20)	253 (12)	438 (23)	496 (25)
	Subject	361 (24)	473 (29)	425 (25)	439 (21)	311 (17)	466 (22)	522 (29)
No interference pronoun	Object	277 (20)	521 (32)	496 (27)	441 (20)	263 (13)	424 (20)	505 (24)
	Subject	328 (23)	449 (23)	411 (24)	451 (23)	307 (15)	465 (20)	546 (25)
					REGRESSIONS OUT			
Interference nominal	Object	NA	18 (3)	24 (3)	2 (1)	15 (3)	5 (2)	42 (4)
	Subject	NA	12 (3)	20 (3)	2 (1)	12 (3)	1 (1)	44 (4)
Interference pronoun	Object	NA	13 (3)	26 (3)	2 (1)	7 (2)	2 (1)	45 (4)
	Subject	NA	17 (3)	26 (3)	2 (1)	7 (2)	2 (1)	49 (4)
No interference pronoun	Object	NA	23 (3)	24 (3)	3 (1)	10 (2)	5 (2)	47 (4)
	Subject	NA	14 (3)	25 (3)	3 (1)	11 (2)	5 (2)	52 (4)
					REGRESSIONS IN			
Interference nominal	Object	68 (6)	32 (4)	5 (2)	22 (3)	8 (2)	35 (4)	NA
	Subject	54 (6)	29 (4)	4 (2)	16 (3)	4 (1)	37 (4)	NA
Interference pronoun	Object	62 (6)	35 (4)	7 (2)	12 (3)	3 (1)	36 (4)	NA
	Subject	66 (5)	38 (4)	11 (2)	13 (3)	4 (2)	36 (4)	NA
No interference pronoun	Object	78 (4)	34 (4)	10 (2)	16 (3)	4 (2)	36 (4)	NA
	Subject	60 (5)	34 (4)	13 (3)	18 (3)	5 (2)	42 (4)	NA

longer go past times than object correlates on the remnant region ($M_{Subject}$ = 328 ms, SE = 10; M_{Object} = 295 ms, SE = 9), t = 2.72, and on the sentence final region, ($M_{Subject}$ = 960 ms, SE = 43; M_{Object} = 860 ms, SE = 38), t = 2.12. Fixed effects of models substantiating the above effects are provided in **Table 6**.

Further, violating Locality manifested in a persistent penalty for total times, as indefinite subject correlates were significantly longer in the sentential subject region ($M_{Subject}$ = 337 ms, SE = 40; M_{Object} = 265, SE = 11), t = 4.65, the remnant ($M_{Subject}$ = 324, SE = 9; M_{Object} = 287, SE = 8), t = 3.40, and the final region ($M_{Subject}$ = 524 ms, SE = 15; M_{Object} = 491 ms, SE = 14), t = 2.20. Models computed for total times are provided in **Table 7**. In addition, there was a marginally significant penalty for indefinite subject correlates (M = 147 ms, SE = 10) compared to indefinite object correlates (M = 125, SE = 10) on the spill-over region in second pass re-reading times, t = 1.95. A summary of the main effects on the remnant is provided in **Figure 1**.

However, a few measures showed a cost for indefinite object correlates. There were more regressions into the sentential subject region for items with indefinite object correlates (M = 70%, SE = 3) than indefinite subject correlates (M = 60%, SE = 3), t = 2.64, $p < 0.01$. The increased rate of regressions into the subject region may correspond to increased global re-reading for indefinite object correlates, as opposed to regressing back into a specific region. In addition, indefinite object correlates elicited longer total times than indefinite subject correlates did in the region containing the object noun (M_{Object} = 471 ms, SE = 13; $M_{Subject}$ = 401 ms, SE = 13), t = −4.73. This effect could be explained if total times corresponded to additional re-reading of the correlate. However, it is unclear whether such an explanation can be strongly maintained without supporting evidence from regressions in and second-pass reading measures, of which there is little evidence.

Nominal Advantage

For the second prediction, we expect that nominal restrictors (*which wines/tourists*) should receive a processing benefit over pronominal restrictors (*which ones*), due to greater specificity of the retrieval cue (also known as "cue diagnosticity"; see Martin and McElree, 2009, 2011; Van Dyke, 2011). Indeed, we find the expected advantage for nominal probes in a variety of measures. In go past times, there was a 117 ms advantage for nominal restrictors over pronominal ones on the final region, t = −2.42; see the *Interference nominal* row in **Table 6**. There were fewer regressions into the object region when a nominal cue in the remnant followed (M = 5%, SE = 1) as compared to a pronominal cue in the restrictor (M = 11%, SE = 2), t = −2.93, $p < 0.01$.

Further, the advantage for nominal cue conditions was considerable in second pass re-reading times of every region of the matrix clause. Nominal restrictors elicited shorter second pass times in the subject region ($M_{Nominal}$ = 87, SE = 9; $M_{Pronominal}$ = 119, SE = 11), t = −2.61, the verb region ($M_{Nominal}$ = 103, SE = 12; $M_{Pronominal}$ = 145, SE = 14), t = −3.21, and the object region ($M_{Nominal}$ = 34, SE = 6; $M_{Pronominal}$ = 80, SE = 10), t = −4.31; see **Figure 2**. There

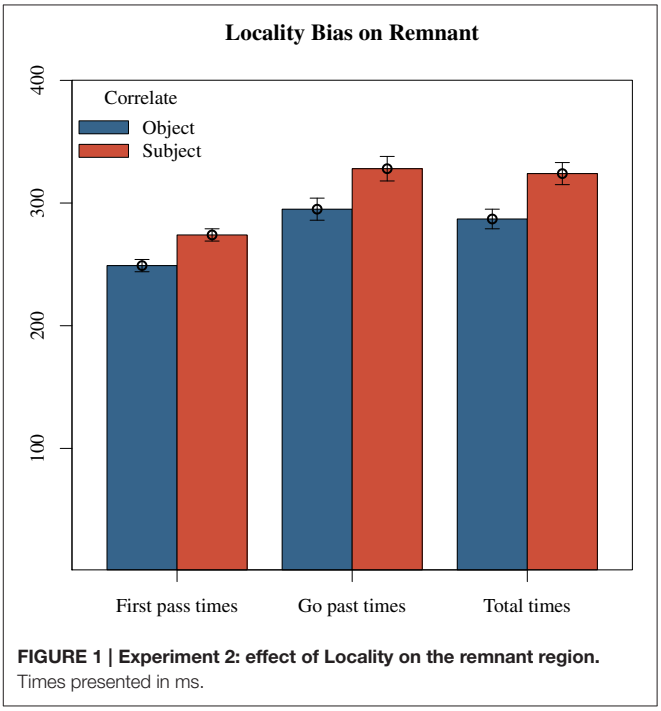


FIGURE 1 | Experiment 2: effect of Locality on the remnant region. Times presented in ms.

TABLE 6 | Experiment 2: linear mixed effects regression models for first fixation and go past times on Remnant and go past times on the sentence final region.

	First pass times on remnant			Go past times on remnant			Go past times on final region		
	Estimate	**Std. error**	***t*-value**	**Estimate**	**Std. error**	***t*-value**	**Estimate**	**Std. error**	***t*-value**
(Intercept)	231.863	18.748	12.368	266.994	35.923	7.432	899.026	80.558	11.16
Interference nominal	−7.037	5.745	−1.225	22.199	15.588	1.424	**−83.699**	**34.589**	**−2.42**
Interference pronoun	**18.517**	**8.139**	**2.275**	−18.053	10.95	−1.649	54.613	41.688	1.31
Locality	**12.827**	**3.303**	**3.883**	**17.02**	**6.254**	**2.721**	**53.623**	**25.251**	**2.124**
Length	5.502	3.453	1.593	8.485	6.644	1.277	NA	NA	NA
Interference nominal × Locality	1.357	4.674	0.29	−2.685	8.812	−0.305	−48.508	40.662	−1.193
Interference pronoun × Locality	−2.378	4.653	−0.511	1.398	8.865	0.158	58.372	45.79	1.275

Effects with t-values above |2| are shown in bold.

TABLE 7 | Experiment 2: linear mixed effects regression models for total times on the sentence initial, remnant, and sentence final regions.

	Total times on subject region			Total times on remnant			Total times final region		
	Estimate	**Std. error**	***t*-value**	**Estimate**	**Std. error**	***t*-value**	**Estimate**	**Std. error**	***t*-value**
(Intercept)	301.190	24.644	12.222	257.591	32.441	7.940	503.846	36.847	13.674
Interference nominal	−15.269	11.066	−1.380	23.878	14.036	1.701	−20.769	11.479	−1.809
Interference pronominal	13.851	11.069	1.251	−12.632	9.842	−1.284	5.633	11.483	0.491
Locality	**36.310**	**7.816**	**4.646**	**19.117**	**5.620**	**3.402**	**17.834**	**8.107**	**2.200**
Length	NA	NA	NA	9.412	5.996	1.570	NA	NA	NA
Interference nominal × Locality	−2.205	11.079	−0.199	−14.451	7.968	−1.814	0.348	11.494	0.030
Interference pronominal × Locality	12.028	11.071	1.086	10.805	7.958	1.358	−3.258	11.485	−0.284

Effects with t-values above |2| are shown in bold.

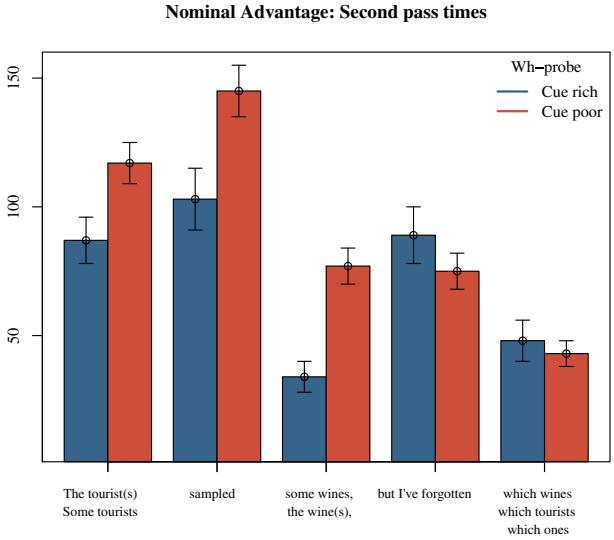


FIGURE 2 | Experiment 2: effect of Locality on the remnant region. Nominal advantage in second pass re-reading times. Times presented in ms.

were also shorter total times for nominal probes in total times for the subject region ($M_{Nominal}$ = 423 ms, SE = 17; $M_{Pronominal}$ = 485 ms, SE = 20), $t = -2.60$.

One region witnessed effects other than were expected: there were more regressions into Region 4 for Nominal than Pronominal conditions ($M_{Nominal}$ = 19%, SE = 2; $M_{Pronominal}$ = 15%, SE = 1), $t = -2.28$, $p < 0.05$. There is no ready explanation of this small effect, as the content of the region was identical in all conditions.

Structure-Dependent Interference

The central prediction of an interaction between similarity-based interference and structural position of the correlate was supported by later eye movement measures. Importantly, a penalty was predicted only for pronominal probes in the remnant like *ones*, but not when the cue was fully specified, as in the case of nominal probes. As expected, there was a greater penalty for second pass times for definite plural object nouns (*the wines*) and a subject correlate (*some tourists*) when the *wh*-restrictor was a pronominal. In the subject region, the penalty for conditions with pronominal restrictors was significantly greater for indefinite subject correlates (d = 70 ms) than for indefinite object correlates ($d = -18$ ms), t = 3.54. Similar effects obtained in the verb region, with a 57 ms penalty for subject correlates over object correlates ($d = -17$ ms), t = 2.26. Both of these effects are shown in the final row in **Table 8**.

On the remnant region, there was again a greater penalty for indefinite subject correlates (d = 33) than for indefinite object correlates (d = 8), t = 2.11; see **Table 9** and **Figure 3**. What's more, nominal restrictors showed a small 13 ms second pass time advantage for indefinite subject correlates and plural distractor objects, $t = -2.08$ in second pass times; **Table 9**. Finally, in the verb region, there was a greater total times penalty for indefinite subject correlates and plural distractors with pronominal retrieval cues (d = 72 ms) than the no interference baseline (d = 44 ms), t = 2.62.

Discussion

Results from the reading experiment support all three predictions of interest. There was early and sustained support for violating the Locality bias, and an advantage for Nominal cues that manifested in go past and second pass measures. These two effects interacted with respect to interference and cue-specificity: there were greater interference effects when the distractor occupied the preferred object position, such that the effect was enhanced when cues for retrieval at the remnant were partial. The results are thus compatible with previous findings of retroactive interference (e.g., Van Dyke and McElree, 2006; Van Dyke and Johns, 2012), but also adds support to the growing body of evidence that retrieval is modulated by the position of the antecedents (e.g., Van Dyke and McElree, 2011; Chow et al., 2014; Dillon et al., 2014; Kush and Phillips, 2014, among others). While the results are clearly consistent with the central predictions of a cue-based parsing system in which the location of targets in the sentence is not ignored during retrieval, the issue of how precisely to utilize such information within cue-based parsing models is far from settled. I return to this question the General Discussion.

As noted by a reviewer, evidence for the central predictions manifested at somewhat different time courses, although we should exercise caution when assigning linking assumptions to eye movement measures (Clifton et al., 2007). Whereas,

TABLE 8 | Experiment 2: nominal advantage in second pass re-reading times.

	Second pass on subject region			Second pass on verb region			Second pass on object region		
	Estimate	**Std. error**	***t*-value**	**Estimate**	**Std. error**	***t*-value**	**Estimate**	**Std. error**	***t*-value**
(Intercept)	107.007	13.958	7.666	131.059	21.721	6.034	63.048	11.175	5.642
Interference nominal	**-20.473**	**7.849**	**-2.608**	**-29.972**	**9.345**	**-3.207**	**-29.309**	**6.794**	**-4.314**
Interference pronominal	7.929	7.849	1.01	15.963	9.348	1.708	11.965	6.794	1.761
Locality	8.137	5.543	1.468	7.064	6.6	1.07	4.291	4.797	0.894
Interference nominal × Locality	−12.287	7.856	−1.564	−7.381	9.357	−0.789	−10.67	6.8	−1.569
Interference pronominal × Locality	**27.778**	**7.851**	**3.538**	**21.11**	**9.349**	**2.258**	7.044	6.795	1.037

Effects with t-values above |2| are shown in bold.

evidence for Locality and Nominal Advantage appeared in various measures, support for Structure-Dependent Interference was only observed in the "late" measure of second-pass times, as subjects re-read portions of the sentence. This delayed effect is compatible with multiple interpretations, including a multiple stage model of anaphoric processing in which measures occurring later in the eye movement record could reflect processing at a secondary stage of discourse integration, perhaps along the lines of Garrod and Sanford's (1990, 1994) bonding and resolution model. If this were the case, Structure-Dependent Interference might reflect difficulty interpreting the link between a poorly specified remnant and a correlate, rather than retrieval difficulty. Alternatively, that the effect appears relatively late in the eye movement record could be attributed to a lag resulting from poor quality matches. In this case, the integration difficulty would directly reflect increased interference from distractors in structurally preferred positions. The results do not arbitrate between these, or any number of, additional possibilities, which must instead be resolved through careful experimental design.

A reviewer proposed that several of the sentences in (12) are ambiguous, in that *which ones* may also co-refer with definite plural nouns. The above design depended on the assumption that definite nouns fail to provide an appropriate antecedent, as discussed in connection with examples (5–6). Yet, there may be a few systematic exceptions to the generalization that the remnant cannot correspond to a definite correlate. Discussion in the literature centers around contrasts like (13) below, which shows that a d-linked *which* remnant can take a definite noun as a correlate, but a simple *wh*-phrase like *what* cannot (Chung et al., 1995; Dayal and Schwarzschild, 2010).

(13) a. John announced he had eaten the asparagus. We didn't know which asparagus.
b. * John announced he had eaten the asparagus. We didn't know what.

These cases have been given various analyses. Chung et al. (1995) suggest that definite nouns are available as correlates whenever they are compatible with the pragmatic contribution of the remnant. They attribute the contrast in (13) to an intuitive conflict in familiarity between the definite *the asparagus* and the novelty imposed by *what* in (13b), a conflict that (13a) avoids.

In contrast, Dayal and Schwarzschild (2010) identify several cases in which presumed speaker knowledge, rather than familiarity, is the distinguishing factor. In their account, (14a) is infelicitous because the speaker has contradicted herself: the knowledge state that permits the speaker to assert that John talked to the detective places the speaker in a sufficient epistemic position to answer the question embedded under the sluice, i.e., *which detective did he talk to*? The corresponding assertion with an indefinite (14b) does not place the speaker in such a specific knowledge state as to warrant a self-contradiction (though see Barker, 2013; Barros, 2013, for recent commentary).

TABLE 9 | Experiment 2: Structure-Dependent Interference effects in second pass re-reading times on the remnant region.

	Second pass times on remnant region		
	Estimate	**Std. Error**	***t*-value**
(Intercept)	31.428	23.059	1.363
Interference nominal	−1.576	10.212	−0.154
Interference pronominal	2.453	7.186	0.341
Locality	4.722	4.121	1.146
Length	2.58	4.347	0.594
Interference nominal × Locality	**−12.126**	**5.843**	**−2.075**
Interference pronominal × Locality	**12.313**	**5.836**	**2.110**

Effects with t-values above |2| are shown in bold.

(14) a. * John talked to the detective. I don't know which detective (he talked to).
b. John talked to a detective. I don't know which detective (he talked to).

They also observe that a definite correlate is sometimes available when it does not carry a uniqueness presupposition, as illustrated by the examples in (15) where there is no requirement that there is a singular, identifiable train or particular hospital in the context (akin to so-called weak definites, e.g., Carlson and Sussman, 2005; Aguilar Guevara, 2014).

(15) a. John is going to take the train, but he doesn't know yet which train (he is going to take).
b. They took him to the hospital. She wouldn't tell us which hospital (they took him to).

As would be predicted, a definite that corresponds to a unique individual within a given context, as in *the Chief of Police* in (16), cannot serve a correlate to the remnant.

(16) *Ed reported the matter to the Chief of Police but Joe couldn't figure out which chief of police (he reported the matter to).

Although it is unclear whether these examples are as acceptable when fully elided, e.g., *John is going to take the train, but he doesn't know which*, the experimental items were reviewed to determine whether the definite noun could sensibly be interpreted as a correlate to the remnant[4]. Two possible types of cases were observed. The first case involved collective nouns whose members could perform an action on behalf of others in the group (the ? mark reflects my judgment that these sentences are somewhat degraded):

(17) a. ? The trustees donated 10 million to the university, but I don't know which ones.
b. ? The professors wrote a letter to the dean, but it doesn't matter which ones.

For example, (17a) could be interpreted so that some of the trustees but not others were responsible for the donation. Here, the definite *the trustees* is interpreted as a collective entity, in which the donation was performed on behalf of the entire group. Although not many items in the experiment permit such a reading, one contender is *The nurses threatened to strike over some contracts, but I'm not sure which ones.*

The second case was one in which the remnant *which ones* does not refer directly to the definite noun phrase. Instead, co-reference appears to be coerced through a partitive interpretation taking the plural definite as the maximal set in order to derive a salient subset (a refset) from it (see Moxey and Sanford, 1993, for terminology). For example, such a reading might paraphrase (12d) as *Some tourists sampled the wines, but I don't know which (ones) of the wines they sampled.* The coercion process could posit a silent or elided partitive phrase, as proposed for bare determiner phrases like *Many (of them) sat down* (Gagnon, 2013). Again, very few plausible cases were found in the experimental items. Two possible cases include the example used as illustration throughout the paper (12), and *Some workers loaded the trucks, but I'm not certain which ones.* Such cases are perhaps strengthened by a distributive semantics of the verb, e.g., a sampling wine involves trying some, but not all, of it.

To assess empirically whether ambiguity could explain the effects observed above, I conducted a *post-hoc* by-items analysis of the results from Experiment 1. Averaging across conditions, no item was biased toward the definite noun completion. However, splitting the data by position of the indefinite revealed that eight items were either biased toward (definite subject: 13, 18, 21; definite object: 12, 24) or on par with (object definite: 2, 10, 18) the indefinite as a correlate. For most measures, there were no differences in the overall statistical effects when these items were removed[5]. However, removing potentially ambiguous items did weaken the interaction between Locality and Interference in second pass times: although the penalty for non-local correlates was still significantly greater for remnants with pronominal restrictors in the sentence-initial region, the interaction did not persist in following region, even though the interaction was still apparent in other measures, including total times. Thus, even though a plausible definite distractor could have engendered a longer lasting interference penalty from the indefinite, it is unlikely to be the primarily source of the effects reported here.

As a final note, ambiguity only becomes a genuine confound if it could otherwise explain the effects attributed to another variable. The only sentences that could have been ambiguous are those with cue-poor (*which ones*) remnants and plural definite distractors, i.e., (12b, d). Several other studies of pronominal ambiguity suggest that competing interpretations do not always result in processing penalties (e.g., attachment ambiguity explored in van Gompel et al., 2001, 2005), especially cases involving pronouns (e.g., Greene et al., 1992). Indeed, in the present case of sluicing, Frazier and Clifton (1998) report that ambiguity between subject and object position correlates did not slow readers down, provided that there was an indefinite correlate in the preferred, object position, like *someone* in (8a). Therefore, it is not yet clear how ambiguity would explain the effects I hope to attribute to interference.

Additionally, although the possibility of ambiguity might challenge whether we can truly interpret the effect of a plural definite as interference *per se*, it cannot fully account for the interaction between the plurality of the definite and its structural location. That is, irrespective of whether or not a definite noun is a possible correlate to *which* remnants, ambiguity does not explain why a plural definite in object position would elicit greater reading penalties than in subject position. Nevertheless, potential ambiguities could be more tightly controlled or even exploited (as in Harris, 2013, 2015) in future studies.

GENERAL DISCUSSION

The central question explored in the experiment above was whether positional information modulates similarity-based interference effects in sluicing structures. There was clear evidence that it does. The central manipulation capitalized on the unique syntactic properties of sluices in two ways. First, the Locality bias was employed to impose a preference for structural position of the correlate to a remnant in the elided clause. Second, the lexical content of the inner restrictor of the remnant was manipulated to examine the role of cue-strength in retrieval.

As previously mentioned, this study is not the first to exploit sluiced sentences in an argument in favor of content-addressable retrieval systems. Martin and McElree (2011) utilized two main properties of an object position correlate in sluiced sentences like (14) in SAT and eye tracking. The correlate appeared on its own

[4]Thanks to Colin Phillips for discussion of this issue and for providing some of the examples that appear in this section.

[5]Other changes were that the slowdown on the Interference pronoun condition in first pass times was marginal, and the penalty for violating the Locality Bias disappeared for go past and total times in the final region, as well as for second pass times in the spill over region. Several other effects were significant once potentially ambiguous items were removed, including an interaction supporting Structure-Dependent Interference in the verb region on second pass times, a slowdown for nominal restrictors in go past times at the remnant, and a previously marginal advantage for nominal restrictors reached significance for total times in the sentence-final region.

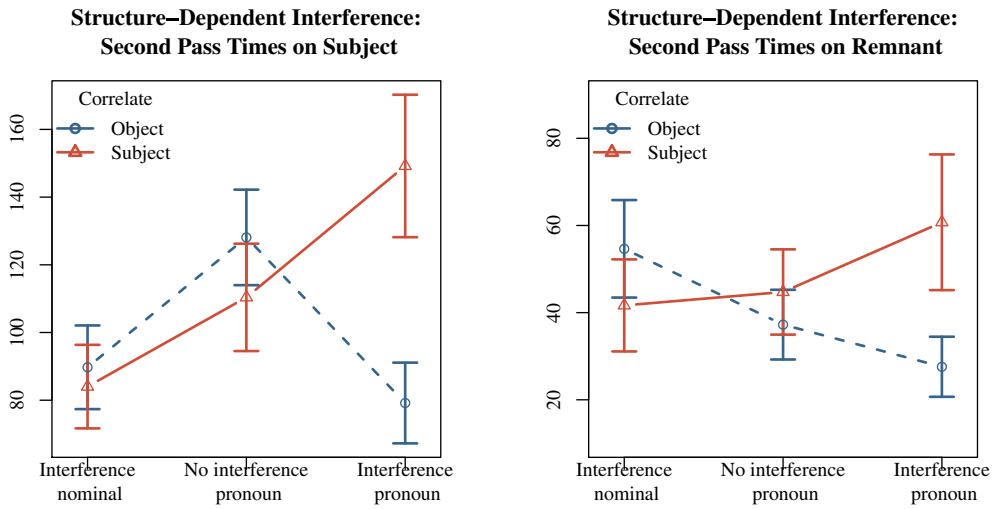


FIGURE 3 | Experiment 2: Structure-Dependent Interference effects on subject and remnant regions. Times presented in ms.

or within a conjunct in object position, and when the correlate was contained within a conjunct, which position of the conjunct it occupied (first or second conjunct position). The verbal syntax of only one member of the conjunct, *typed (something)*, provides a correlate, which varied according to whether the object was overt or not, to associate with the remnant *what*.

(14) a. Michael (slept and) studied (something), but he didn't tell me what$_1$ <~~he typed~~ t$_1$>.
b. Michael studied (something) (and slept), but didn't tell me what$_1$ <~~he typed~~ t$_1$>.

The design varied the distance between the correlate and remnant, along with the size of the elided material that was to be recovered. In keeping with the findings above, they found that readers spent longer re-reading distant antecedents (14b) than local ones (14a), and suggested, as I have, that interfering antecedents degrades the quality of a match with potential antecedents in memory. However, the materials of their study are quite different from the ones above in three respects. First, as only one conjunct provided a proper correlate (which sometimes had to be sprouted) to the remnant, the experiment lacks the conditions for fully investigating similarity-based interference from other noun phrase distractors. Second, the correlate was always in the object position, thereby satisfying the Locality bias, at least in a broad sense. Third, while the remnant varied according to *wh*-element type (*what*, *which*, and *where*), they did not manipulate the properties of the inner restrictor of the remnant to provide explicit cues to guide the dependency formation. The study above therefore contributes very different, yet congruent, evidence in favor of interference effects in retrieving correlates for sluiced sentences.

It is worth comparing Martin and McElree's study to the present one for another reason, as well. They found that the presence of a conjunct over a single noun in object position did not affect retrieval in either reading time or a SAT task, and concluded that retrieval processes access the material for ellipsis directly on the basis of its content via a cost-free pointer mechanism, in line with studies on verb phrase ellipsis (Frazier and Clifton, 2001, 2005; Martin and McElree, 2008, 2009). However, it is possible that the mechanisms responsible for retrieving a correlate for the remnant are distinct from those responsible for recovering the elided IP material. Given the previously discussed dependency between resolving the remnant and determining the appropriate syntax of the ellipsis, it stands to reason that the former might be prioritized over the later, rather than attempting to solve two retrieval problems at once. Although it is theoretically possible that the ellipsis site lacks an explicit syntactic representation (Chung et al., 1995), there is good evidence for syntactic structure in sluicing ellipsis from both theoretical (e.g., Merchant, 2001; van Craenenbroeck, 2010) and experimental (e.g., Frazier and Clifton, 2001, 2005; Poirier et al., 2010) literature, in which case retrieving the ellipsis site is unlikely to reduce to simply pairing a correlate to the remnant of ellipsis.

Finally, one might be concerned that increased temporal distance, and thus decay, between the subject and the remnant might sufficiently explain the Locality bias, thereby eliminating structural information *per se* as a factor in the retrieval process. However, this explanation is unlikely given the results of Poirier et al.'s (2010) cross-modal priming study, in which printed targets related to the subject (*the handyman*) and dative object (*the programmer*) distractors were presented at two probe points in an auditory sentence: immediately after the offset of the remnant $*_1$ or 500 ms downstream $*_2$.

(15) The handyman threw a book to the programmer but I don't know which book $*_1$ and no one $*_2$ else seems to know.

There was no difference between decision times for targets related to subject and dative object nouns until position $*_2$ (which showed a priming effect for the object), suggesting that subject

and the dative object nouns were equally active at the remnant of the ellipsis. Crucially, these effects do not contradict the results of the reading study, since probes related to the indefinite target *a book* could not be tested, given that they were repeated in the inner restrictor of the *wh*-phrase *which book*. If the restrictor were replaced with a cue-poor probe like *which ones*, we would expect an advantage for more local antecedents at, or soon after, the remnant.

Several models of sentence processing could in principle accommodate the findings reported above, models which diverge on how to account for the differences observed between subject and object position correlates. Naturally, the results of a single study cannot determine whether the effect of position reflects temporal precedence, linear distance, or, as I have suggested, structural information. Although various interpretations are possible, structural information has been shown independently to impact the earliest stages of retrieval in several related domains. It stands to reason that retrieval might privilege items located in preferred structural positions, even when the preference is not grammatically controlled. Of course, the nature of the mechanisms that underlie this putative advantage will remain unsettled until an effect of structural privilege is replicated in a design that dissociates structure from other factors, like linear order. Fortunately, sluicing ellipsis offers just the right sort of flexibility to tease such issues apart in the future.

Moreover, uncovering how the processor resolves the multiple dependencies required for interpreting sluiced sentences has only just begun. The configurational possibilities of sluicing ellipsis provide a rich testing ground for disentangling the retrieval processes that are charged with recovering linguistic antecedents and integrating them into a representation as it unfolds during real-time comprehension. While numerous questions remain, one major challenge is the stage at which semantic and discourse information informs dependency formation in ellipsis, and whether information structural cues or strongly biased contexts can favor potential antecedents in the same way that structural information can. At the minimum, the present study provides additional support for converging evidence for cue-based parsing, and that the mechanisms underlying such retrieval are not wholly blind to the structural location of products in memory.

FUNDING

The author gratefully acknowledges financial support from Pomona College.

ACKNOWLEDGMENTS

This research benefited from conversations with Lyn Frazier, Carson Schütze, Sarah VanWagenen, and the audiences at the 28th Annual CUNY Conference on Human Sentence Processing held at the University of Southern California, as well as at a psycholinguistics lab meeting at the University of Maryland. Many thanks to Brian Dillon, Colin Phillips, and the two reviewers for their generous comments on previous drafts. Any mistakes or errors should be attributed to me alone. I thank Karin Denton for her assistance in running subjects for Experiment 2.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpsyg.2015.01839

REFERENCES

Aguilar Guevara, A. (2014). *Weak Definites. Semantics, Lexicon and Pragmatics*. Ph.D. thesis, Utrecht University, The Netherlands.

Alcocer, P., and Phillips, C. (2012). *Using Relational Syntactic Constraints in Content-addressable Memory Architectures for Sentence Parsing*. College Park, MD: Ms. University of Maryland.

Anderson, J. R. (1974). Retrieval of propositional information from long-term memory. *Cogn. Psychol.* 6, 451–474. doi: 10.1016/0010-0285(74)90021-8

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., and Qin, Y. (2004). An integrated theory of the mind. *Psychol. Rev.* 111, 1036–1060. doi: 10.1037/0033-295X.111.4.1036

Anderson, J. R., and Reder, L. M. (1999). The fan effect: new results and new theories. *J. Exp. Psychol. Gen.* 128, 186–197. doi: 10.1037/0096-3445.128.2.186

Anderson, M. C., and Neely, J. H. (1996). "Interference and inhibition in memory retrieval," in *Handbook of Perception and Memory, Vol. 10, Memory,* eds E. L. Bjork and R. A. Bjork (San Diego, CA: Academic Press), 237–313.

Arnold, J. E. (1998). *Reference form and Discourse Patterns*. Ph.D. thesis, Stanford University.

Autry, K. S., and Levine, W. H. (2014). A fan effect in anaphor processing: effects of multiple distractors. *Front. Psychol.* 5:818. doi: 10.3389/fpsyg.2014.00818

Badecker, W., and Straub, K. (2002). The processing role of structural constraints on interpretation of pronouns and anaphors. *J. Exp. Psychol. Learn. Mem. Cogn.* 28, 748–769. doi: 10.1037/0278-7393.28.4.748

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., et al. (2007). The English lexicon project. *Behav. Res. Methods* 39, 445–459. doi: 10.3758/BF03193014

Barker, C. (2013). Scopability and sluicing. *Linguist. Philos.* 36, 187–223. doi: 10.1007/s10988-013-9137-1

Barr, D., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure in mixed-effects models: keep it maximal. *J. Mem. Lang.* 68, 255–278. doi: 10.1016/j.jml.2012.11.001

Barros, M. V. (2013). "Harmonic sluicing: Which remnant/correlate pairs work and why," in *Proceedings of the 23rd Semantics and Linguistic Theory,* ed T. Snider (Santa Cruz, CA: University of California), 295–315.

Bates, D., and Maechler, M. (2009). *lme4: Linear Mixed-Effects Models using S4 Classes. R Package Version 0.999375–999331*. Available online at: https://github.com/lme4/lme4/

Caplan, D., and Waters, G. (2013). Memory mechanisms supporting syntactic comprehension. *Psychon. Bull. Rev.* 20, 243–268. doi: 10.3758/s13423-012-0369-9

Carlson, G., and Sussman, R. (2005). "Seemingly indefinite definites," in *Linguistic Evidence: Empirical, Theoretical, and Computational Perspectives*, eds S. Kepser and M. Reis (Tübingen: De Gruyter Mouton), 71–85.

Carlson, K. (2002). *Parallelism and Prosody in the Processing of Ellipsis Sentences*. Outstanding Dissertations in Linguistics Series. Routledge, New York.

Carlson, K. (2013). The role of only in contrasts in and out of context. *Discourse Process.* 50, 249–275. doi: 10.1080/0163853X.2013.778167

Carlson, K., Dickey, M. W., Frazier, L., and Clifton, C. Jr. (2009). Information structure expectations in sentence comprehension. *Q. J. Exp. Psychol.* 62, 114–139. doi: 10.1080/17470210701880171

Carlson, K., Dickey, M. W., and Kennedy, C. (2005). Structural economy in the processing and representation of gapping sentences. *Syntax* 8, 208–228. doi: 10.1111/j.1467-9612.2005.00079.x

Chafe, W. K. (1976). "Givenness, contrastiveness, definiteness, subjects, topics, and point of view," in *Subject and Topic,* ed C. N. Li (New York, NY: Academic Press), 25–55.

Chen, Z., Jäger, L., and Vasishth, S. (2012). "How structure-sensitive is the parser? Evidence from Mandarin Chinese," in *Empirical Approaches to Linguistics Theory. Studies in Meaning and Structure*, eds B. Stolterfoht and S. Featherston (Berlin: De Gruyter Mouton), 43–62.

Chow, W. Y., Lewis, S., and Phillips, C. (2014). Immediate sensitivity to structural constraints in pronoun resolution. *Front. Psychol.* 5:630. doi: 10.3389/fpsyg.2014.00630

Chung, S., Ladusaw, W. A., and McCloskey, J. (1995). Sluicing and logical form. *Nat. Lang. Semantics* 3, 239–282. doi: 10.1007/BF01248819

Cinque, G. (1993). A null theory of phrase and compound stress. *Linguist. Inq.* 24, 239–297.

Clark, S. E., and Gronlund, S. D. (1996). Global matching models of recognition memory: how the models match the data? *Psychon. Bull. Rev.* 3, 37–60. doi: 10.3758/BF03210740

Clifton, C., Frazier, L., and Deevy, P. (1999). Feature manipulation in sentence comprehension. *Rivisita Linguistica* 11, 11–39.

Clifton, C. Jr., Staub, A., and Rayner, K. (2007). "Eye movements in reading words and sentences," in *Eye Movement Research: A Window on Mind and Brain,* eds R. Van Gompel, M. Fisher, W. Murray, and R. L. Hill (Oxford: Elsevier Ltd.), 341–372.

Crowder, R. G. (1976). *Principles of Learning and Memory*. Hillsdale, NJ: Erlbaum.

Cunnings, I., Patterson, C., and Felser, C. (2015). Structural constraints on pronoun binding and coreference: evidence from eye movements during reading. *Front. Psychol.* 6:840. doi: 10.3389/fpsyg.2015.00840

Dayal, V., and Schwarzschild, R. (2010). "Definite inner antecedents and *wh*-correlates in sluices," in *Rutgers Working Papers in Linguistics*, Vol. 3, eds P. Staroverov, D. Altshuler, A. Braver, C.A. Fasola, and S. Murray (New Brunswick, NJ), 92–114.

Dickey, M. W., and Bunger, A. C. (2011). Comprehension of elided structure: evidence from sluicing. *Lang. Cogn. Process.* 26, 63–78. doi: 10.1080/01690961003691074

Dillon, B., Chow, W. Y., Wagers, M., Guo, T., Liu, F., and Phillips, C. (2014). The structure-sensitivity of memory access: evidence from Mandarin Chinese. *Front. Psychol.* 5:1025. doi: 10.3389/fpsyg.2014.01025

Dillon, B., Mishler, A., Sloggett, S., and Phillips, C. (2013). Contrasting intrusion profiles for agreement and anaphora: experimental and modeling evidence. *J. Mem. Lang.* 69, 85–103. doi: 10.1016/j.jml.2013.04.003

Dosher, B. A., and McElree, B. (1992). "Memory search: Retrieval processes in short-term and long-term recognition," in *Encyclopedia of Learning and Memory*, ed L. R. Squire (New York, NY: MacMillan), 398–406.

Fedorenko, E., Gibson, E., and Rohde, D. (2006). The nature of working memory capacity in sentence comprehension: evidence against domain-specific working memory resources. *J. Mem. Lang.* 54, 541–553. doi: 10.1016/j.jml.2005.12.006

Foraker, S., and McElree, B. (2007). The role of prominence in pronoun resolution: Active versus passive representations. *J. Mem. Lang.* 56, 357–383. doi: 10.1016/j.jml.2006.07.004

Foraker, S., and McElree, B. (2011). Comprehension of linguistic dependencies: speed−accuracy tradeoff evidence for direct−access retrieval from memory. *Lang. Linguist. Compass* 5, 764–783. doi: 10.1111/j.1749-818X.2011.00313.x

Frazier, L. (1978). *On Comprehending Sentences: Syntactic Parsing Strategies*. Ph.D. thesis, University of Connecticut, Storrs, CT.

Frazier, L. (2008). "Processing ellipsis: a processing solution to the undergeneration problem," in *The Proceedings of the 26th West Coast Conference on Formal Linguistics*, eds C. B. Chang and H. J. Haynie (Somerville, MA: Cascadilla Proceedings Project), 21–32.

Frazier, L., and Clifton, C. Jr. (1998). Comprehension of sluiced sentences. *Lang. Cogn. Process.* 13, 499–520. doi: 10.1080/016909698386474

Frazier, L., and Clifton, C. Jr. (2001). Parsing coordinates and ellipsis: Copy α. *Syntax* 4, 1–22. doi: 10.1111/1467-9612.00034

Frazier, L., and Clifton, C. Jr. (2005). The syntax-discourse divide: processing ellipsis. *Syntax* 8, 121–174. doi: 10.1111/j.1467-9612.2005.00077.x

Frazier, L., and Clifton, C. Jr. (2011). "D-linking and memory retrieval: the annoying case of sluicing," in *University of Massachusetts Occasional Papers in Linguistics: Processing Linguistic Structure.* Vol. 38, eds J. A. Harris and M. Grant (Amherst, MA: GLSA Publications), 37–52.

Frazier, L., Taft, L., Roeper, T., Clifton, C. Jr., and Ehrlich, K. (1984). Parallel structure: a source of facilitation in sentence comprehension. *Mem. Cogn.* 12, 421–430. doi: 10.3758/BF03198303

Gagnon, M. (2013). "Part and parcel of eliding partitives," in *Proceedings of the 23rd Semantics and Linguistic Theory Conference,* ed T. Snider (Santa Cruz, CA: University of Santa Cruz), 316–335.

Garrod, S., and Sanford, A. J. (1994). "Resolving sentences in a discourse context: how discourse representation affects language understanding," in *Handbook of Psycholinguistics,* ed M. Gernsbacher (New York, NY: Academic Press), 675–698.

Garrod, S. C., and Sanford, A. J. (1990). "Referential processing in reading: focusing on roles and individuals," in *Comprehension Processes in Reading,* eds D. A. Balota, G. B. Flores d'Arcias, and K. Rayner (Hillsdale, NJ: Erlbaum), 465–486.

Gordon, P. C., Hendrick, R., and Johnson, M. (2001). Memory interference during language processing. *J. Exp. Psychol.* 27, 1411–1423. doi: 10.1037/0278-7393.27.6.1411

Gordon, P. C., Hendrick, R., and Johnson, M. (2004). Effects of noun phrase type on sentence complexity. *J. Mem. Lang.* 51, 97–114. doi: 10.1016/j.jml.2004.02.003

Gordon, P. C., Hendrick, R., Johnson, M., and Lee, Y. (2006). Similarity-based interference during language comprehension: evidence from eye tracking during reading. *J. Exp. Psychol.* 32, 1304–1321. doi: 10.1037/0278-7393.32.6.1304

Gordon, P. C., Hendrick, R., and Levine, W. H. (2002). Memory-load interference in syntactic processing. *Psychol. Sci.* 13, 425–430. doi: 10.1111/1467-9280.00475

Greene, S. B., McKoon, G., and Ratcliff, R. (1992). Pronoun resolution and discourse models. *J. Exp. Psychol.* 18, 266–283. doi: 10.1037/0278-7393.18.2.266

Grosz, B., Joshi, A., and Weinstein, S. (1995). Centering: A framework for modeling the local coherence of discourse. *Comput. Ling.* 21, 203–225.

Gundel, J. (1985). 'Shared knowledge' and topicality. *J. Pragmat.* 9, 83–107. doi: 10.1016/0378-2166(85)90049-9

Gundel, J., Hedberg, N., and Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language* 69, 274–307. doi: 10.2307/416535

Harris, J. A. (2013). "Discourse accessibility and structural bias: Processing d-linked phrases in sluices," in *Poster at the 26th Annual CUNY Human Sentence Processing Conference* (Columbia, SC: University of South Carolina).

Harris, J. A. (2015). *Alternatives on Demand and the Locality Bias: Resolving discourse-linked wh-phrases in sluiced structures*. Los Angeles, CA: University of California.

Harris, J. A., and Carlson, K. (2015). Keep it local (and final): remnant preferences for let alone ellipsis. *Q. J. Exp. Psychol*. doi: 10.1080/17470218.2015.1062526. [Epub ahead of print].

Jäger, L. A., Benz, L., Roeser, J., Dillon, B. W., and Vasishth, S. (2015). Teasing apart retrieval and encoding interference in the processing of anaphors. *Front. Psychol.* 6:506. doi: 10.3389/fpsyg.2015.00506

Keppel, G., and Underwood, B. J. (1962). Proactive inhibition in short-term retention of single items. *J. Verbal Learn. Verbal Behav.* 1, 153–161. doi: 10.1016/S0022-5371(62)80023-1

Kush, D. (2013). *Respecting relations: Memory Access and Antecedent Retrieval in Incremental Sentence Processing*. Ph.D. thesis. University of Maryland, College Park, MD.

Kush, D., and Phillips, C. (2014). Local anaphor licensing in an SOV language: implications for retrieval strategies. *Front. Psychol.* 5:1252. doi: 10.3389/fpsyg.2014.01252

Lewis, R. L. (1996). Interference in short-term memory: the magical number two (or three) in sentence processing. *J. Psycholinguist. Res.* 25, 93–115. doi: 10.1007/BF01708421

Lewis, R. L., and Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cogn. Sci.* 29, 375–419. doi: 10.1207/s15516709cog0000_25

Lewis, R. L., Vasishth, S., and Van Dyke, J. A. (2006). Computational principles of working memory in sentence comprehension. *Trends Cogn. Sci.* 10, 447–454. doi: 10.1016/j.tics.2006.08.007

Link, G. (1983). "The logical analysis of plurals and mass terms: a lattice-theoretical approach," in *Meaning, Use, and the Interpretation of Language*, eds R. Bäuerle, C. Schwarze, and A. von Stechow (Berlin: Walter de Gruyter and Co), 303–323.

Martin, A. E. (2010). *Memory Operations and Structures in Sentence Comprehension: Evidence from Ellipsis*. Ph.D. thesis, New York University, New York, NY.

Martin, A. E., and McElree, B. (2008). A content-addressable pointer mechanism underlies comprehension of verb-phrase ellipsis. *J. Mem. Lang.* 58, 879–906. doi: 10.1016/j.jml.2007.06.010

Martin, A. E., and McElree, B. (2009). Memory operations that support language comprehension: evidence from verb-phrase ellipsis. *J. Exp. Psychol.* 35, 1231–1239. doi: 10.1037/a0016271

Martin, A. E., and McElree, B. (2011). Direct-access retrieval during sentence comprehension: evidence from sluicing. *J. Mem. Lang.* 64, 327–343. doi: 10.1016/j.jml.2010.12.006

Martin, A. E., Nieuwland, M. S., and Carreiras, M. (2012). Event-related brain potentials index cue-based retrieval interference during sentence comprehension. *Neuroimage* 59, 1859–1869. doi: 10.1016/j.neuroimage.2011.08.057

Martin, A. E., Nieuwland, M. S., and Carreiras, M. (2014). Agreement attraction during comprehension of grammatical sentences: ERP evidence from ellipsis. *Brain Lang.* 135, 42–51. doi: 10.1016/j.bandl.2014.05.001

McElree, B. (2000). Sentence comprehension is mediated by content-addressable memory. *J. Psycholinguist. Res.* 29, 111–123. doi: 10.1023/A:1005184709695

McElree, B. (2001). Working memory and focal attention. *J. Exp. Psychol.* 27, 817–835. doi: 10.1037/0278-7393.27.3.817

McElree, B. (2006). "Accessing recent events," in *The Psychology of Learning and Motivation,* Vol. 46, ed B. H. Ross (San Diego, CA: Academic Press), 155–200. doi: 10.1016/S0079-7421(06)46005-9

McElree, B., and Dosher, B. A. (1989). Serial position and set size in short-term memory: time course of recognition. *J. Exp. Psychol.* 118, 346–373. doi: 10.1037/0096-3445.118.4.346

McElree, B., and Dosher, B. A. (1993). Serial retrieval processes in the recovery of order information. *J. Exp. Psychol.* 122, 291–315. doi: 10.1037/0096-3445.122.3.291

McElree, B., Foraker, S., and Dyer, L. (2003). Memory structures that subserve sentence comprehension. *J. Mem. Lang.* 48, 67–91. doi: 10.1016/S0749-596X(02)00515-6

Merchant, J. (2001). *The Syntax of Silence: Sluicing, Islands, and the Theory of Ellipsis*. Oxford: Oxford University Press.

Moxey, L. M., and Sanford, A. J. (1993). *Communicating Quantities: A Psychological Perspective*. Hove, UK: Lawrence Erlbaum Associates, Inc.

Nairne, J. S. (2002). Remembering over the short-term: the case against the standard model. *Annu. Rev. Psychol.* 53, 53–81. doi: 10.1146/annurev.psych.53.100901.135131

Nicol, J., and Swinney, D. (1989). The role of structure in coreference assignment during sentence comprehension. *J. Psycholinguist. Res.* 18, 5–19. doi: 10.1007/BF01069043

Öztekin, I., and McElree, B. (2007). Retrieval dynamics of proactive interference: PI slows retrieval by eliminating fast assessments of familiarity. *J. Mem. Lang.* 57, 126–149. doi: 10.1016/j.jml.2006.08.011

Phillips, C., Wagers, M. W., and Lau, E. F. (2011). "Grammatical illusions and selective fallibility in real-time language comprehension," in *Experiments at the Interfaces. Syntax and Semantics,* Vol. 37, ed J. Runner (Bingley: Emerald Group Publishing Limited), 147–180.

Poirier, J., Wolfinger, K., Spellman, L., and Shapiro, L. P. (2010). The real-time processing of sluiced sentences. *J. Psycholinguist. Res.* 39, 411–427. doi: 10.1007/s10936-010-9148-9

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychol. Bull.* 124, 372–422. doi: 10.1037/0033-2909.124.3.372

R Development Core Team (2014). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Romero, M. (1998). *Focus and Reconstruction Effects in WH-Phrases*. Ph.D. thesis, University of Massachusetts Amherst.

Ross, J. R. (1967). *Constraints on Variables in Syntax*. Ph.D. thesis, MIT, Cambridge, MA.

Ross, J. R. (1969). "Guess who," in *Proceedings of the 5th Annual Chicago Linguistic Society,* eds R. I. Binnick, A. Davison, G. M. Green, and J. L. Morgan (Chicago, IL), 252–286.

Selkirk, E. O. (1984). *Phonology and Syntax: The Relation between Sound and Structure*. Cambridge: MIT Press.

Sprouse, J., Wagers, M., and Phillips, C. (2012). A test of the relation between working memory capacity and syntactic island effects. *Language* 88, 82–123. doi: 10.1353/lan.2012.0004

Staub, A., and Rayner, K. (2007). "Eye movements and on-line comprehension processes," in *The Oxford Handbook of Psycholinguistics,* ed G. Gaskell (Oxford: Oxford University Press), 327–342.

Sternberg, S. (1966). High-speed scanning in human memory. *Science* 153, 652–654. doi: 10.1126/science.153.3736.652

Sternberg, S. (1975). Memory-scanning: new findings and current controversies. *Q. J. Exp. Psychol.* 27, 1–32. doi: 10.1080/14640747508400459

Sturt, P. (2003). The time-course of the application of binding constraints in reference resolution. *J. Mem. Lang.* 48, 542–562. doi: 10.1016/S0749-596X(02)00536-3

van Craenenbroeck, J. (2010). *The Syntax of Ellipsis: Evidence from Dutch Dialects*. New York, NY: Oxford University Press.

Van Dyke, J. A. (2007). Interference effects from grammatically unavailable constituents during sentence processing. *J. Exp. Psychol.* 33, 407–430. doi: 10.1037/0278-7393.33.2.407

Van Dyke, J. A. (2011). Cue-dependent interference in comprehension. *J. Mem. Lang.* 65, 247–263. doi: 10.1016/j.jml.2011.05.002

Van Dyke, J. A., and Johns, C. L. (2012). Memory interference as a determinant of language comprehension. *Lang. Linguist. Compass* 6, 193–211. doi: 10.1002/lnc3.330

Van Dyke, J. A., and Lewis, R. L. (2003). Distinguishing effects of structure and decay on attachment and repair: a retrieval interference theory of recovery from misanalyzed ambiguities. *J. Mem. Lang.* 49, 285–413 doi: 10.1016/S0749-596X(03)00081-0

Van Dyke, J. A., and McElree, B. (2006). Retrieval interference in sentence comprehension. *J. Mem. Lang.* 55, 157–166. doi: 10.1016/j.jml.2006.03.007

Van Dyke, J. A., and McElree, B. (2011). Cue-dependent interference in comprehension. *J. Mem. Lang.* 65, 247–263. doi: 10.1016/j.jml.2011.05.002

van Gompel, R. P. G., Pickering, M. J., Pearson, J., and Liversedge, S. P. (2005). Evidence against competition during syntactic ambiguity resolution. *J. Mem. Lang.* 52, 284–307. doi: 10.1016/j.jml.2004.11.003

van Gompel, R. P. G., Pickering, M. J., and Traxler, M. J. (2001). Reanalysis in sentence processing: evidence against current constraint-based and two-stage models. *J. Mem. Lang.* 45, 225–258. doi: 10.1006/jmla.2001.2773

Vasishth, S., Brüssow, S., Lewis, R. L., and Drenhaus, H. (2008). Processing polarity: how the ungrammatical intrudes on the grammatical. *Cogn. Sci.* 32, 685–712. doi: 10.1080/03640210802066865

Wagers, M. W., Lau, E. F., and Phillips, C. (2009). Agreement attraction in comprehension: representations and processes. *J. Mem. Lang.* 61, 206–237. doi: 10.1016/j.jml.2009.04.002

Waugh, N. C., and Norman, D. A. (1965). Primary memory. *Psychol. Rev.* 72, 89–104. doi: 10.1037/h0021797

Xiang, M., Dillon, B., and Phillips, C. (2009). Illusory licensing effects across dependency types: ERP evidence. *Brain Lang.* 108, 40–55. doi: 10.1016/j.bandl.2008.10.002

Xiang, M., Grove, J., and Giannakidou, A. (2013). Dependency dependent interference: NPI interference, agreement attraction, and global pragmatic inferences. *Front. Psychol.* 4:708. doi: 10.3389/fpsyg.2013.00708

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2015 Harris. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 26 January 2016
doi: 10.3389/fpsyg.2016.00027

Filling the Silence: Reactivation, not Reconstruction

Dario L. J. F. Paape *

Department of Linguistics, University of Potsdam, Potsdam, Germany

In a self-paced reading experiment, we investigated the processing of sluicing constructions ("sluices") whose antecedent contained a known garden-path structure in German. Results showed decreased processing times for sluices with garden-path antecedents as well as a disadvantage for antecedents with non-canonical word order downstream from the ellipsis site. A *post-hoc* analysis showed the garden-path advantage also to be present in the region right before the ellipsis site. While no existing account of ellipsis processing explicitly predicted the results, we argue that they are best captured by combining a local antecedent mismatch effect with memory trace reactivation through reanalysis.

Keywords: ellipsis processing, garden-path effect, German, retrieval, reconstruction, self-paced reading

1. INTRODUCTION

Besides verb-phrase ellipsis, sluicing (Ross, 1969) is probably the most-studied ellipsis variety in both theoretical linguistics (e.g., Chung et al., 1995; Merchant, 2001; Potsdam, 2007) and psycholinguistics (e.g., Poirier et al., 2010; Dickey and Bunger, 2011; Yoshida et al., 2013). In sluicing, an entire clause is left out and a wh-element remains behind, as in (1).

(1) John saw Mary, but I don't remember when ____.
____ = John saw Mary

Sluicing is anaphoric: to interpret (1), the semantics of the antecedent (*John saw Mary*) must somehow be inserted into the gap behind the word *when* to derive the meaning *I don't remember when John saw Mary*. We write "meaning" because deriving an interpretation is the fundamental goal of sentence processing, not because it is necessarily clear that the relevant representation of the antecedent is semantic in nature. There is an ongoing debate as to whether syntactic structure is also present at ellipsis sites (cf. Cai et al., 2013, and references therein), or whether one should adopt a more discourse-centered approach to the gap-filling process (e.g., Hardt, 1993; Kehler, 2000). Since the evidence to date, at least in our view, does not unequivocally favor any of these views, we will not take a stance with regard to the representation question. We will, however, use syntactic terminology throughout the article for ease of reference.

Even with the question of *what* is inserted into the gap set aside, another point of debate has been *how* it ends up there. Ross (1967) was perhaps the first to explicitly propose a deletion approach to ellipsis (in this case, verb-phrase ellipsis): the missing bit of structure is assumed to be underlyingly present, but its phonological representation is erased under identity with the antecedent.[1] Under the approach taken by Williams (1977), ellipsis involves copying. Like Ross (1967), Williams assumes invisible syntax at the gap, but the terminal symbols of this structure are null elements

[1]There is no condition of strict identity, however, as several kinds of mismatch can be observed, as in *The car was supposed to be washed but nobody did* (e.g., Kertz, 2000; Merchant, 2013, submitted).

OPEN ACCESS

Edited by:
Colin Phillips,
University of Maryland, USA

Reviewed by:
Elsi Kaiser,
University of Southern California, USA
Jesse Harris,
University of California, Los Angeles, USA

***Correspondence:**
Dario L. J. F. Paape
paape@uni-potsdam.de

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: *25 August 2015*
Accepted: *07 January 2016*
Published: *26 January 2016*

Citation:
Paape DLJF (2016) Filling the Silence: Reactivation, not Reconstruction. Front. Psychol. 7:27. doi: 10.3389/fpsyg.2016.00027

(Wasow, 1972). The ellipsis is interpreted by copying the terminals (that is, words) from the antecedent to the appropriate positions within the gap.

From a processing perspective, it is not enough to claim that the syntax is there in the silence: the processor must have some way of creating it. A reader of (1) would have to first infer that deletion has applied, then identify the antecedent and finally reconstruct it at the gap. The main aim of the current study is to investigate how this "reconstruction" is to be conceived of: does the parser rebuild the antecedent's structure at the ellipsis site, or does it come to be there by virtue of some other mechanism?

One might think of dispensing with the idea of invisible structure altogether. The approach of Hardt (1993) is explicitly non-syntactic in nature and treats ellipsis as an unstructured proform that refers to a stored meaning in a discourse model. The notion of copying does not enter into the picture; ellipsis acts rather like a pointer or a hyperlink into memory than as an entity of its own. This conception can be related to the processing of other types of anaphors: It is not commonly assumed that in a sentence such as *The man from England drank tea, but he didn't drink coffee*, the pronoun *he* will contain the syntactic structure of the NP *the man from England* at any level of representation. Instead, an identity of *reference* between the two expressions seems to obtain (cf. Grinder and Postal, 1971, p. 269).

Note that the opposition between copying and the "memory pointer" approach is orthogonal to that between syntactic and semantic/discourse representations (cf. Phillips and Parker, 2014). Semantic representations could also be copied, just as syntactic representations could be pointed to. The processing literature has focused mainly on the copying/pointing dichotomy, even though some studies have also tested whether there is syntactic priming from ellipsis sites, with mixed results (Cai et al., 2013; Xiang et al., 2014). Murphy (1985) appears to have been the first to systematically look for effects of antecedent length on reading times for elliptical clauses, in this case the sentence *Later, his uncle did too* in (2).

(2) a. Jimmy swept the floor. Later, his uncle did too.
b. Jimmy swept the tile floor behind the chairs free of hair and cigarettes. Later, his uncle did too.

Despite being concerned with verb-phrase ellipsis, we assume that this study is informative with regard to sluicing as well, since the most parsimonious hypothesis would be that all types of ellipsis are processed in the same way. The reasoning behind Murphy's manipulation was that "[l]onger antecedents would be expected to affect a copying process, since the longer the string that must be copied onto the anaphor, the longer it should take to understand the anaphor" (p. 293). If there was no copying, so the argument goes, then reading times for the second sentence should not differ between (2a,b). Murphy found that reading times for the elliptical sentence were increased by about 260 ms when the antecedent was long rather than short. Interestingly, this difference disappeared when another sentence was inserted between antecedent and ellipsis[2].

[2]Murphy was concerned that the observed complexity effect was simply due to processing spillover from the antecedent sentence into the ellipsis sentence, but the intervening sentence did not show any effects either.

The system Murphy proposes is one in which there are two processes, namely copying and discourse-based "plausible reasoning," which operate in parallel, with the process that finishes first supplying the antecedent. When the antecedent is far away, the speed and/or availability of copying suffers and readers fall back on plausible reasoning, which by assumption is not influenced by complexity effects. Tanenhaus and Carlson (1990, p. 261) remain unconvinced by Murphy's (1985) evidence for copying, arguing that the length manipulation "also introduced potential scope and attachment ambiguities"[3]. The authors favor a pointer-based approach, while allowing for the possibility that there are both a syntax- and a discourse-based process at work.

Two additional important findings come from an experiment by Frazier and Clifton (2000) and a series of experiments by Martin and McElree (2008), all on verb-phrase ellipsis.

(3) Frazier and Clifton (2000), Experiment 1 B
a. Sarah left her boyfriend last May. Tina did too.
b. Sarah got up the courage to leave her boyfriend last May. Tina did too.

(4) Martin and McElree (2008), Experiment 3
a. The history professor understood Roman mythology, …
b. The history professor understood Rome's swift and brutal destruction of Carthage, …

…but the principal was displeased to learn that the over-worked students attending summer session did not.

Frazier and Clifton's study used self-paced reading and found no difference in reading times between (3a,b) for the sentence *Tina did too*. Martin and McElree's Experiment 3, which used sentences such as (4a,b), employed a speed-accuracy trade-off paradigm with end-of-sentence acceptability judgments. No effect of antecedent complexity on processing times was observed in this study and two further experiments, which the authors interpret as evidence for a pointer-based approach.

Here is where terminology becomes an issue, as Frazier and Clifton (2001) explain their earlier results by means of a mechanism called Copy α. Copy α becomes available when the scope of an ellipsis can be uniquely identified and serves as a shortcut to syntactic structure: instead of being built step-by-step, which would be computationally costly, the silent syntax is copied from the antecedent. As this process is assumed to be "cost-free," the complexity of the copied structure has no influence on processing time. Frazier and Clifton's use of the copying metaphor is not very intuitive (cf. Martin and McElree, 2008, p. 882f.), as it should take more time to copy a larger amount of information, in concordance with Murphy (1985) prediction[4].

[3]It is not obvious which ambiguities the authors are referring to, or how they would impact processing under an approach without copying. It should be pointed out, however, that interpreting the ellipsis with the long antecedent in (2) requires an additional assumption, namely that the floor became dirty again between the first and the second sweeping.

[4]A possible analogy would be the copying of a file from one location on a hard drive to another, which becomes more time-consuming as file size increases.

Indeed, Frazier and Clifton (2001, p. 17) themselves explain that a pointer would be a possible implementation of Copy α and in a later paper (Frazier and Clifton, 2005) describe Copy α as equivalent to "sharing" one structure between antecedent and ellipsis (cf. also Murguia, 2004). We will thus treat pointer-based approaches, Copy α and "sharing" as variants of one and the same idea, namely that the antecedent's structure is available in memory and can be retrieved from there as-is, without any additional costly computations.

Phillips and Parker (2014, p. 91) make note of several methodological problems in both of the above studies. Frazier and Clifton's (2000) experiment used only a small number of experimental items, all of which had the ellipsis at the very end of a sentence, where wrap-up effects might mask an influence of antecedent complexity. Additionally, comprehension questions were not asked after every trial and never targeted the interpretation of the ellipsis[5]. The ungrammatical sentences in Martin and McElree's (2008) study replaced the subject of the elliptical clause by an inanimate NP (*the overly worn books*), thus making the judgments fairly easy and possibly leading subjects to engage in superficial processing. Given these concerns, Phillips and Parker judge the results to be inconclusive, but also point out that it would be difficult to design an experiment that would provide convincing evidence for or against complexity effects.

Given this state of affairs, we think it worthwhile to look back at Frazier and Clifton's (2001) distinction between a syntactic structure that is computed step-by-step and one that is retrieved from memory. What happens when the antecedent is structured in a way that is known to fool the "normal" incremental parsing mechanism, that is, if it contains a garden path? Assuming a serial parsing architecture, recovering from a syntactic misanalysis involves reanalyzing the ambiguous region and assigning the same structure that would be computed for an unambiguous control sentence. Since the final memory representations for ambiguous and unambiguous sentences are the same, pointer-based approaches and Copy α would predict that there should be no difference in processing times at the ellipsis site. If, on the other hand, ellipsis is not resolved by linking the gap to a complete structure in memory, different scenarios are possible. One would be that the antecedent is accessed in memory as a word string, and that syntax and semantics are assigned to this string in the usual way, that is, incrementally. However, as verbatim memory is known to be highly fallible even in recognition tests (Sachs, 1967; Murphy and Shapiro, 1994), it may be unrealistic to assume that strings are *recalled* literally for ellipsis processing. The account of Kim et al. (2011) proposes that not the words themselves but their features are accessed by the parser at the ellipsis site, and that "derivations in an initial conjunct [are allowed] to do double-duty in a second conjunct" (p. 346). Their account states that "once [...] an appropriate antecedent is found, [its derivation] becomes available to the parser, just as if it were located at the elision point in the input string" (p. 346), essentially claiming that the derivation is carried out twice. Now, if the sentence processor has no way of "remembering" that it was garden-pathed by the antecedent, there is a chance that it will be garden-pathed again at the ellipsis site.

A model that is, in principle, compatible with both the pointer/sharing approach and the "reparsing" account is the cue-based retrieval parser of Lewis and Vasishth (2005). In this model, syntactic phrases are stored in working memory as chunks than can be retrieved if needed. For complex phrases, both the phrase itself and its constituent parts, such as the subject of a verb phrase, are stored, along with their grammatical relations. When an ellipsis site is encountered, the parser would thus have the opportunity to retrieve either the whole antecedent as one chunk, as under a pointer-based account, or to retrieve whatever chunks are contained within the antecedent and build a new structure, as under the "reparsing" view. The latter possibility may become especially attractive in cases of antecedent-ellipsis mismatch, where a strict isomorphism condition cannot be upheld (e.g., Merchant, 2001). As in the case of Kim et al. (2011) chunks are conceived of as feature bundles and thus no verbatim memory of the antecedent is required for retrieval. In fact, both Kim et al. (2011) and Lewis and Vasishth (2005) explicitly assume that the linear order of constituents is not represented in the syntax.

The "parse twice" approach might seem counterintuitive, but is in fact no less parsimonious than Frazier and Clifton's Copy α, given that it needs no special machinery besides access to grammatical features inside the antecedent structure. One would not expect the garden-path effect at the ellipsis site to be of the same strength as the one observed for the antecedent, just as one would not expect the reading time for *when* in (1) to be equal to that of *John saw Mary*. Several steps involved in lexical access can be omitted during ellipsis processing. Simner and Smyth (1999) suggest that instead of using lexemes, ellipsis targets word lemmas, which would be compatible with the "feature bundle" view described above. Additionally, ellipsis normally occurs in environments that feature a high amount of syntactic parallelism. If a parallel structure is expected, the relevant routines may be activated beforehand or at least be assigned a higher rank when the parser decides which structure to build at the ellipsis site, which can be seen as an instance of syntactic priming (Dubey et al., 2008; Dickey and Bunger, 2011). Given this assumption, however, it might be that in case of a garden path the preferred but incorrect structure will feature into the calculation, making the ellipsis more difficult to process than in cases where the antecedent's structure is unambiguous. While Arai et al. (2014) found evidence that resolving an ambiguity in a prime sentence makes processing of the same ambiguity in the target sentence easier when the same verb is repeated (see also Branigan et al., 2005), it is unclear whether ellipsis constitutes "repetition."

In our experiment, we used a known garden-path structure in German to test the—equivalent—predictions of pointer- and sharing-based approaches against those of a reconstruction-based approach of ellipsis processing. The former two predict that garden-pathing within the antecedent clause should have no

[5]An anonymous reviewer suggests that this might have been in order not to risk making the subjects aware of the experimental manipulation. While this is a fair point, it has been shown that subjects adapt their processing strategy to task demands, trying to minimize effort through underspecification (e.g., Foertsch and Gernsbacher, 1994; Swets et al., 2008). If one intends to investigate "deep" processing, we believe that the latter risk outweighs the former, being aware that the opposite stance is equally tenable.

effect at the ellipsis site while the latter predicts that the pattern observed at the point of disambiguation will reappear, although the effect size may be significantly smaller. To anticipate the results, we found an unpredicted pattern that was inconsistent with a reconstruction approach, but compatible with pointer- and sharing-based accounts if additional assumptions are made.

2. MATERIALS AND METHODS

2.1. Stimuli

It is known that German readers prefer to assign a subject interpretation to a sentence-initial NP that is ambiguous between a subject and an object reading, which results in a garden path when it is disambiguated toward an object role (cf. also Hemforth, 1993, among others). Different explanations for the subject preference have been proposed. For instance, Gorrell's (1996) approach assumes that the parser favors structural simplicity; under his analysis, deriving an OVS structure requires more movement operations (and thus more traces) than deriving an SVO structure, where the object presumably remains in the position at which it is base-generated. Schlesewsky et al. (2000) consider the possibility that the subject preference is due to a frequency-based "tuning" effect (e.g., Mitchell et al., 1995), reporting over 90% nominative-initial main clauses in a corpus study. Still other possibilities are that subject-first is a default parsing assumption, as has been proposed for English (e.g., Bever, 1970; Grodzinsky, 1986; Fishbein and Harris, 2014). If one follows the current standard analysis of German clause structure, where S(O)V word order is assumed to be basic and all other word orders are derived through movement (e.g., Schwartz and Vikner, 2007), the reanalysis of an object-initial structure will minimally involve removing co-indexation between an assumed trace position for the subject and the initial noun phrase, as well as postulating a trace position for an object.

The garden-path effect incurred by the non-canonical structure is stronger when disambiguation is achieved through agreement on the finite verb rather than through case marking on another NP (Meng and Bader, 2000). As shown in (5), we used indefinite NPs instead of the wh-marked NPs employed by Meng and Bader. Case marking on the *sympathizer* NP is either ambiguous (5a/b) or unambiguous (5c/d). The auxiliary *hatte(n)*, "had," agrees either with the singular *sympathizer* or with the plural *rebels* NP, thereby signaling either OVS (5a/c) or SVO word order (5b/d). The result is a 2 × 2 design with the factors word order and case marking. Diamonds indicate the boundaries of presentation regions in the experiment, subscripts indicate region coding for the statistical analysis.

(5) a. **Ambiguous / OVS**

Eine Sympathisantin der Opposition$_{np1}$ ◇
A sympathizer.fem.**nom/acc** of the opposition
hatten$_{aux}$ ◇ die Rebellen$_{np2}$ ◇ …
had.pl the rebels.nom/acc

b. **Ambiguous / SVO**

Eine Sympathisantin der Opposition$_{np1}$ ◇
A sympathizer.fem.**nom/acc** of the opposition
hatte$_{aux}$ ◇ die Rebellen$_{np2}$ ◇ …
had.sg the rebels.nom/acc

c. **Unambiguous / OVS**

Einen Sympathisanten der Opposition$_{np1}$ ◇
A sympathizer.masc.**acc** of the opposition
hatten$_{aux}$ ◇ die Rebellen$_{np2}$ ◇ …
had.pl the rebels.nom/acc

d. **Unambiguous / SVO**

Ein Sympathisant der Opposition$_{np1}$ ◇
A sympathizer.masc.**nom** of the opposition
hatte$_{aux}$ ◇ die Rebellen$_{np2}$ ◇ …
had.sg the rebels.nom/acc

… laut einem Bericht$_{adj}$ ◇ maßgeblich unterstützt,$_{vp}$
according to a report decisively supported
◇ aber ◇ die Regierung ◇ konnte ◇ nicht ◇
but the government could not
nachweisen,$_{wh-1}$ ◇ wie,$_{wh}$ ◇ so sehr$_{wh+1}$ ◇ sich$_{wh+2}$ ◇
substantiate **how** so greatly itself
die Untersuchungskommission$_{wh+3}$ ◇ auch ◇
the investigative commission too
bemühte.
struggled

"The rebels had supported a sympathizer (OVS, a/c)/A sympathizer had supported the rebels (SVO, b/d), but the government could not substantiate how, no matter how hard the investigative commission tried."

The antecedent clause ends at *unterstützt*, "supported." It is conjoined with a second clause by *aber*, "but," which contains a sluicing site (or "sluice") at *wie*, "how." All wh-phrases in the experiment were "sprouted" (Chung et al., 1995), that is, they had no explicit correlate in the antecedent. We only used adjunct wh-phrases since argument wh-phrases are case-marked in German, which would have introduced a potential confound. The other wh-phrases used were several expressions meaning "why" (*warum*, *weshalb*, *wieso*), *wo*, "where," *wann*, "when," *womit*, "with what," *wozu* "to what (end)," and *wobei*, "at what" (combined with the verb *unterstützen*, "to support"). The part of the sentence following the sluicing site was intended as a spillover region. We could have used only conditions (5a) and (5c) to look for an effect of reanalysis, but decided to also include (5b) and (5d) as control conditions since otherwise reanalysis would be completely confounded with the gender of the initial NP. Additionally, even though condition (5b) is initially ambiguous, there should be no reanalysis as readers will assume SVO order by default (cf. Meng and Bader, 2000); we can thus control for temporarily ambiguous antecedents being processed differently from unambiguous ones. Thirty-two sentences were created according to this schema for use in the experiment. A complete list of the experimental materials is given in the appendix. The

stimuli were combined with ninety-six filler sentences featuring various constructions.

We expected a garden-path effect to occur at the auxiliary of the antecedent clause in the form of a word order × case marking interaction. Meng and Bader (2000) observed longer reaction times in a grammaticality judgment task for OVS than for SVO sentences, indicating that OVS order is overall more difficult to process. In (5a), however, the *sympathizer* NP presumably has to be reanalyzed from subject to object, which should further increase processing time. If ellipsis acts as a pointer into memory, no interaction between the experimental factors should appear at *wie*, "how," as neither the scope of the ellipsis nor the availability of a completely analyzed antecedent structure vary between conditions. If, however, the syntax of the ellipsis site has to be constructed by normal parsing routines, the garden-path effect should reappear at this position, though most likely with reduced magnitude.

We had no specific predictions as to possible effects of OVS vs. SVO word order at the ellipsis site, but a *post-hoc* hypothesis will be developed in the discussion section. A complication concerning the predictions of both accounts that did not become apparent to us until after the experiment is that inserting a verb-second antecedent into the ellipsis site verbatim is impossible in our stimuli, as German subordinate clauses are generally required to be verb-final. The predictions outlined above are valid for well-formed antecedents, but should pertain to mismatched antecedents as well if certain additional assumptions are made, as will be explained shortly.

2.2. Participants

Sixty students from the University of Potsdam were recruited for the study. All subjects were native speakers of German and were either paid 6 € or received course credit for the participation. Informed consent was obtained from all participants prior to testing.

2.3. Procedure

The sentences were presented using the moving window self-paced reading technique (Just et al., 1982), which was implemented using the Linger software (Rohde, 2003; http://tedlab.mit.edu/~dr/Linger/). Participants sat in front of a PC in a quiet room and were instructed to read silently and at their own pace. Sentences were presented in 20 pt Courier New font according to a latin square procedure. At the beginning of each trial, all characters were masked with underscores. Participants completed two practice trials before the experiment proper. The order of fillers and experimental sentences was randomized at runtime. Each trial was followed by a comprehension test which took one of two forms: either a statement about the preceding sentence had to be judged as true or false, or a gap in a statement had to be filled by selecting one out of four options. Some test statements targeted the argument structure of the antecedent (*Rebels had supported a sympathizer of the opposition. [Yes/No]*), while others targeted other kinds of information from the sentence. The ratio of true to false statements for the judgment test was balanced. For a subset of fill-in-the-gap statements appearing after experimental sentences, participants had to supply the critical wh-pronoun[6].

3. RESULTS AND DISCUSSION

3.1. Data Analysis

After 15 participants had completed the experiment, it was noticed that three experimental items contained a typographical error in one condition each. The errors were removed and data from the corresponding trials were excluded from the statistical analysis. The remaining data were analyzed using the `R` software environment (R Core Team, 2015) by fitting linear mixed-effects models to individual regions of interest with the `lme4` package (Bates et al., 2014). The models included varying intercepts and slopes by subjects and by items. The code and data will be released with the publication of this paper. When the estimate for a slope adjustment was zero, the random effect was dropped from the model, along with any associated higher-order effects. When a model failed to converge, random effects were removed, starting with the effect that accounted for the smallest amount of variance, until convergence was obtained. Sum contrasts were defined for the experimental factors word order and case marking and entered into the models as fixed effects. For word order, the OVS conditions were coded as 1 and the SVO conditions as −1, respectively. For case marking, the ambiguous conditions were coded as 1 and the unambiguous conditions as −1. Since processing spillover is a known concern in self-paced reading, the reading time for the immediately preceding region was also entered into all models after being appropriately transformed (see below) and subsequently centered. The addition of this parameter improved model fit for all regions of interest[7], but the method is by no means guaranteed to eliminate spillover entirely, for instance if subjects postpone processing and keep "tapping" the button at fixed time intervals (Witzel et al., 2012).

An underlying assumption in linear modeling is that the residuals are approximately normally distributed. As this was not the case when raw reading times were used as the dependent variable, we applied the Box-Cox procedure (Box and Cox, 1964; Venables and Ripley, 2002), which suggested a reciprocal transformation (1/RT). Reciprocal reading times were multiplied by −1000 to make the parameters easier to interpret. Additionally, all data points corresponding to reading times below 150 ms were removed, which resulted in a loss of less than one per cent of data in all cases. Effects were judged as significant if $t \geqslant 2$. Model output is shown in **Table 2**.

3.2. Comprehension Accuracy

Participants' overall comprehension accuracy was at 90%, though accuracy for experimental items was somewhat lower at 82%. Overall, subjects were most accurate at supplying the wh-pronoun (92% accuracy) and least accurate at verifying

[6]Though the specific example in (5) was not accompanied by this kind of test, a possible fill-the-gap statement could have been *The government could not substantiate ____ rebels had supported a sympathizer of the opposition. [why/how/when/if]*.

[7]Improvement of fit was assessed through likelihood ratio tests comparing models with and without the spillover predictor.

statements about the argument structure of the antecedent (72% accuracy), with the rest of the comprehension tests falling in between (86% accuracy). All further analyses were conducted without distinguishing between question types, unless otherwise noted. A linear mixed-effects model was fit to question response times using the same procedure described above for reading times. The analysis revealed no significant effects of the experimental manipulation. An analogous model with reciprocal response time as an additional predictor was fit to response accuracies using a logit link function. The fit showed an effect of response time such that accuracy dropped with increased delay ($\hat{\beta} = -0.13$, $se = 0.03$, $t = -5.18$), as well as a significant word order × case marking interaction ($\hat{\beta} = -0.18$, $se = 0.07$, $t = -2.74$), which nested contrasts[8] revealed to be driven by the OVS/ambiguous condition eliciting more incorrect responses than the SVO/ambiguous condition ($\hat{\beta} = -0.27$, $se = 0.13$, $t = -2.09$). To investigate further, we created a new contrast between questions that queried the role of the arguments in the antecedent and questions that did not. When this distinction was entered into the model[9], it turned out to be highly predictive of accuracy ($\hat{\beta} = -0.66$, $se = 0.16$, $t = -4.24$), indicating that questions about argument structure were more difficult to answer than other question types. At the same time, the word order × case marking interaction was significant ($\hat{\beta} = -0.17$, $se = 0.07$, $t = -2.63$), but there was no three−way interaction. There was thus no indication that comprehension failure for questions targeting argument structure was limited to garden-path sentences. Why answering questions about garden-path sentences should be difficult even when the temporary ambiguity is not targeted remains mysterious for the time being.

3.3. Reading Times

Table 1 shows the mean raw reading times for the analyzed regions of interest. **Figure 1** shows residual mean reading times for each region of the antecedent. Residualization was carried out by fitting a linear mixed-effects model with region length as a fixed effect and random slopes by subject. Unresidualized reciprocal reading times (see above) were used in the statistical analysis. A main effect of word order appeared at the auxiliary ($\hat{\beta} = 0.03$, $se = 0.01$, $t = 2.07$), such that OVS was processed more slowly than SVO, which is likely due to the additional plural suffix in the OVS conditions. On the second NP, there were main effects of word order ($\hat{\beta} = 0.04$, $se = 0.01$, $t = 3.02$) and case marking ($\hat{\beta} = 0.04$, $se = 0.01$, $t = 3.3$), such that SVO was read faster than OVS and unambiguous sentences were read faster than ambiguous ones. There was also a significant interaction between the factors ($\hat{\beta} = 0.02$, $se = 0.01$, $t = 2.12$), which nested contrasts revealed to be driven by OVS clauses taking longer to read in the presence of ambiguous case marking ($\hat{\beta} = 0.07$, $se = 0.02$, $t = 3.68$). The preverbal adjunct again showed a main effect of word order ($\hat{\beta} = -0.02$, $se = 0.01$, $t = -2.38$); at this position, OVS clauses were read faster than SVO clauses[10].

TABLE 1 | Untrimmed raw mean reading times in milliseconds by condition for antecedent, ellipsis and spillover regions, standard errors in parantheses.

		OVS/ amb.	SVO/ amb.	OVS/ unamb.	SVO/ unamb.
A sympathizer ...	np1	1793 (48)	1760 (39)	1830 (41)	1651 (39)
had.sg/pl	aux	519 (17)	474 (8)	499 (12)	474 (10)
the rebels	np2	1021 (28)	976 (28)	913 (23)	921 (27)
according to ...	adj	1041 (26)	1107 (29)	1066 (28)	1135 (31)
decisively supported ...	vp	892 (23)	887 (24)	868 (22)	900 (26)
substantiate	wh−1	471 (8)	485 (10)	493 (9)	486 (10)
how	wh	423 (7)	427 (7)	422 (6)	434 (7)
so greatly	wh+1	437 (7)	452 (8)	449 (9)	449 (8)
itself	wh+2	578 (15)	564 (15)	591 (16)	584 (18)
the ... commission	wh+3	571 (18)	580 (16)	604 (17)	590 (17)

Figure 2 shows the mean reading times from the region right before the ellipsis site to three words after the ellipsis site, again in residualized form. No significant effects appeared at the wh-pronoun or in the immediately following region. In the next region (wh+2), there was a main effect of word order ($\hat{\beta} = 0.03$, $se = 0.01$, $t = 2.02$), such that OVS clauses took longer to read than SVO clauses. For this position, closer inspection of the model revealed one very short reading time (177 ms) to be highly influential in the fit, and removing this value resulted in the effect merely approaching significance ($\hat{\beta} = 0.02$, $se = 0.01$, $t = 1.89$). In the third region after the wh-pronoun (wh+3), a word order × case marking interaction reached significance ($\hat{\beta} = -0.03$, $se = 0.01$, $t = -2.02$), due to the OVS/ambiguous condition being read faster than the OVS/unambiguous condition, with no single condition driving the interaction. During data analysis we noticed that five experimental sentences featured gender-marked pronouns at position wh+2, which presents a possible confound. Adding the presence vs. absence of a pronoun as a sum-coded predictor did, however, not change the results found at regions wh+2 and wh+3.

One might think that the interaction found at position wh+3 stemmed from occasional processing breakdowns in the OVS/ambiguous sentences. We assume that these would be due to failures in processing the antecedent, which would leave the parser without an adequate retrieval target for the ellipsis. To test this hypothesis, we added the reading time for the second NP, which is expected to reflect the difficulty of the garden path, to the reading time model for position wh+3 on the same trial. While this measure turned out to be a highly significant predictor ($\hat{\beta} = 0.13$, $se = 0.02$, $t = 5.51$), the word order × case marking interaction also stayed significant and indeed became stronger ($\hat{\beta} = -0.03$, $se = 0.01$, $t = -2.21$). This suggests that while the time spent processing the garden-path influences retrieval

[8]For this analysis, case marking was treated as nested within word order. One sum contrast compared the two ambiguous conditions, one compared the two unambiguous conditions, and a third one the OVS vs. SVO conditions.

[9]The fixed effect of reciprocal response time was removed from this model as it consistently led to convergence failure.

[10]Speculatively, this effect may be due to readers trying to make up for lost time after having been slowed down.

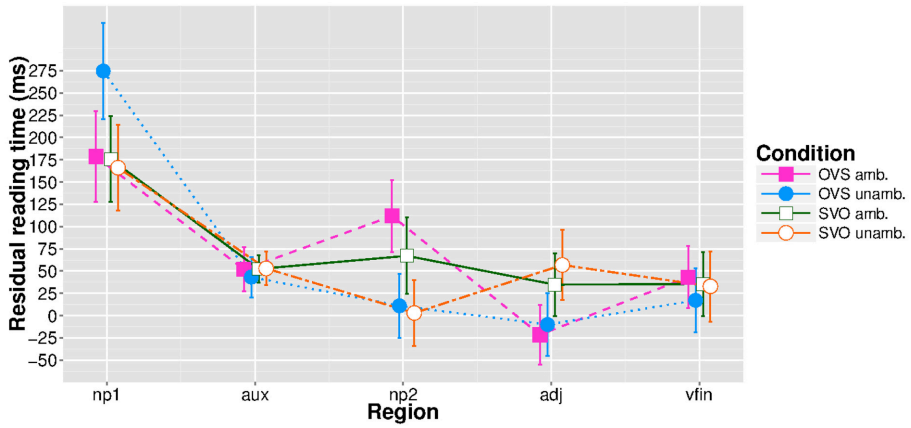


FIGURE 1 | Residual reading times for the antecedent regions, extreme values removed. Error bars represent 95% confidence intervals.

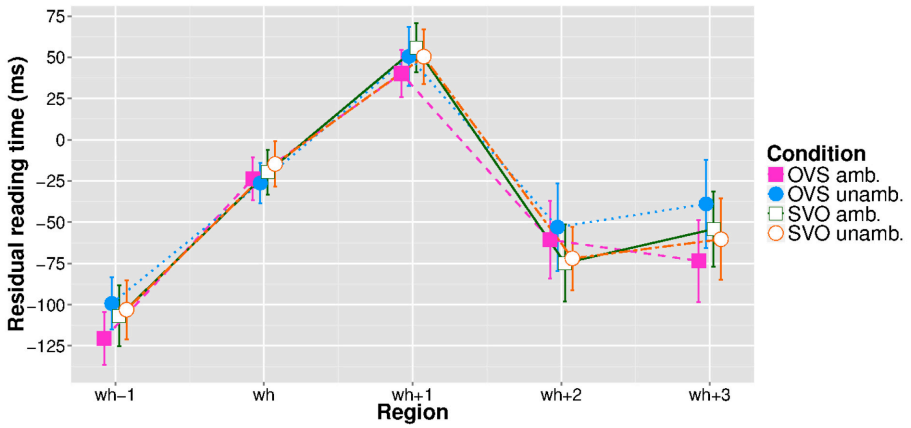


FIGURE 2 | Residual reading times for the pre-ellipsis, ellipsis, and spillover regions, extreme values removed. Error bars represent 95% confidence intervals.

difficulty, there are factors above and beyond this measure which determine processing effort at the ellipsis site. In a further test, we added reading times for both the second NP and position wh+3 to the response accuracy model reported above. The reasoning behind this was that processing failure at either position could lead to incorrect responses. Adding these parameters did, however, not change the result. We also compared the median reading time in the OVS/ambiguous condition for position wh+3 with the overall median reading time for the experimental items. The difference lay within reasonable bounds (439 ms, *se* 18 ms vs. 473 ms, *se* 2 ms), indicating that very short RTs from processing failures were not pushing down the median. Congruently with this, a visual inspection of a density plot of RTs at position wh+3 did not indicate a mode or tail of fast reading times, nor did Hartigan's Dip Test (Hartigan and Hartigan, 1985) yield any evidence for bimodality. Finally, we removed all trials with incorrect responses to the comprehension test, which amounted to 18% of the data for position wh+3, and refit our model. Note that an incorrect answer does not necessarily mean that parsing failed; misinterpretations could, for instance, arise from fragments of discarded analyses in memory (see below). Nevertheless, the results of the comprehension test are the only pertinent measure available to us. With one fifth of data removed, the word order × case marking interaction stayed near the significance threshold ($\hat{\beta} = -0.02$, $se = 0.01$, $t = -1.62$) and became marginally significant when antecedent reading time was added as a predictor ($\hat{\beta} = -0.03$, $se = 0.01$, $t = -1.86$). The loss of significance is not particularly unexpected given the loss of statistical power incurred by removing data. To our minds, these results do not indicate that processing failure was a factor in decreasing reading times for the OVS/ambiguous condition.

3.4. Discussion

The expected garden-path effect for the antecedent appeared one region later than predicted, at the second NP, showing that the experimental manipulation was successful. While no effects were found at the ellipsis site itself, OVS antecedents led to longer reading times two regions downstream from the wh-pronoun.

TABLE 2 | Coefficient estimates, standard errors and *t*-values for the linear mixed-effects models fit to reciprocal reading times at the indicated regions of interest.

	aux				np2		
	Estimate	**Std. Error**	***t*-value**		**Estimate**	**Std. Error**	***t*-value**
(Intercept)	−2.19	0.07	−32.82	(Intercept)	−1.32	0.08	−17.08
Case marking	0.01	0.01	0.51	Case marking	0.04	0.01	3.30
Word order	0.03	0.01	2.07	Word order	0.04	0.01	3.02
Spillover	−0.08	0.04	−1.75	Spillover	0.15	0.02	6.48
Case marking:word order	−0.01	0.01	−0.91	Case marking:word order	0.02	0.01	2.12
	adj				**wh+2**		
(Intercept)	−1.17	0.07	−15.96	(Intercept)	−1.98	0.10	−20.65
Case marking	0.00	0.01	0.26	Case marking	−0.02	0.01	−1.37
Word order	−0.02	0.01	−2.38	Word order	0.03	0.01	2.02
Spillover	−0.01	0.02	−0.83	Spillover	0.25	0.02	10.22
Case marking:word order	−0.00	0.01	−0.35	Case marking:word order	0.01	0.01	0.89
	wh+3						
(Intercept)	−2.11	0.10	−21.19				
Case marking	−0.03	0.01	−1.86				
Word order	−0.00	0.02	−0.12				
Spillover	0.06	0.02	3.05				
Case marking:word order	−0.03	0.01	−2.01				

Note that this cannot be explained by a "global spillover effect" from the antecedent: earlier regions did not show the pattern, and there is no reason to assume that antecedents in the OVS/unambiguous condition were extremely difficult to process. Furthermore, an interaction between the experimental factors appeared at position wh+3, albeit in a surprising form: sentences in the OVS/ambiguous condition were read faster than those in the OVS/unambiguous condition, with the two SVO conditions lying in between. We assume that the observed pattern reflects delayed processing of the ellipsis, either as the consequence of subjects "tapping" the space bar at fixed time intervals (Witzel et al., 2012; see Discussion below) or as spillover that was not factored out by the statistical model. As the OVS/ambiguous condition was responsible for the garden-path effect within the antecedent clause, the processing advantage is unexpected with regard to the reconstruction hypothesis, which had predicted the same pattern to reappear at the ellipsis site. The result is also not straightforwardly explained by a pointer-based approach, which would have predicted no differences between the conditions. We will argue below that what we are observing at positions wh+2 and wh+3 is the interaction of two factors: antecedent-ellipsis mismatch and memory trace reactivation through reanalysis.

3.4.1. German Word Order and Antecedent-Ellipsis Mismatch

As we've pointed out in the introduction, German subordinate clauses are required to be verb-final[11] while main clauses invariably have the finite verb in second position. As the sluicing structures in the present study appeared in subordinate clauses, all antecedent clauses would therefore have had to be verb-final instead of verb-second to be compatible with the gap. Given that sluicing is still perfectly acceptable in all of our stimuli, we seem to be seeing a case of "acceptable ungrammaticality" (Frazier, 2008). Both SVO and OVS antecedents were, to use the terminology of Arregui et al. (2006), "flawed," but possibly not in the same way.

OVS order in German main clauses can be derived through topicalization, with the object occupying the so-called *Vorfeld* ("prefield," e.g., Müller, 2005)[12]. As this strategy is not available in subordinate clauses, non-canonical word orders must be derived via *scrambling*, which moves constituents within the so-called *Mittelfeld* ("middle field," e.g., Hinterhölzl, 2006). The slightly simplified examples in (6) illustrate this. SOV order in (6a) is unproblematic, but scrambled OSV in (6b) is, at the very least, highly marked[13].

(6) Die Regierung konnte nicht nachweisen, …
the government could not substantiate

a. **SOV subordinate clause**
… wie die Rebellen einen Sympathisanten
how the rebels a sympathizer.acc
unterstützt hatten.
supported had.pl

[11]The only exception to this rule occurs when the verb takes a sentential complement, which was not the case in our experiment.

[12]The *Feldertheorie* of German sentence structure was first developed by Drach (1937), and is also known as the *Topological Model*.

[13]Apart from not being licensed by information structure, moving the object in (6b) also violates a constraint dictating that definite noun phrases should appear before indefinite ones (see Müller, 1999 for an optimality-theoretic account).

b. **OSV subordinate clause**

… ?? wie einen Sympathisanten$_i$ die Rebellen t$_i$
how a sympathizer.acc the rebels
unterstützt hatten.
supported had.pl

The Recycling Hypothesis proposed by Arregui et al. (2006) predicts that ellipses are more difficult to process the more the antecedent mismatches the ellipsis site. Arregui et al. assume "repair" operations as the source of the difficulty. Assuming the verb-second antecedents have already been partly repaired by moving the verb to the end, (6b) would still need to be transformed into an SOV structure like (6a), presumably be reversing the movement operation. The increased reading times for sentences with object-initial antecedents observed at position wh+2 would be expected under the assumption that the mismatch between an OVS antecedent and an SOV sluice is greater than for SVO antecedents, where the repair process does not need to change the order of the arguments.

Two alternative suggestions made by an anonymous reviewer merit discussion. One is that the processor simply fills the ellipsis site with a verb-second clause, deriving a structure that would have no grammatical surface equivalent. There would be no reason to invoke the Recycling Hypothesis in this case, and the OVS disadvantage would need to be explained either by constraints on topicalization or possibly by invoking working memory factors. Both of these possibilities present problems. It has been found that surprising or unusual stimuli lead to better recall performance (Hirshman et al., 1989), which would lead us to expect that the more uncommon OVS antecedents should be easier instead of more difficult to retrieve. Additionally, the claim that ungrammatical structures can be derived during ellipsis processing seems extreme given that the observed effects can be explained through other means. The reviewer's second suggestion is that garden-pathing in the antecedent might result in its memory representation being more difficult to access, allowing a slower discourse-based mechanism like Murphy's (1985) to dominate during processing. However, seeing that *unambiguous* OVS antecedents also led to longer reading times at position wh+2, this does not seem like a plausible alternative to us.

3.4.2. Antecedent Reactivation through Reanalysis

A reviewer points out that there is some evidence that initial misinterpretations of garden-path sentences persist beyond the point of disambiguation, leading to structural priming, (van Gompel et al., 2006) systematic errors during paraphrasing (Patson et al., 2009) and in comprehension tests (Christianson et al., 2001), as well as competition effects when late-arriving plausibility information contradicts the initial parse (Slattery et al., 2013). One explanation for these effects is that the initial parse of the sentence remains active in memory to some degree even after it has been discarded. In the case of our experiment, if a remnant of the discontinued subject-initial analysis remains behind in the OVS/ambiguous condition, it might be conceivable that this memory trace is considered as a possible antecedent for the ellipsis, possibly blocking access to the "real," reanalyzed antecedent. Research on agreement processing, reflexives and subject-verb dependencies has shown that such memory interference may turn out to make processing easier or more difficult, depending on the phenomenon under study and the exact setup of the experiment (see Engelmann et al., submitted for a review). While the observed speedup in the current study may, in principle, be explained through facilitative interference, the results of Martin and McElree (2009) suggest that the availability of multiple candidate antecedents does not influence the time-course of ellipsis processing in any way. As it is unclear why the interference effect should visible in our experiment but not in theirs, we will present an alternative explanation of our results.

We suggest that the pattern at position wh+3 should be analyzed in terms of a reactivation of the antecedent's memory trace that outweighs the mismatch penalty created by the word order manipulation. As explained in the introduction section, the cue-based retrieval parser of Lewis and Vasishth (2005) incorporates the assumption that syntactic phrases are stored in working memory as chunks. If a chunk is retrieved in order to make an attachment, its activation level increases, which makes subsequent retrievals easier. A reanalysis such as the one required for sentences in the OVS/ambiguous condition should reactivate the antecedent's memory chunk as its structure needs to be changed. Later, at the ellipsis site, it should thus be retrieved faster than the other types of antecedents, to which reanalysis has not applied[14]. The mismatch effect explained above can also be accounted for through an extension of the Lewis and Vasishth (2005) model: If the wh-pronoun sets retrieval cues for a verb-final antecedent in order to match the local clausal configuration, there will be no matching chunk in memory. In order to be able to complete the retrieval, the processor may then attempt to retrieve chunks which do not match the cues perfectly, such as the main clauses in the current study. Due to the matching relative order of subject and object, an SVO chunk may resonate more strongly with the SOV cue than one with OVS word order, as schematized in (7).

(7) a. **OVS antecedent, resonates weakly with SOV cue (O-S $\neq$ SO)**

[Einen Sympathisanten hatten die Rebellen
A sympathizer had.pl the rebels
unterstützt]$_{OVS}$ …
supported

b. **SVO antecedent, resonates more strongly with SOV cue (S-O $\sim$ SO)**

[Ein Sympathisant hatte die Rebellen
A sympathizer had.pl the rebels
unterstützt]$_{SVO}$ …
supported

[14]This presupposes that trace decay has not reduced the activation of the antecedent to zero in any case by the time the ellipsis is processed. The model of Lewis and Vasishth (2005) assumes that the activation of chunks than have been reaccessed is higher even after complete decay.

***wie* in subordinate clause sets SOV cue**

… aber die Regierung konnte nicht nachweisen,
but the government could not substantiate
wie [ ]$_{SOV}$ …
how

A lower retrieval latency would then be expected for SVO chunks, thereby predicting the observed OVS disadvantage at position wh+2[15]. The reactivation/mismatch approach is thus able to account for the observed pattern of results, but due its status as a *post-hoc* argument is in need of further empirical validation.

One might think of yet another explanation for the result, namely that reconstruction *is* taking place and that syntactic priming is responsible for the advantage in the OVS/ambiguous condition. However, such an approach would not fit with the fact that the antecedent's structure is, strictly speaking, incompatible with the word order required at the gap: As the derivations of main and subordinate clauses involve different steps, it is not obvious what exactly would be primed. One would have to make a very specific set of assumptions: First, the parser would need to blindly reconstruct the syntax of the antecedent at the ellipsis site *before* checking for possible mismatches, similarly to the anonymous reviewer's suggestion that was discussed earlier. Secondly, garden-path sentences would need to prime their final structure more strongly than unambiguous controls, which to our knowledge has not been demonstrated to date. Ambiguous/OVS antecedents would then initially gain an advantage through increased priming while both kinds of OVS antecedents would be disadvantaged during the mismatch checking phase.

3.4.3. Sluicing and Predictive Processing

We believe that one additional result is worth mentioning, even though it was only arrived at *post-hoc*. It fits with the proposal by Yoshida et al. (2013) that predictive processing may be involved in the interpretation of sluicing structures. Yoshida et al. compared sentences in which it was either possible or impossible to analyze a specific wh-phrase as part of a sluice. The evidence suggested that as soon as the wh-phrase in question was encountered, the parser started building a sluicing structure, presumably because it is preferred over other possible continuations.

We took the implication of predictive processing as an incentive to analyze reading times for the region directly *preceding* the wh-pronoun in our own experiment: If sluicing is the preferred continuation after a wh-pronoun has been encountered, it is not unlikely that it will also rank fairly highly before that point. This is especially likely given that subordinate clauses in German require a comma, which was thus present in the pre-wh region in all of our stimuli, excluding a vast range of alternative continuations that would have been likely in Yoshida et al.'s materials.

The fitting of a linear mixed-effects model (see above) at position wh-1 revealed a significant interaction between word order and case marking ($\hat{\beta} = -0.03$, $se = 0.01$, $t = -2.3$) which had the same sign as the one observed at position wh+3[16]. **Table 3** shows the model output. However, unlike at the later position, nested contrasts showed that the interaction was driven by the OVS/unambiguous condition being read more slowly than the SVO/unambiguous condition ($\hat{\beta} = 0.04$, $se = 0.02$, $t = 2.24$), even though the numerical pattern in raw reading times was the same as for position wh+3. We have no ready explanation for this finding. Speculatively, a heuristic may be used to estimate the fit between the sluice and the antecedent. Such a heuristic might work better when case is overtly marked, and might operate more quickly when word order is canonical. In our opinion this kind of predictive strategy makes it unlikely that processing proceeds according to the priming-based account described above, in which local constraints do not influence the initial structure assignment for the ellipsis.

TABLE 3 | Coefficient estimates, standard errors and *t*-values for the linear mixed-effects model fit to reciprocal reading times at region wh-1.

	wh-1		
	Estimate	**Std. Error**	***t*-value**
(Intercept)	−2.10	0.06	−35.17
Case marking	−0.02	0.01	−1.60
Word order	0.01	0.01	0.93
Spillover	0.36	0.02	15.21
Case marking: word order	−0.03	0.01	−2.30

To further investigate the notion that a sluice was the expected structure in our materials, we ran a sentence completion study with thirty-five new participants. It has been suggested that the speech production system may be responsible for generating linguistic expectations in comprehension (Pickering and Garrod, 2007). As sentence continuation preferences have been shown to be predictive of processing difficulty in self-paced reading (Smith and Levy, 2011), we assume that a preference for sluicing continuations in our reading study should translate into a corresponding preference in sentence completions. The stimuli consisted of the 32 sentences used in the current reading study, along with 32 sentences from a different experiment and 96 fillers. Sentences were presented using a modified version of Linger's masked auto-paced reading (otherwise known as rapid serial visual presentation or RSVP). The stimuli from the current study were cut off right before the ellipsis site and participants were asked to complete the sentences using the first continuation that came to mind. Due to the nature of the presentation, participants could not reread the sentences while they were typing their continuation. Results showed a total of only five per cent sluicing continuations. Another 54% of continuations were non-sluiced wh-clauses, followed by *if*-clauses at seventeen per cent and *that*-clauses at seven per cent. Assuming that this pattern is not due to idiosyncrasies of the production system, the observed outcome casts some doubt on the assumption that a sluicing continuation was, in fact, highly expected in

[15]In order to derive grammatical structures, repair processes that change the word order to verb-final would still need to apply after retrieval.

[16]As a sanity check, we also analyzed reading times at position wh-2, finding no significant effects.

our stimuli. However, subjects in the production experiment could choose their preferred continuation freely, which may conceivably have led to more conscious deliberation on their part. It is entirely possible that sluicing is only one of several possible continuations which are pre-activated during reading, which might be enough to explain the findings of Yoshida et al. (2013) and the interaction we observed at position wh-1 in the self-paced reading study. Despite the limited scope of the production experiment, given the earlier findings by (Smith and Levy, 2011), we feel that it was important to investigate whether the predictive processing seen in comprehension maps directly onto language users' preferences in production. This is apparently not the case under the conditions tested here.

4. GENERAL DISCUSSION

The current experiment investigated the processing of a sluicing construction in cases where the antecedent is a garden-path structure, in this instance a clause with a subject/object ambiguity. We observed reduced reading times for sentences with garden-path antecedents three regions downstream from the ellipsis as well as directly before the ellipsis. Furthermore, there was an overall pattern of elevated reading times in the spillover region for antecedents that mismatched the canonical word order of the ellipsis site. Our results are best compatible with accounts of ellipsis resolution that can be implemented in the form of a memory pointer mechanism (Frazier and Clifton, 2001, 2005; Martin and McElree, 2008), which would need to be augmented to account for reactivation assumed by the cue-based retrieval parser of Lewis and Vasishth (2005). The evidence for a mismatch effect is in line with the predictions of the Recycling Hypothesis proposed by Arregui et al. (2006). However, given that we have observed no evidence for reconstruction in our experiment, we do not subscribe to Arregui et al.'s assumption that "flawed" antecedents are "repaired" in a way that is similar to syntactic reanalysis (p. 242). The mismatch effect may be better approached along the lines of the wh-pronoun setting a retrieval cue for an antecedent that matches the word order requirements of the local clause, opting for the closest candidate upon failure. Alternatively, one could follow the proposal of Kim et al. (2011), in which ellipses with non-canonical antecedents violate parsing heuristics that are based on construction frequency and expectation. Under an approach without reconstruction, we would claim that it is not a parsing heuristic that is violated, but a local expectation as to what an antecedent targeted by retrieval should look like. If the expectation were global, no mismatch effect would be expected, given that the antecedent has already been encountered in the input. The local expectation account fits with the pattern observed by Yoshida et al. (2013) as well as with the effect found in the pre-sluice region (wh-1) in the current study.

Still, why did we observe a pattern in which the experimental manipulation seemed to have an effect before and after, but not *at* the ellipsis site? We assume that this is due to either insufficient statistical power, to our subjects' reading strategies, or both. Power is always an issue when effect sizes are as small as in the current study: the mean reading time difference between the unambiguous/OVS and the ambiguous/OVS conditions at position wh+3 was only 30 ms. Given this value and the associated standard errors, the *post-hoc* power to detect a real effect was at 45%, which is comparable to Frazier and Clifton's (2000) study, where the computation yields 43% *post-hoc* power[17]. The bottom line is that sample size needs to be significantly increased in order to convincingly argue that there really is *no* effect of the manipulation, even though this might be construed as trying to "force significance."

The concern related to reading strategies comes from the fact that while non-cumulative self-paced reading more closely resembles data from natural reading than the cumulative variant does (Just et al., 1982), it is by no means certain that subjects will not adopt a "wait and see" strategy at least on some trials, meaning that they will press the button at a fixed rate and only then start processing. Witzel et al. (2012), suspecting such rhythmic "tapping" in their data, tried to remove its influence by calculating the standard deviation of the response time by subject and excluding the participants with the smallest variability, which did, however, not change their statistical result. The authors conclude that either 'tapping' was not a factor in their data or their method was not suitable to account for it, leaving the issue for future research. We will do the same here.

There is also a slightly different explanation for the delay we observed, namely that subjects *did* process the words the words as they were revealed, but postponed the processing of the ellipsis until they had more information. Such a strategy might make sense considering that an embedded question (i.e., an interrogative clause that serves as a complement, as in *. . . , but the government could not substantiate how, . . .*) in itself usually imparts no relevant information apart from the fact that some piece of information is missing. As the contents of the spillover region put this information in context (*. . . , because/so that/even though/until . . .*), the relevance may have become apparent, causing the observed processing pattern.

A final objection to our study would be that there was no control condition without ellipsis. It should be noted that it is extremely difficult to create closely matched controls for our sentences, given that possible continuations are limited to complement clauses, which usually feature more than one word. Other studies on ellipsis processing also lack controls [e.g., Frazier and Clifton, 2000, 2005 (except Experiments 2 and 3), Poirier et al., 2010], leaving open the possibility that any observed effects do not actually stem from the antecedent being recovered due to a perceived gap in the sentence but from some other mechanism. While this criticism can be met by pointing to the localization of the effects, as well as to the unavailability of a plausible alternative explanation, it would be desirable to include controls in future studies to strengthen the conclusions drawn from the data.

Further investigations into the interaction between antecedent ambiguity and ellipsis processing are already underway in our laboratory. We are currently aiming to find further evidence for the reactivation effect using different kinds of temporary

[17]Note that this is not the *true* power of the experiments, which depends on the unknown true effect size.

ambiguities and ellipses, as well as experimental procedures other than self-paced reading (e.g., eye tracking).

FUNDING

This research was funded by the University of Potsdam.

ACKNOWLEDGMENTS

To author wishes to thank Shravan Vasishth, Lena A. Jäger, Barbara Hemforth, the Vasishth Lab team, and the audience at CUNY 2015 for helpful comments and suggestions, as well as Johanna Thieke for assistance with data collection.

REFERENCES

Arai, M., Nakamura, C., and Mazuka, R. (2014). Predicting the unbeaten path through syntactic priming. *J. Exp. Psychol. Learn. Mem. Cogn.* 41, 482–500. doi: 10.1037/a0038389

Arregui, A., Clifton, C. Jr., Frazier, L., and Moulton, K. (2006). Processing elided verb phrases with flawed antecedents: the recycling hypothesis. *J. Mem. Lang.* 55, 232–246. doi: 10.1016/j.jml.2006.02.005

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2014). *lme4: Linear Mixed-Effects Models Using Eigen and S4*. R package version 1.1-7.

Bever, T. G. (1970). "The cognitive basis for linguistic structures," in *Cognition and the Development of Language*, ed J. R. Hayes (New York, NY: Wiley), 279–362.

Box, G. E., and Cox, D. R. (1964). An analysis of transformations. *J. R. Stat. Soc. Ser. B Methodol.* 26, 211–252.

Branigan, H. P., Pickering, M. J., and McLean, J. F. (2005). Priming prepositional-phrase attachment during comprehension. *J. Exp. Psychol. Learn. Mem. Cogn.* 31, 468–481. doi: 10.1037/0278-7393.31.3.468

Cai, Z. G., Pickering, M. J., and Sturt, P. (2013). Processing verb-phrase ellipsis in Mandarin Chinese: evidence against the syntactic account. *Lang. Cogn. Process.* 28, 810–828. doi: 10.1080/01690965.2012.665932

Christianson, K., Hollingworth, A., Halliwell, J. F., and Ferreira, F. (2001). Thematic roles assigned along the garden path linger. *Cogn. Psychol.* 42, 368–407. doi: 10.1006/cogp.2001.0752

Chung, S., Ladusaw, W. A., and McCloskey, J. (1995). Sluicing and logical form. *Nat. Lang. Semant.* 3, 239–282.

Dickey, M. W., and Bunger, A. C. (2011). Comprehension of elided structure: evidence from sluicing. *Lang. Cogn. Process.* 26, 63–78. doi: 10.1080/01690961003691074

Drach, E. (1937). *Grundgedanken der Deutschen Satzlehre*. Frankfurt/Main: Diesterweg.

Dubey, A., Keller, F., and Sturt, P. (2008). A probabilistic corpus-based model of syntactic parallelism. *Cognition* 109, 326–344. doi: 10.1016/j.cognition.2008.09.006

Fishbein, J., and Harris, J. A. (2014). Making sense of Kafka: structural biases induce early sense commitment for metonyms. *J. Mem. Lang.* 76, 94–112. doi: 10.1016/j.jml.2014.06.005

Foertsch, J., and Gernsbacher, M. A. (1994). In search of complete comprehension: getting minimalistsİ to work. *Discourse Process.* 18, 271–296.

Frazier, L. (2008). "Processing ellipsis: a processing solution to the undergeneration problem," in *Proceedings of the 26th West Coast Conference on Formal Linguistics*, eds C. B. Chang and H. J. Haynie (Somerville, MA: Cascadilla Press), 21–32.

Frazier, L., and Clifton, C. Jr. (2000). On bound variable interpretations: the LF-only hypothesis. *J. Psycholinguist. Res.* 29, 125–140. doi: 10.1023/A:1005136826534

Frazier, L., and Clifton, Jr. C. (2001). Parsing coordinates and ellipsis: copy α. *Syntax* 4, 1–22. doi: 10.1111/1467-9612.00034

Frazier, L., and Clifton, Jr. C. (2005). The syntax-discourse divide: processing ellipsis. *Syntax* 8, 121–174. doi: 10.1111/j.1467-9612.2005.00077.x

Gorrell, P. (1996). Parsing theory and phrase-order variation in German V2 clauses. *J. Psycholinguist. Res.* 25, 135–156.

Grinder, J., and Postal, P. M. (1971). Missing antecedents. *Linguist. Inq.* 2, 269–312.

Grodzinsky, Y. (1986). Language deficits and the theory of syntax. *Brain Lang.* 27, 135–159.

Hardt, D. (1993). *Verb Phrase Ellipis: Form, Meaning, and Processing*. Ph.D. thesis, University of Pennsylvania, Philadelphia.

Hartigan, J. A., and Hartigan, P. (1985). The dip test of unimodality. *Ann. Stat.* 13, 70–84.

Hemforth, B. (1993). *Kognitives Parsing: Repräsentation und Verarbeitung sprachlichen Wissens*. Sankt Augustin: Infix.

Hinterhölzl, R. (2006). *Scrambling, Remnant Movement, and Restructuring in West Germanic: Oxford Studies in Comparative Syntax*. Oxford: Oxford University Press.

Hirshman, E., Whelley, M. M., and Palij, M. (1989). An investigation of paradoxical memory effects. *J. Mem. Lang.* 28, 594–609.

Just, M. A., Carpenter, P. A., and Woolley, J. D. (1982). Paradigms and processes in reading comprehension. *J. Exp. Psychol. Gen.* 111, 228–238.

Kehler, A. (2000). Coherence and the resolution of ellipsis. *Linguist. Philos.* 23, 533–575. doi: 10.1023/A:1005677819813

Kertz, L. (2000). *Ellipsis Reconsidered*. Ph.D. thesis, University of San Diego.

Kim, C. S., Kobele, G. M., Runner, J. T., and Hale, J. T. (2011). The acceptability cline in VP ellipsis. *Syntax* 14, 318–354. doi: 10.1111/j.1467-9612.2011.00160.x

Lewis, R. L., and Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cogn. Sci.* 29, 375–419. doi: 10.1207/s15516709cog0000_25

Martin, A. E., and McElree, B. (2008). A content-addressable pointer mechanism underlies comprehension of verb-phrase ellipsis. *J. Mem. Lang.* 58, 879–906. doi: 10.1016/j.jml.2007.06.010

Martin, A. E., and McElree, B. (2009). Memory operations that support language comprehension: Evidence from verb-phrase ellipsis. *J. Exp. Psychol. Learn. Mem. Cogn.* 35, 1231. doi: 10.1037/a0016271

Meng, M., and Bader, M. (2000). Mode of disambiguation and garden-path strength: an investigation of subject-object ambiguities in German. *Lang. Speech* 43, 43–74. doi: 10.1177/00238309000430010201

Merchant, J. (2001). *The Syntax of Silence: Sluicing, Islands, and the Theory of Ellipsis*. Oxford: Oxford University Press.

Merchant, J. (2013). Voice and ellipsis. *Linguist. Inq.* 44, 77–108. doi: 10.1162/LING_a_00120

Mitchell, D. C., Cuetos, F., Corley, M. M., and Brysbaert, M. (1995). Exposure-based models of human parsing: evidence for the use of coarse-grained (nonlexical) statistical records. *J. Psycholinguist. Res.* 24, 469–488.

Müller, G. (1999). Optimality, markedness, and word order in German. *Linguistics* 37, 777–818.

Müller, S. (2005). Zur analyse der deutschen satzstruktur. *Linguist. Berichte* 201, 3–39. Available online at: https://hpsg.fu-berlin.de/~stefan/Pub/satz-lb.html

Murguia, E. (2004). *Syntactic Identity and Locality Restrictions on Verbal Ellipsis*. Ph.D. thesis, University of Maryland.

Murphy, G. L. (1985). Processes of understanding anaphora. *J. Mem. Lang.* 24, 290–303.

Murphy, G. L., and Shapiro, A. M. (1994). Forgetting of verbatim information in discourse. *Mem. Cogn.* 22, 85–94.

Patson, N. D., Darowski, E. S., Moon, N., and Ferreira, F. (2009). Lingering misinterpretations in garden-path sentences: evidence from a paraphrasing task. *J. Exp. Psychol. Learn. Mem. Cogn.* 35, 280. doi: 10.1037/a0014276

Phillips, C., and Parker, D. (2014). The psycholinguistics of ellipsis. *Lingua* 151, 78–95. doi: 10.1016/j.lingua.2013.10.003

Pickering, M. J., and Garrod, S. (2007). Do people use language production to make predictions during comprehension? *Trends Cogn. Sci.* 11, 105–110. doi: 10.1016/j.tics.2006.12.002

Poirier, J., Wolfinger, K., Spellman, L., and Shapiro, L. P. (2010). The real-time processing of sluiced sentences. *J. Psycholinguist. Res.* 39, 411–427. doi: 10.1007/s10936-010-9148-9

Potsdam, E. (2007). Malagasy sluicing and its consequences for the identity requirement on ellipsis. *Nat. Lang. Linguist. Theory* 25, 577–613. doi: 10.1007/s11049-006-9015-4

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rohde, D. (2003). *Linger, Version 2.94.*

Ross, J. R. (1967). *Constraints on Variables in Syntax*. Ph.D. thesis, Massachusetts Institute of Technology.

Ross, J. R. (1969). "Guess who?," in *Papers from the 5th Regional Meeting of the Chicago Linguistic Society*, eds R. Binnick, A. Davison, G. Green, and J. Morgan (Chicago, IL: Chicago Linguistic Society), 252–286.

Sachs, J. S. (1967). Recognition memory for syntactic and semantic aspects of connected discourse. *Percept. Psychophys.* 2, 437–442.

Schlesewsky, M., Fanselow, G., Kliegl, R., and Krems, J. (2000). "The subject preference in the processing of locally ambiguous wh-questions in German," in *German Sentence Processing*, eds B. Hemforth and L. Konieczny (Dordrecht: Kluwer), 65–93.

Schwartz, B., and Vikner, S. (2007). "The verb always leaves IP in V2 clauses," in *Comparative Grammar - Critical Concepts in Linguistics, Vol. 5*, ed I. Roberts (London: Routledge), 244–302.

Simner, J., and Smyth, R. (1999). Phonological activation in Anaphoric Lexical Access (ALA). *Brain Lang.* 68, 40–45.

Slattery, T. J., Sturt, P., Christianson, K., Yoshida, M., and Ferreira, F. (2013). Lingering misinterpretations of garden path sentences arise from competing syntactic representations. *J. Mem. Lang.* 69, 104–120. doi: 10.1016/j.jml.2013.04.001

Smith, N. J., and Levy, R. (2011). "Cloze but no cigar: the complex relationship between cloze, corpus, and subjective probabilities in language processing," in *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, 1637–1642.

Swets, B., Desmet, T., Clifton, C., and Ferreira, F. (2008). Underspecification of syntactic ambiguities: evidence from self-paced reading. *Mem. Cogn.* 36, 201–216. doi: 10.3758/MC.36.1.201

Tanenhaus, M. K., and Carlson, G. N. (1990). Comprehension of deep and surface verbphrase anaphors. *Lang. Cogn. Process.* 5, 257–280.

van Gompel, R. P., Pickering, M. J., Pearson, J., and Jacob, G. (2006). The activation of inappropriate analyses in garden-path sentences: Evidence from structural priming. *J. Mem. Lang.* 55, 335–362. doi: 10.1016/j.jml.2006.06.004

Venables, W. N., and Ripley, B. D. (2002). *Modern Applied Statistics with S*. New York, NY: Springer.

Wasow, T. (1972). *Anaphoric Relations in English*. Ph.D. thesis, Massachusetts Institute of Technology.

Williams, E. S. (1977). Discourse and logical form. *Linguist. Inq.* 8, 101–139.

Witzel, N., Witzel, J., and Forster, K. (2012). Comparisons of online reading paradigms: eye tracking, moving-window, and maze. *J. Psycholinguistic Res.* 41, 105–128. doi: 10.1007/s10936-011-9179-x

Xiang, M., Grove, J., and Merchant, J. (2014). *Ellipsis Sites Induce Structural Priming Effects*. Manuscript, University of Chicago.

Yoshida, M., Dickey, M. W., and Sturt, P. (2013). Predictive processing of syntactic structure: sluicing and ellipsis in real-time sentence processing. *Lang. Cogn. Process.* 28, 272–302. doi: 10.1080/01690965.2011.622905

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2016 Paape. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

APPENDIX - EXPERIMENTAL MATERIALS

Eine Vertreterin der Gewerkschaft * hatten * die anwesenden Minister * während der Sitzung * scharf attackiert, * aber * der gesprächige Parlamentarier * wusste * selbst * nicht, * warum, * denn * er * war * nicht * dabei gewesen.

Eine Vertreterin der Gewerkschaft * hatte * die anwesenden Minister * während der Sitzung * scharf attackiert, * aber * der gesprächige Parlamentarier * wusste * selbst * nicht, * warum, * denn * er * war * nicht * dabei gewesen.

Einen Vertreter der Gewerkschaft * hatten * die anwesenden Minister * während der Sitzung * scharf attackiert, * aber * der gesprächige Parlamentarier * wusste * selbst * nicht, * warum, * denn * er * war * nicht * dabei gewesen.

Ein Vertreter der Gewerkschaft * hatte * die anwesenden Minister * während der Sitzung * scharf attackiert, * aber * der gesprächige Parlamentarier * wusste * selbst * nicht, * warum, * denn * er * war * nicht * dabei gewesen. Eine Vertraute des Bürgermeisters * hatten * die Ratsmitglieder * kurz vor der Wahl * auffallend häufig angerufen, * aber * heute * weiß * niemand * mehr, * warum, * wie * eine Zeitung * kürzlich * in einem Kommentar * schrieb.

Eine Vertraute des Bürgermeisters * hatte * die Ratsmitglieder * kurz vor der Wahl * auffallend häufig angerufen, * aber * heute * weiß * niemand * mehr, * warum, * wie * eine Zeitung * kürzlich * in einem Kommentar * schrieb. Einen Vertrauten des Bürgermeisters * hatten * die Ratsmitglieder * kurz vor der Wahl * auffallend häufig angerufen, * aber * heute * weiß * niemand * mehr, * warum, * wie * eine Zeitung * kürzlich * in einem Kommentar * schrieb. Ein Vertrauter des Bürgermeisters * hatte * die Ratsmitglieder * kurz vor der Wahl * auffallend häufig angerufen, * aber * heute * weiß * niemand * mehr, * warum, * wie * eine Zeitung * kürzlich * in einem Kommentar * schrieb.

Eine Kellnerin des Lokals * hatten * die Stammgäste * über das geplante Skatturnier * ausgefragt, * aber * der Wirt * konnte * nicht * sagen, * warum, * da * er * offenbar * an jenem Abend * sehr beschäftigt gewesen war.

Eine Kellnerin des Lokals * hatte * die Stammgäste * über das geplante Skatturnier * ausgefragt, * aber * der Wirt * konnte * nicht * sagen, * warum, * da * er * offenbar * an jenem Abend * sehr beschäftigt gewesen war.

Einen Kellner des Lokals * hatten * die Stammgäste * über das geplante Skatturnier * ausgefragt, * aber * der Wirt * konnte * nicht * sagen, * warum, * da * er * offenbar * an jenem Abend * sehr beschäftigt gewesen war. Ein Kellner des Lokals * hatte * die Stammgäste * über das geplante Skatturnier * ausgefragt, * aber * der Wirt * konnte * nicht * sagen, * warum, * da * er * offenbar * an jenem Abend * sehr beschäftigt gewesen war.

Eine Beraterin des Präsidenten * hatten * die Ermittler * offensichtlich * mit Erfolg getäuscht, * aber * man * fand * nie * heraus, * wie, * denn * es * galt * nach wie vor * die höchste Geheimhaltungsstufe.

Eine Beraterin des Präsidenten * hatte * die Ermittler * offensichtlich * mit Erfolg getäuscht, * aber * man * fand * nie * heraus, * wie, * denn * es * galt * nach wie vor * die höchste Geheimhaltungsstufe.

Einen Berater des Präsidenten * hatten * die Ermittler * offensichtlich * mit Erfolg getäuscht, * aber * man * fand * nie * heraus, * wie, * denn * es * galt * nach wie vor * die höchste Geheimhaltungsstufe.

Ein Berater des Präsidenten * hatte * die Ermittler * offensichtlich * mit Erfolg getäuscht, * aber * man * fand * nie * heraus, * wie, * denn * es * galt * nach wie vor * die höchste Geheimhaltungsstufe.

Eine Sprecherin des Pharmakonzerns * hatten * die Sportler * nach Angaben der Presse * persönlich getroffen, * aber * die Quelle * konnte * nicht * mitteilen, * wo, * sodass * die Geschichte * den meisten Lesern * wahrscheinlich * nicht sehr glaubwürdig erschien.

Eine Sprecherin des Pharmakonzerns * hatte * die Sportler * nach Angaben der Presse * persönlich getroffen, * aber * die Quelle * konnte * nicht * mitteilen, * wo, * sodass * die Geschichte * den meisten Lesern * wahrscheinlich * nicht sehr glaubwürdig erschien.

Einen Sprecher des Pharmakonzerns * hatten * die Sportler * nach Angaben der Presse * persönlich getroffen, * aber * die Quelle * konnte * nicht * mitteilen, * wo, * sodass * die Geschichte * den meisten Lesern * wahrscheinlich * nicht sehr glaubwürdig erschien.

Ein Sprecher des Pharmakonzerns * hatte * die Sportler * nach Angaben der Presse * persönlich getroffen, * aber * die Quelle * konnte * nicht * mitteilen, * wo, * sodass * die Geschichte * den meisten Lesern * wahrscheinlich * nicht sehr glaubwürdig erschien.

Eine Sympathisantin der Opposition * hatten * die Rebellen * laut einem Bericht * maßgeblich unterstützt, * aber * die Regierung * konnte * nicht * nachweisen, * wie, * so sehr * sich * die Untersuchungskommission * auch * bemühte.

Eine Sympathisantin der Opposition * hatte * die Rebellen * laut einem Bericht * maßgeblich unterstützt, * aber * die Regierung * konnte * nicht * nachweisen, * wie, * so sehr * sich * die Untersuchungskommission * auch * bemühte.

Einen Sympathisanten der Opposition * hatten * die Rebellen * laut einem Bericht * maßgeblich unterstützt, * aber * die Regierung * konnte * nicht * nachweisen, * wie, * so sehr * sich * die Untersuchungskommission * auch * bemühte.

Ein Sympathisant der Opposition * hatte * die Rebellen * laut einem Bericht * maßgeblich unterstützt, * aber * die Regierung * konnte * nicht * nachweisen, * wie, * so sehr * sich * die Untersuchungskommission * auch * bemühte.

Eine Gönnerin des Künstlers * hatten * die etwas seltsamen Verwandten * zu Anfang * des Mordes verdächtigt, * aber * aus den Tagebüchern * geht * nicht * hervor, * warum, * zumal * es * sich * relativ eindeutig * um Suizid handelte.

Eine Gönnerin des Künstlers * hatte * die etwas seltsamen Verwandten * zu Anfang * des Mordes verdächtigt, * aber * aus den Tagebüchern * geht * nicht * hervor, * warum, * zumal * es * sich * relativ eindeutig * um Suizid handelte.

Einen Gönner des Künstlers * hatten * die etwas seltsamen Verwandten * zu Anfang * des Mordes verdächtigt, * aber * aus den Tagebüchern * geht * nicht * hervor, * warum, * zumal * es * sich * relativ eindeutig * um Suizid handelte.

Ein Gönner des Künstlers * hatte * die etwas seltsamen Verwandten * zu Anfang * des Mordes verdächtigt, * aber * aus

den Tagebüchern * geht * nicht * hervor, * warum, * zumal * es * sich * relativ eindeutig * um Suizid handelte.
Eine Schülerin des Schachmeisters * hatten * die Schiedsrichter * während des Turniers * sehr genau beobachtet, * aber * der aufmerksame Zuschauer * fragte * sich * noch immer, * warum, * als * er * am Abend * endlich * nach Hause kam.
Eine Schülerin des Schachmeisters * hatte * die Schiedsrichter * während des Turniers * sehr genau beobachtet, * aber * der aufmerksame Zuschauer * fragte * sich * noch immer, * warum, * als * er * am Abend * endlich * nach Hause kam.
Einen Schüler des Schachmeisters * hatten * die Schiedsrichter * während des Turniers * sehr genau beobachtet, * aber * der aufmerksame Zuschauer * fragte * sich * noch immer, * warum, * als * er * am Abend * endlich * nach Hause kam.
Ein Schüler des Schachmeisters * hatte * die Schiedsrichter * während des Turniers * sehr genau beobachtet, * aber * der aufmerksame Zuschauer * fragte * sich * noch immer, * warum, * als * er * am Abend * endlich * nach Hause kam.
Eine Spielerin des Vereins * hatten * die aufdringlichen Fans * nach dem Auswärtsspiel * grob beleidigt, * aber * der Trainer * konnte * nicht * verstehen, * warum, * sodass * er * nur * enttäuscht * den Kopf schüttelte.
Eine Spielerin des Vereins * hatte * die aufdringlichen Fans * nach dem Auswärtsspiel * grob beleidigt, * aber * der Trainer * konnte * nicht * verstehen, * warum, * sodass * er * nur * enttäuscht * den Kopf schüttelte.
Einen Spieler des Vereins * hatten * die aufdringlichen Fans * nach dem Auswärtsspiel * grob beleidigt, * aber * der Trainer * konnte * nicht * verstehen, * warum, * sodass * er * nur * enttäuscht * den Kopf schüttelte.
Ein Spieler des Vereins * hatte * die aufdringlichen Fans * nach dem Auswärtsspiel * grob beleidigt, * aber * der Trainer * konnte * nicht * verstehen, * warum, * sodass * er * nur * enttäuscht * den Kopf schüttelte.
Eine Geschworene des Gerichts * hatten * die beiden Angeklagten * trotz richterlicher Verwarnung * direkt angesprochen, * aber * niemand im Saal * verstand * wohl * so recht, * weshalb, * bevor * die Verhandlung * überraschend * auf unbestimmte Zeit * vertagt wurde.
Eine Geschworene des Gerichts * hatte * die beiden Angeklagten * trotz richterlicher Verwarnung * direkt angesprochen, * aber * niemand im Saal * verstand * wohl * so recht, * weshalb, * bevor * die Verhandlung * überraschend * auf unbestimmte Zeit * vertagt wurde.
Einen Geschworenen des Gerichts * hatten * die beiden Angeklagten * trotz richterlicher Verwarnung * direkt angesprochen, * aber * niemand im Saal * verstand * wohl * so recht, * weshalb, * bevor * die Verhandlung * überraschend * auf unbestimmte Zeit * vertagt wurde.
Ein Geschworener des Gerichts * hatte * die beiden Angeklagten * trotz richterlicher Verwarnung * direkt angesprochen, * aber * niemand im Saal * verstand * wohl * so recht, * weshalb, * bevor * die Verhandlung * überraschend * auf unbestimmte Zeit * vertagt wurde.
Eine Mitarbeiterin der maroden Firma * hatten * die Geschäftsführer * in das raffinierte Veruntreuungssystem * eingeweiht, * aber * es * herrscht * Uneinigkeit * darüber, * wann, * denn * von den belastenden Dokumenten * trägt * keines * ein Datum.
Eine Mitarbeiterin der maroden Firma * hatte * die Geschäftsführer * in das raffinierte Veruntreuungssystem * eingeweiht, * aber * es * herrscht * Uneinigkeit * darüber, * wann, * denn * von den belastenden Dokumenten * trägt * keines * ein Datum.
Einen Mitarbeiter der maroden Firma * hatten * die Geschäftsführer * in das raffinierte Veruntreuungssystem * eingeweiht, * aber * es * herrscht * Uneinigkeit * darüber, * wann, * denn * von den belastenden Dokumenten * trägt * keines * ein Datum.
Ein Mitarbeiter der maroden Firma * hatte * die Geschäftsführer * in das raffinierte Veruntreuungssystem * eingeweiht, * aber * es * herrscht * Uneinigkeit * darüber, * wann, * denn * von den belastenden Dokumenten * trägt * keines * ein Datum.
Eine Aufseherin des Gefängnisses * hatten * die verdächtigen Häftlinge * durch ein erfundenes Alibi * gedeckt, * aber * keinem der Beteiligten * war * damals * zu entlocken, * wieso, * denn * eine Aussage * hätte * wohl * gegen die Ehre verstoßen.
Eine Aufseherin des Gefängnisses * hatte * die verdächtigen Häftlinge * durch ein erfundenes Alibi * gedeckt, * aber * keinem der Beteiligten * war * damals * zu entlocken, * wieso, * denn * eine Aussage * hätte * wohl * gegen die Ehre verstoßen.
Einen Aufseher des Gefängnisses * hatten * die verdächtigen Häftlinge * durch ein erfundenes Alibi * gedeckt, * aber * keinem der Beteiligten * war * damals * zu entlocken, * wieso, * denn * eine Aussage * hätte * wohl * gegen die Ehre verstoßen.
Ein Aufseher des Gefängnisses * hatte * die verdächtigen Häftlinge * durch ein erfundenes Alibi * gedeckt, * aber * keinem der Beteiligten * war * damals * zu entlocken, * wieso, * denn * eine Aussage * hätte * wohl * gegen die Ehre verstoßen.
Eine Angestellte des städtischen Verkehrsunternehmens * hatten * die Fahrgäste * mit unverschämten äußerungen * belästigt, * aber * das Team von Soziologen * konnte * nicht * erklären, * wieso, * sodass * der Zwischenfall * für die Wissenschaft * bis heute * rätselhaft bleibt.
Eine Angestellte des städtischen Verkehrsunternehmens * hatte * die Fahrgäste * mit unverschämten äußerungen * belästigt, * aber * das Team von Soziologen * konnte * nicht * erklären, * wieso, * sodass * der Zwischenfall * für die Wissenschaft * bis heute * rätselhaft bleibt.
Einen Angestellten des städtischen Verkehrsunternehmens * hatten * die Fahrgäste * mit unverschämten äußerungen * belästigt, * aber * das Team von Soziologen * konnte * nicht * erklären, * wieso, * sodass * der Zwischenfall * für die Wissenschaft * bis heute * rätselhaft bleibt.
Ein Angestellter des städtischen Verkehrsunternehmens * hatte * die Fahrgäste * mit unverschämten äußerungen * belästigt, * aber * das Team von Soziologen * konnte * nicht * erklären, * wieso, * sodass * der Zwischenfall * für die Wissenschaft * bis heute * rätselhaft bleibt.
Eine Dolmetscherin des Botschafters * hatten * die Gastgeber * während der Begrüßungszeremonie * empfindlich gekränkt, * aber * damals * konnte * niemand * nachvollziehen, * womit, * obwohl * die kulturellen Gepflogenheiten * der jeweils anderen Seite * auf jeden Fall * hinreichend bekannt waren.

Eine Dolmetscherin des Botschafters * hatte * die Gastgeber * während der Begrüßungszeremonie * empfindlich gekränkt, * aber * damals * konnte * niemand * nachvollziehen, * womit, * obwohl * die kulturellen Gepflogenheiten * der jeweils anderen Seite * auf jeden Fall * hinreichend bekannt waren.
Einen Dolmetscher des Botschafters * hatten * die Gastgeber * während der Begrüßungszeremonie * empfindlich gekränkt, * aber * damals * konnte * niemand * nachvollziehen, * womit, * obwohl * die kulturellen Gepflogenheiten * der jeweils anderen Seite * auf jeden Fall * hinreichend bekannt waren.
Ein Dolmetscher des Botschafters * hatte * die Gastgeber * während der Begrüßungszeremonie * empfindlich gekränkt, * aber * damals * konnte * niemand * nachvollziehen, * womit, * obwohl * die kulturellen Gepflogenheiten * der jeweils anderen Seite * auf jeden Fall * hinreichend bekannt waren.
Eine Spionin des Inlandsgeheimdienstes * hatten * die Informanten * im Vorfeld der Verhandlungen * enttarnt, * aber * nicht einmal Experten * wussten * letztlich * zu sagen, * wie, * bis * irgendwann * eine Reinigungskraft * im Schutz der Anonymität * den entscheidenden Hinweis gab.
Eine Spionin des Inlandsgeheimdienstes * hatte * die Informanten * im Vorfeld der Verhandlungen * enttarnt, * aber * nicht einmal Experten * wussten * letztlich * zu sagen, * wie, * bis * irgendwann * eine Reinigungskraft * im Schutz der Anonymität * den entscheidenden Hinweis gab.
Einen Spion des Inlandsgeheimdienstes * hatten * die Informanten * im Vorfeld der Verhandlungen * enttarnt, * aber * nicht einmal Experten * wussten * letztlich * zu sagen, * wie, * bis * irgendwann * eine Reinigungskraft * im Schutz der Anonymität * den entscheidenden Hinweis gab.
Ein Spion des Inlandsgeheimdienstes * hatte * die Informanten * im Vorfeld der Verhandlungen * enttarnt, * aber * nicht einmal Experten * wussten * letztlich * zu sagen, * wie, * bis * irgendwann * eine Reinigungskraft * im Schutz der Anonymität * den entscheidenden Hinweis gab.
Eine Redakteurin der Tageszeitung * hatten * die maskierten Aktivisten * zu einer geheimen Videokonferenz * eingeladen, * aber * niemand * konnte * überzeugend * begründen, * wieso, * nachdem * das Vorhaben * unbeabsichtigterweise * der Öffentlichkeit * bekannt geworden war.
Eine Redakteurin der Tageszeitung * hatte * die maskierten Aktivisten * zu einer geheimen Videokonferenz * eingeladen, * aber * niemand * konnte * überzeugend * begründen, * wieso, * nachdem * das Vorhaben * unbeabsichtigterweise * der Öffentlichkeit * bekannt geworden war.
Einen Redakteur der Tageszeitung * hatten * die maskierten Aktivisten * zu einer geheimen Videokonferenz * eingeladen, * aber * niemand * konnte * überzeugend * begründen, * wieso, * nachdem * das Vorhaben * unbeabsichtigterweise * der Öffentlichkeit * bekannt geworden war.
Ein Redakteur der Tageszeitung * hatte * die maskierten Aktivisten * zu einer geheimen Videokonferenz * eingeladen, * aber * niemand * konnte * überzeugend * begründen, * wieso, * nachdem * das Vorhaben * unbeabsichtigterweise * der Öffentlichkeit * bekannt geworden war.
Eine Sachverständige aus Osteuropa * hatten * die Investoren * in der Planungsphase * eigenständig hinzugezogen, * aber * im Nachhinein * fragte * sich * so mancher Gutachter, * wieso, * da * das Ergebnis * augenscheinlich * nicht * verbessert wurde.
Eine Sachverständige aus Osteuropa * hatte * die Investoren * in der Planungsphase * eigenständig hinzugezogen, * aber * im Nachhinein * fragte * sich * so mancher Gutachter, * wieso, * da * das Ergebnis * augenscheinlich * nicht * verbessert wurde.
Einen Sachverständigen aus Osteuropa * hatten * die Investoren * in der Planungsphase * eigenständig hinzugezogen, * aber * im Nachhinein * fragte * sich * so mancher Gutachter, * wieso, * da * das Ergebnis * augenscheinlich * nicht * verbessert wurde.
Ein Sachverständiger aus Osteuropa * hatte * die Investoren * in der Planungsphase * eigenständig hinzugezogen, * aber * im Nachhinein * fragte * sich * so mancher Gutachter, * wieso, * da * das Ergebnis * augenscheinlich * nicht * verbessert wurde.
Eine Biologin mit Doktortitel * hatten * die Naturschützer * auf einer Fachkonferenz * äußerst heftig kritisiert, * aber * die anderen Teilnehmer * erinnerten * sich * nicht, * wieso, * zumal * die Diskussion * offenbar * abseits des Podiums * stattfand.
Eine Biologin mit Doktortitel * hatte * die Naturschützer * auf einer Fachkonferenz * äußerst heftig kritisiert, * aber * die anderen Teilnehmer * erinnerten * sich * nicht, * wieso, * zumal * die Diskussion * offenbar * abseits des Podiums * stattfand.
Einen Biologen mit Doktortitel * hatten * die Naturschützer * auf einer Fachkonferenz * äußerst heftig kritisiert, * aber * die anderen Teilnehmer * erinnerten * sich * nicht, * wieso, * zumal * die Diskussion * offenbar * abseits des Podiums * stattfand.
Ein Biologe mit Doktortitel * hatte * die Naturschützer * auf einer Fachkonferenz * äußerst heftig kritisiert, * aber * die anderen Teilnehmer * erinnerten * sich * nicht, * wieso, * zumal * die Diskussion * offenbar * abseits des Podiums * stattfand.
Eine Patientin mit unklaren Symptomen * hatten * die Krankenschwestern * dem behandelnden Arzt zufolge * mehrfach angeschrien, * aber * es * war * nicht * zu ergründen, * wieso, * obwohl * seitdem * schon * mehrere Gespräche * geführt wurden.
Eine Patientin mit unklaren Symptomen * hatte * die Krankenschwestern * dem behandelnden Arzt zufolge * mehrfach angeschrien, * aber * es * war * nicht * zu ergründen, * wieso, * obwohl * seitdem * schon * mehrere Gespräche * geführt wurden.
Einen Patienten mit unklaren Symptomen * hatten * die Krankenschwestern * dem behandelnden Arzt zufolge * mehrfach angeschrien, * aber * es * war * nicht * zu ergründen, * wieso, * obwohl * seitdem * schon * mehrere Gespräche * geführt wurden.
Ein Patient mit unklaren Symptomen * hatte * die Krankenschwestern * dem behandelnden Arzt zufolge * mehrfach angeschrien, * aber * es * war * nicht * zu ergründen, * wieso, * obwohl * seitdem * schon * mehrere Gespräche * geführt wurden.
Eine Teenagerin ohne Schulabschluss * hatten * die Talentsucher * in der Bewerbungsphase * angeschrieben, * aber * der Programmverantwortliche * fragte * sich * ernsthaft, * wozu, * denn * bemerkenswerte Fähigkeiten * wurden * an keiner Stelle * erwähnt.
Eine Teenagerin ohne Schulabschluss * hatte * die Talentsucher * in der Bewerbungsphase * angeschrieben, * aber * der

Programmverantwortliche * fragte * sich * ernsthaft, * wozu, * denn * bemerkenswerte Fähigkeiten * wurden * an keiner Stelle * erwähnt.
Einen Teenager ohne Schulabschluss * hatten * die Talentsucher * in der Bewerbungsphase * angeschrieben, * aber * der Programmverantwortliche * fragte * sich * ernsthaft, * wozu, * denn * bemerkenswerte Fähigkeiten * wurden * an keiner Stelle * erwähnt.
Ein Teenager ohne Schulabschluss * hatte * die Talentsucher * in der Bewerbungsphase * angeschrieben, * aber * der Programmverantwortliche * fragte * sich * ernsthaft, * wozu, * denn * bemerkenswerte Fähigkeiten * wurden * an keiner Stelle * erwähnt.
Eine Straßenhündin mit schwarzem Fell * hatten * die Kinder * bis an den Rand des Dorfes * verfolgt, * aber * niemand * konnte * sich * erklären, * weshalb, * zumal * das Tier * sich * normalerweise * vor Menschen versteckte.
Eine Straßenhündin mit schwarzem Fell * hatte * die Kinder * bis an den Rand des Dorfes * verfolgt, * aber * niemand * konnte * sich * erklären, * weshalb, * zumal * das Tier * sich * normalerweise * vor Menschen versteckte.
Einen Straßenhund mit schwarzem Fell * hatten * die Kinder * bis an den Rand des Dorfes * verfolgt, * aber * niemand * konnte * sich * erklären, * weshalb, * zumal * das Tier * sich * normalerweise * vor Menschen versteckte.
Ein Straßenhund mit schwarzem Fell * hatte * die Kinder * bis an den Rand des Dorfes * verfolgt, * aber * niemand * konnte * sich * erklären, * weshalb, * zumal * das Tier * sich * normalerweise * vor Menschen versteckte.
Eine Violinistin des Nationalorchesters * hatten * die Konzertbesucher * während der halbstündigen Pause * heimlich fotografiert, * aber * der Beitrag * verriet * leider * nicht, * weshalb, * sondern * befasste * sich * eher * mit der Bildqualität.
Eine Violinistin des Nationalorchesters * hatte * die Konzertbesucher * während der halbstündigen Pause * heimlich fotografiert, * aber * der Beitrag * verriet * leider * nicht, * weshalb, * sondern * befasste * sich * eher * mit der Bildqualität.
Einen Violinisten des Nationalorchesters * hatten * die Konzertbesucher * während der halbstündigen Pause * heimlich fotografiert, * aber * der Beitrag * verriet * leider * nicht, * weshalb, * sondern * befasste * sich * eher * mit der Bildqualität.
Ein Violinist des Nationalorchesters * hatte * die Konzertbesucher * während der halbstündigen Pause * heimlich fotografiert, * aber * der Beitrag * verriet * leider * nicht, * weshalb, * sondern * befasste * sich * eher * mit der Bildqualität.
Eine Korrespondentin des erfolgreichen Nachrichtensenders * hatten * die Kollegen * vor laufender Kamera * schlechtgemacht, * aber * in einem Gespräch * konnte * nicht * festgestellt werden, * weshalb, * sodass * der Konflikt * trotz aller Entschuldigungen * ohne Zweifel * weiterhin bestehen blieb.
Eine Korrespondentin des erfolgreichen Nachrichtensenders * hatte * die Kollegen * vor laufender Kamera * schlechtgemacht, * aber * in einem Gespräch * konnte * nicht * festgestellt werden, * weshalb, * sodass * der Konflikt * trotz aller Entschuldigungen * ohne Zweifel * weiterhin bestehen blieb.
Einen Korrespondenten des erfolgreichen Nachrichtensenders * hatten * die Kollegen * vor laufender Kamera * schlechtgemacht, * aber * in einem Gespräch * konnte * nicht * festgestellt werden, * weshalb, * sodass * der Konflikt * trotz aller Entschuldigungen * ohne Zweifel * weiterhin bestehen blieb.
Ein Korrespondent des erfolgreichen Nachrichtensenders * hatte * die Kollegen * vor laufender Kamera * schlechtgemacht, * aber * in einem Gespräch * konnte * nicht * festgestellt werden, * weshalb, * sodass * der Konflikt * trotz aller Entschuldigungen * ohne Zweifel * weiterhin bestehen blieb.
Eine Autorin aus Bolivien * hatten * die vier Literaturwissenschaftler * in einem 2500-Seiten-Werk * zitiert, * aber * noch * kann * niemand * sagen, * wo, * da * der Text * bislang * seltsamerweise * verschollen blieb.
Eine Autorin aus Bolivien * hatte * die vier Literaturwissenschaftler * in einem 2500-Seiten-Werk * zitiert, * aber * noch * kann * niemand * sagen, * wo, * da * der Text * bislang * seltsamerweise * verschollen blieb.
Einen Autor aus Bolivien * hatten * die vier Literaturwissenschaftler * in einem 2500-Seiten-Werk * zitiert, * aber * noch * kann * niemand * sagen, * wo, * da * der Text * bislang * seltsamerweise * verschollen blieb.
Ein Autor aus Bolivien * hatte * die vier Literaturwissenschaftler * in einem 2500-Seiten-Werk * zitiert, * aber * noch * kann * niemand * sagen, * wo, * da * der Text * bislang * seltsamerweise * verschollen blieb.
Eine Studentin mit außergewöhnlichen Leistungen * hatten * die Professoren * laut Stellungnahme des Instituts * tatkräftig unterstützt, * aber * es * war * nicht * zu erfahren, * wobei, * da * der Projektverantwortliche * nicht * für Nachfragen * zu erreichen ist.
Eine Studentin mit außergewöhnlichen Leistungen * hatte * die Professoren * laut Stellungnahme des Instituts * tatkräftig unterstützt, * aber * es * war * nicht * zu erfahren, * wobei, * da * der Projektverantwortliche * nicht * für Nachfragen * zu erreichen ist.
Einen Studenten mit außergewöhnlichen Leistungen * hatten * die Professoren * laut Stellungnahme des Instituts * tatkräftig unterstützt, * aber * es * war * nicht * zu erfahren, * wobei, * da * der Projektverantwortliche * nicht * für Nachfragen * zu erreichen ist.
Ein Student mit außergewöhnlichen Leistungen * hatte * die Professoren * laut Stellungnahme des Instituts * tatkräftig unterstützt, * aber * es * war * nicht * zu erfahren, * wobei, * da * der Projektverantwortliche * nicht * für Nachfragen * zu erreichen ist.
Eine Schwimmerin mit zwei Beinprothesen * hatten * die Komiteemitglieder * bezüglich der geplanten Werbekampagne * kontaktiert, * aber * es * bleibt * äußerst * schleierhaft, * wann, * zumal * das Schriftstück * angeblich * zwischenzeitlich * verloren gegangen ist.
Eine Schwimmerin mit zwei Beinprothesen * hatte * die Komiteemitglieder * bezüglich der geplanten Werbekampagne * kontaktiert, * aber * es * bleibt * äußerst * schleierhaft, * wann, * zumal * das Schriftstück * angeblich * zwischenzeitlich * verloren gegangen ist.
Einen Schwimmer mit zwei Beinprothesen * hatten * die Komiteemitglieder * bezüglich der geplanten Werbekampagne * kontaktiert, * aber * es * bleibt * äußerst * schleierhaft, *

wann, * zumal * das Schriftstück * angeblich * zwischenzeitlich * verloren gegangen ist.
Ein Schwimmer mit zwei Beinprothesen * hatte * die Komiteemitglieder * bezüglich der geplanten Werbekampagne * kontaktiert, * aber * es * bleibt * äußerst * schleierhaft, * wann, * zumal * das Schriftstück * angeblich * zwischenzeitlich * verloren gegangen ist.
Eine Mathematikerin mit Programmierkenntnissen * hatten * die Seitenbetreiber * über die Sicherheitslücke * informiert, * aber * der Staatsanwalt * wollte * genau * wissen, * wann, * da * dies * für den Tathergang * womöglich * äußerst entscheidend war.
Eine Mathematikerin mit Programmierkenntnissen * hatte * die Seitenbetreiber * über die Sicherheitslücke * informiert, * aber * der Staatsanwalt * wollte * genau * wissen, * wann, * da * dies * für den Tathergang * womöglich * äußerst entscheidend war.
Einen Mathematiker mit Programmierkenntnissen * hatten * die Seitenbetreiber * über die Sicherheitslücke * informiert, * aber * der Staatsanwalt * wollte * genau * wissen, * wann, * da * dies * für den Tathergang * womöglich * äußerst entscheidend war.
Ein Mathematiker mit Programmierkenntnissen * hatte * die Seitenbetreiber * über die Sicherheitslücke * informiert, * aber * der Staatsanwalt * wollte * genau * wissen, * wann, * da * dies * für den Tathergang * womöglich * äußerst entscheidend war.
Eine Abgeordnete der Landtagsfraktion * hatten * die Finanzbeamten * in einem offenen Brief * gemaßregelt, * aber * fünfzig Jahre später * erscheint * es * unverständlich, * weshalb, * da * aus heutiger Sicht * wohl * kein Fehlverhalten * vorlag.
Eine Abgeordnete der Landtagsfraktion * hatte * die Finanzbeamten * in einem offenen Brief * gemaßregelt, * aber * fünfzig Jahre später * erscheint * es * unverständlich, * weshalb, * da * aus heutiger Sicht * wohl * kein Fehlverhalten * vorlag.
Einen Abgeordneten der Landtagsfraktion * hatten * die Finanzbeamten * in einem offenen Brief * gemaßregelt, * aber * fünfzig Jahre später * erscheint * es * unverständlich, * weshalb, * da * aus heutiger Sicht * wohl * kein Fehlverhalten * vorlag.
Ein Abgeordneter der Landtagsfraktion * hatte * die Finanzbeamten * in einem offenen Brief * gemaßregelt, * aber * fünfzig Jahre später * erscheint * es * unverständlich, * weshalb, * da * aus heutiger Sicht * wohl * kein Fehlverhalten * vorlag.
Eine Sanitäterin des Rettungsteams * hatten * die Feuerwehrleute * nachdrücklich * um Hilfe gebeten, * aber * man * verstand * später * nicht, * warum, * bis * schließlich * Bildmaterial vom Unglücksort * das Ausmaß der Verwüstung * verständlich machte.
Eine Sanitäterin des Rettungsteams * hatte * die Feuerwehrleute * nachdrücklich * um Hilfe gebeten, * aber * man * verstand * später * nicht, * warum, * bis * schließlich * Bildmaterial vom Unglücksort * das Ausmaß der Verwüstung * verständlich machte.
Einen Sanitäter des Rettungsteams * hatten * die Feuerwehrleute * nachdrücklich * um Hilfe gebeten, * aber * man * verstand * später * nicht, * warum, * bis * schließlich * Bildmaterial vom Unglücksort * das Ausmaß der Verwüstung * verständlich machte.
Ein Sanitäter des Rettungsteams * hatte * die Feuerwehrleute * nachdrücklich * um Hilfe gebeten, * aber * man * verstand * später * nicht, * warum, * bis * schließlich * Bildmaterial vom Unglücksort * das Ausmaß der Verwüstung * verständlich machte.
Eine Befürworterin der Steuerreform * hatten * die Leiter der betroffenen Behörden * wiederholt * verbal angegriffen, * aber * es * bleibt * völlig * im Dunkeln, * weshalb, * da * das Wortgefecht * von beiden Seiten * überaus unsachlich * geführt wurde.
Eine Befürworterin der Steuerreform * hatte * die Leiter der betroffenen Behörden * wiederholt * verbal angegriffen, * aber * es * bleibt * völlig * im Dunkeln, * weshalb, * da * das Wortgefecht * von beiden Seiten * überaus unsachlich * geführt wurde.
Einen Befürworter der Steuerreform * hatten * die Leiter der betroffenen Behörden * wiederholt * verbal angegriffen, * aber * es * bleibt * völlig * im Dunkeln, * weshalb, * da * das Wortgefecht * von beiden Seiten * überaus unsachlich * geführt wurde.
Ein Befürworter der Steuerreform * hatte * die Leiter der betroffenen Behörden * wiederholt * verbal angegriffen, * aber * es * bleibt * völlig * im Dunkeln, * weshalb, * da * das Wortgefecht * von beiden Seiten * überaus unsachlich * geführt wurde.
Eine Gegnerin des umstrittenen Staudammprojekts * hatten * die Planer * schließlich * doch noch überzeugt, * aber * es * herrscht * Stillschweigen * darüber, * wie, * weil * niemand * sich * dem Verdacht der Bestechlichkeit * aussetzen will.
Eine Gegnerin des umstrittenen Staudammprojekts * hatte * die Planer * schließlich * doch noch überzeugt, * aber * es * herrscht * Stillschweigen * darüber, * wie, * weil * niemand * sich * dem Verdacht der Bestechlichkeit * aussetzen will.
Einen Gegner des umstrittenen Staudammprojekts * hatten * die Planer * schließlich * doch noch überzeugt, * aber * es * herrscht * Stillschweigen * darüber, * wie, * weil * niemand * sich * dem Verdacht der Bestechlichkeit * aussetzen will.
Ein Gegner des umstrittenen Staudammprojekts * hatte * die Planer * schließlich * doch noch überzeugt, * aber * es * herrscht * Stillschweigen * darüber, * wie, * weil * niemand * sich * dem Verdacht der Bestechlichkeit * aussetzen will.
Eine Soldatin der gegnerischen Streitkräfte * hatten * die ausgesandten Kundschafter * offenbar * in die Irre geführt, * aber * der Befehlshaber * begriff * einfach * nicht, * wie, * obwohl * ihm * die Finte * mehrmals * erklärt worden war.
Eine Soldatin der gegnerischen Streitkräfte * hatte * die ausgesandten Kundschafter * offenbar * in die Irre geführt, * aber * der Befehlshaber * begriff * einfach * nicht, * wie, * obwohl * ihm * die Finte * mehrmals * erklärt worden war.
Einen Soldaten der gegnerischen Streitkräfte * hatten * die ausgesandten Kundschafter * offenbar * in die Irre geführt, * aber * der Befehlshaber * begriff * einfach * nicht, * wie, * obwohl * ihm * die Finte * mehrmals * erklärt worden war.
Ein Soldat der gegnerischen Streitkräfte * hatte * die ausgesandten Kundschafter * offenbar * in die Irre geführt, * aber * der Befehlshaber * begriff * einfach * nicht, * wie, * obwohl * ihm * die Finte * mehrmals * erklärt worden war.

frontiers in Psychology

ORIGINAL RESEARCH
published: 23 March 2015
doi: 10.3389/fpsyg.2015.00312

Working memory differences in long-distance dependency resolution

Bruno Nicenboim[1]*, *Shravan Vasishth*[1], *Carolina Gattei*[2], *Mariano Sigman*[3,4] and *Reinhold Kliegl*[5]

[1] Department of Linguistics, University of Potsdam, Potsdam, Germany, [2] Grupo de Lingüística y Neurobiología Experimental del Lenguaje, Instituto de Ciencias Humanas, Sociales y Ambientales, Consejo Nacional de Investigaciones Científicas y Técnicas, Mendoza, Argentina, [3] Departamento de Física, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires/Instituto de Física de Buenos Aires, Consejo Nacional de Investigaciones Científicas y Técnicas, Buenos Aires, Argentina, [4] Escuela de Negocios, Universidad Torcuato Di Tella, Buenos Aires, Argentina, [5] Department of Psychology, University of Potsdam, Potsdam, Germany

There is a wealth of evidence showing that increasing the distance between an argument and its head leads to more processing effort, namely, locality effects; these are usually associated with constraints in working memory (DLT: Gibson, 2000; activation-based model: Lewis and Vasishth, 2005). In SOV languages, however, the opposite effect has been found: antilocality (see discussion in Levy et al., 2013). Antilocality effects can be explained by the expectation-based approach as proposed by Levy (2008) or by the activation-based model of sentence processing as proposed by Lewis and Vasishth (2005). We report an eye-tracking and a self-paced reading study with sentences in Spanish together with measures of individual differences to examine the distinction between expectation- and memory-based accounts, and within memory-based accounts the further distinction between DLT and the activation-based model. The experiments show that (i) antilocality effects as predicted by the expectation account appear only for high-capacity readers; (ii) increasing dependency length by interposing material that modifies the head of the dependency (the verb) produces stronger facilitation than increasing dependency length with material that does not modify the head; this is in agreement with the activation-based model but not with the expectation account; and (iii) a possible outcome of memory load on low-capacity readers is the increase in regressive saccades (locality effects as predicted by memory-based accounts) or, surprisingly, a speedup in the self-paced reading task; the latter consistent with good-enough parsing (Ferreira et al., 2002). In sum, the study suggests that individual differences in working memory capacity play a role in dependency resolution, and that some of the aspects of dependency resolution can be best explained with the activation-based model together with a prediction component.

Keywords: locality, antilocality, working memory capacity, individual differences, Spanish, activation, DLT, expectation

OPEN ACCESS

Edited by:
Matthew Wagers, University of California, Santa Cruz, USA

Reviewed by:
Leticia Pablos, Leiden University, Netherlands
Kiel Christianson, University of Illinois, USA

***Correspondence:**
Bruno Nicenboim, Departament of Linguistics, University of Potsdam, Haus 14, Karl-Liebknecht-Str. 24–25, Potsdam D-14476, Germany
bruno.nicenboim@uni-potsdam.de

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

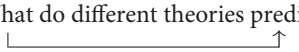
Received: *14 November 2014*
Accepted: *04 March 2015*
Published: *23 March 2015*

Citation:
Nicenboim B, Vasishth S, Gattei C, Sigman M and Kliegl R (2015) Working memory differences in long-distance dependency resolution. Front. Psychol. 6:312. doi: 10.3389/fpsyg.2015.00312

1. Introduction

Long-distance dependencies (also called non-local, filler-gap, or unbounded dependencies) have been investigated since Fodor's (1978) work on parsing strategies, but many questions remain

unanswered or only partially answered. It is uncontroversial that the distance over which a dependency is resolved, shown in (1) with an arrow, is a primary determinant of the speed and the accuracy of the dependency resolution (among others: Gibson, 2000; McElree et al., 2003; Lewis and Vasishth, 2005; Levy, 2008). It is controversial, however, how increasing this distance affects the speed and the accuracy of the resolution.

(1) What do different theories predict?


1.1. Memory-Based Explanations

There is a wealth of evidence showing that increasing the distance between an argument and its head hinders underlying memory processes in some way. This is supported by research that shows that longer dependencies produced (i) locality effects, that is, a slowdown (or increase of regressive saccades) at the region of the dependency resolution when the distance between dependent and head or subcategorizing verb (or gap) is increased (either in self-paced reading, eye-tracking experiments, or both; among others: Gibson, 2000; Grodner and Gibson, 2005; Demberg and Keller, 2008; Bartek et al., 2011; Vasishth and Drenhaus, 2011); (ii) Event Related Potential (ERP) measures associated with difficulty (Kluender and Kutas, 1993; Fiebach et al., 2002; but see: Phillips et al., 2005); and (iii) deterioration of response accuracy in speed-accuracy trade-off (SAT) experiments (McElree, 2000; McElree et al., 2003). The underlying memory process that is adversely affected when distance is increased is subject to debate. Here we discuss two theories that account for the memory-based locality effects: dependency locality theory (DLT; Gibson, 2000) and the activation-based model (Lewis and Vasishth, 2005).

DLT posits two separate components of a sentence's processing cost: storage and integration costs. Storage cost is argued to depend on the number of syntactic heads required to complete the current input as a grammatical sentence (Gibson, 2000) and seems to be independent of the amount of time that an incomplete dependency is held in memory (Gibson et al., 2005). On the other hand, integration cost is locality-based, that is, the cost is based on the distance between the dependent and its head; this distance is based on the number of new intervening discourse referents (Gibson, 2000).

In contrast to DLT, which is a theory specific to sentence comprehension processes, the activation-based model is based on a general cognitive model. In the activation-based model, linguistic items in memory are represented as feature bundles that suffer from decay and interference from the features of other linguistic items. Under this model, locality effects can be explained in terms of difficulty in the retrieval of a non-local argument; retrieval is driven by cues that are set at the moment of dependency resolution. Since the access to the argument involves a match of retrieval cue features against candidate memory items (Lewis et al., 2006), this access is adversely affected when (i) more time has passed from the encoding of the argument (decay); and (ii) when there are other items with similar features that serve as distractors (similarity-based interference). The activation-based model excludes the possibility of storage costs as proposed by DLT, but stored memories have their observable effects through interference (Van Dyke and Lewis, 2003; Lewis et al., 2006).

Thus, in cases such as (2), both DLT and the activation-based model predict that as the distance between the displaced argument *who* and the subcategorizing verb *supervised* increases, the retrieval of the argument will be harder. This is supported by the evidence of locality effects in relative clauses (Grodner and Gibson, 2005; Bartek et al., 2011).

(2) From Experiment 2 of Grodner and Gibson (2005)
   a. The administrator **who** the nurse **supervised**...
   b. The administrator **who** the nurse from the clinic **supervised**...
   c. The administrator **who** the nurse who was from the clinic **supervised**...

In spite of the evidence for locality effects, there is a growing body of evidence showing the opposite effect: antilocality. Studies on SOV structures (in Hindi: Vasishth, 2003; Vasishth and Lewis, 2006; and in German: Konieczny, 2000; Konieczny and Döring, 2003; Levy and Keller, 2013) showed that increasing distance can produce a speedup at the site of the dependency completion. In many cases the speedup can be accommodated in the activation-based model since the interposed material can help to strengthen the representation of the upcoming head by activating it through modification (Vasishth and Lewis, 2006). This would entail that the processing of the head will be facilitated since it has already been generated; we will express that here by saying that the VP has been *preactivated*. This is specially relevant for SOV languages, where the arguments of the VP appear preverbally, modifying the VP before the head is parsed. So, in cases such as (3), where the extra material belongs to the VP, the activation-based account will predict that increasing distance should, in fact, result in a speedup (but only if the decay does not offset the benefit of activation; Lewis et al., 2006).

(3) From Vasishth and Lewis (2006)
   a. Vo kaagaz **jisko** us laṛke-ne **dekhaa** bahut puraanaa thaa.
      that paper **which** that boy-ERG **saw** very old was
      'That paper which that boy saw was very old.' (Object relative, no intervening discourse referents)

   b. Vo kaagaz **jisko** us laṛke-ne mez-ke piiche gire.hue **dekhaa** bahut puraanaa thaa.
      that paper **which** that boy-ERG table-GEN behind fallen **saw** very old was
      'That paper which that boy saw fallen behind a/the table was very old.' (Object relative, two intervening discourse referents)

1.2. Expectation-Based Explanations

As in other aspects of cognition, predictions play an important role in language, and evidence from different sources supports the view that language processing does not only depend on bottom-up processes (for a review of prediction in language see: Kutas

et al., 2011). It has been shown that a syntactically constraining context can lead to facilitation when a word is predicted either (i) because of local syntactic constraints related to characteristic of verbs, as proposed by Trueswell et al. (1993), and Konieczny (2000); or (ii) because the parser is able to build structure in a top-down manner, using grammatical or probabilistic information, as proposed by Jurafsky (1996) and Hale (2001). The latter idea was developed further in an expectation-based theory of processing (Levy, 2008) where the main source of difficulty is determined by the surprisal (negative log of the conditional probability) of a word given its context (as proposed by Hale, 2001). The surprisal metric proposed by Hale (2001) formalizes the idea that a more surprising lexical content is also less predictable.

Long-distance dependency resolution is a situation where the comprehender knows that a subcategorizing verb has to appear, but does not know exactly when. Since each constituent of a given category that is integrated after the dependent (a wh-element in this case) eliminates most of the expectation for seeing another constituent of the same type next, each constituent that is read increases the expectation for seeing a constituent of one of the remaining types. Because the subcategorizing verb is one of the remaining types, the expectations of finding it will increase monotonically, and being more expected it will also be processed more easily. In other words, given that the clause has a finite length, the probability that the next word will be the subcategorizing verb rises as the number of words after finding the wh-element increases (in a similar way to an increasing hazard function as proposed for visual search by Peterson et al., 2001, and for the anticipation function of environmental cues in macaques by Janssen and Shadlen, 2005).

Thus, also in the cases where memory-based accounts will predict locality effects (due to integration or retrieval costs), the expectation-based account of dependency resolution will predict the opposite effect: antilocality. The predictions of the expectation-based account for non-local dependency resolution were borne out specially in studies using languages with SOV structures, which showed antilocality effects. However, as mentioned before, in many cases the predicted antilocality effects could also be explained either with local syntactic constraints (Konieczny, 2000) or with an activation-based account (Vasishth, 2003; Vasishth and Lewis, 2006). Independent support for the expectation-based account of antilocality in dependency resolution would come from cases where the length manipulation is independent of material that belongs to the VP and appears preverbally. Cases like this can be found in length manipulations such as (4): object wh-questions where the dependency crosses over a sentence boundary. This is examined in more depth in the experiments of this paper.

(4) a. **Who** has John **called**?
b. **Who** does Mary think that John has **called**?

1.3. Individual Differences

1.3.1. Working Memory Capacity and the Parsing of Unbounded Dependencies

Memory-based accounts of locality effects assume, either implicitly or explicitly, that if more working memory capacity (WMC) is required for processing than is available, longer processing times and/or a higher proportion of errors will result during retrieval or integration. This prediction is implicit in DLT, where the upper limits on storage and integration cost (Gibson and Thomas, 1999; Gibson, 2000) should depend on WMC; and it is explicit in the activation-based model, where low capacity is argued to result in hindered ability to complete a retrieval (Daily et al., 2001). One plausible implication is that low-capacity readers may be more affected by locality effects, showing stronger effects than high-capacity readers.

However, the effect of individual differences in WMC influencing dependency resolution processes has been neglected in the literature (but see: Van Dyke et al., 2014). This absence of work is surprising given that there is considerable evidence for the interaction of individual differences with syntactic and semantic processes (Just and Carpenter, 1992; Pearlmutter and MacDonald, 1995; Traxler et al., 2005, 2012; von der Malsburg and Vasishth, 2013), and there is also evidence for a reduction in performance during long-distance dependency resolution and memory dual-tasks (Fedorenko et al., 2006, 2013).

Regarding the influence of working memory on expectation-based parsing, the predictions are less clear. The studies showing that expectations may play a dominant role only when working memory load is relatively low (Levy, 2008; Levy and Keller, 2013; Husain et al., 2014) suggest that the processes involved in the anticipation of upcoming material may also depend on working memory. This is so because comprehenders' expectations depend on the accumulating information (Levy, 2008). Low-WMC readers, who have a reduced ability to temporarily store and manipulate information, may then be less able to adequately expect upcoming lexical material, relative to high-WMC readers. To our knowledge, the only evidence for this claim, however, comes from Otten and Van Berkum's (2009) ERP study where low-WMC participants showed an additional later negativity (900–1500 ms) to unexpected content.

1.3.2. WMC and Reading Skills

Differences in WMC can successfully explain individual differences in comprehension performance (Daneman and Carpenter, 1980); and this measure of individual differences seems to be the right candidate to account for differential effects in processes related to dependency resolution. There is ample evidence showing that lower WMC reflects higher limitations in attention allocation for goals (Engle, 2002), and several studies have shown the predictive power of WMC for language comprehension ability (for a meta-analysis of 77 studies till the mid-nineties: Daneman and Merikle, 1996). Furthermore, some studies have shown that individuals with lower capacity are less successful in integrating information over distance in a text (Daneman and Carpenter, 1980; Yuill et al., 1989), and have greater comprehension deficits, in part, because they are less able to maintain on-task thought (McVay and Kane, 2011). Moreover, low-capacity participants seem to have a greater disadvantage than high-capacity participants when they face difficult sentences (for garden-path vs. non-garden path sentences: Christianson et al., 2006; for comprehension reaction times in subject- vs. object-relative clauses: King and Just, 1991; Vos et al., 2001). The reason for differences in WMC may be rooted in the variability in either a limited amount

of activation (Just and Carpenter, 1992; van Rij et al., 2013), computational resources available or processing efficiency (among others: Daneman and Carpenter, 1980; Daneman and Carpenter, 1983), the ability to overcome interference (Hasher and Zacks, 1988; Unsworth and Engle, 2007), or the efficiency of retrieval cues present in the active portion of working memory (Ericsson and Kintsch, 1995).

It is possible, however, that individual differences in capacity only reflect experience and not intrinsic capacity differences (MacDonald and Christiansen, 2002; Wells et al., 2009). Readers characterized as high-capacity may indeed be more sensitive to the semantic cues available to them, as proposed by Pearlmutter and MacDonald (1995), but mainly because these readers also have more language experience. In fact, recent work by Traxler et al. (2012) raises the concern that WMC correlates with many other reader characteristics. According to Traxler et al., fast readers, who read more often than slow readers, will have greater experience with language; this would in turn make them more sensitive to semantic cues in the syntactic analysis. In a new set of analyses based on Traxler et al.'s (2005) data set, Traxler and colleagues found that WMC interacted with sentence-characteristic variables only when reading speed was not included in the model (since they assumed that reading speed was a measure of reading skills).

In order to obtain a reliable measure of working memory that is not correlated with reading speed and experience, we chose to use the operation span task (Turner and Engle, 1989; Conway et al., 2005). In addition, we adopted the rapid automatized naming task (Denckla and Rudel, 1976), since it has been shown that it predicts reading speed, comprehension, and other characteristics associated with reading skills (among others: Kuperman and Van Dyke, 2011). The inclusion of both tasks can therefore help to determine whether it is WMC and/or reading experience that account for differences in dependency resolution processes.

2. Experiments

The experiments have two main objectives. The first objective is to disentangle memory- and expectation-based explanations on the processing of long-distance dependencies. While both the expectation and activation accounts may predict antilocality effects, the activation-based model predicts that facilitation should occur when intervening material modifies an upcoming head, whereas the expectation account predicts facilitation regardless of what the intervening material modifies. Even though this is an oversimplification of the expectation account as defined by Hale (2001) and Levy (2008), it should hold for the type of sentences we included in our stimuli.

The second objective is to examine the effect of individual differences in dependency resolution: if working memory constraints are involved, participants with different WMC should show differential locality or antilocality effects.

In order to address these objectives, we measured WMC and reading skills of (Argentinean) Spanish native speakers, and we used both self-paced reading and eyetracking methodologies to provide converging evidence. The use of Spanish stimuli allowed us to investigate antilocality effects in an SVO language. In addition, because of the relatively free order and long sentences permitted by Spanish, we could do a manipulation that is more common in studies that investigate antilocality in *SOV* structures: increasing the dependent-head distance by interposing material that belongs to the verbal phrase (VP) but appears prior to the verb.

The design of the stimuli is exemplified by (5). The distance between the wh-element and the head verb (*had fired*) was manipulated by including an adverbial phrase (AdvP; *before some days*) that attaches to the different VPs in the sentence. Hence there are two different aspects of the manipulation to consider for each condition: (i) the attachment site of the adverbial phrase (main VP, intermediate VP, and last VP where the dependency is completed) and (ii) the length of the dependency between the wh-word (who.ACC) and the head verb. In (5a), the length of the dependency is the shortest one, since the AdvP is attached to the main clause VP *asked* (henceforth condition VP1). This entails that by the time the dependency is started at the wh-element, the AdvP has already been interpreted. In this condition, the action that was performed *before some days* was the "asking." In both conditions (5b) and (5c) the dependency length is larger than in (5a), since the the AdvP is interposed between the dependent and head verb. However, while in (5b) the AdvP modifies an intermediate VP (henceforth condition VP2), in (5c) it modifies the third VP, which contains the head verb, where the dependency is completed (henceforth condition VP3). So while in condition VP2 the "saying" happened *before some days*, in condition VP3 the "firing" of the dependent "who.ACC" was *before some days*. All the items had as a second verb either *comentar* or *decir* "to say." Even though these two verbs are ditransitive, the ditransitive construction is extremely uncommon in Argentinean Spanish without a clitic. This means that the reading that would allow an indirect object such as *a quién* completing the dependency is very unlikely (for a similar construction in Spanish with clitic left-dislocation, see Pablos, 2006). Since this type of verbs appears in all conditions, and they are not in the region of interest, they should not affect the experiment. Notice, as well, that the head verb position is kept fixed across conditions in order to avoid word-position effects (Ferreira and Henderson, 1993). The characteristics of the stimuli are summarized in **Table 1**.

(5) a. ATTACHMENT AT VP1

Hace algunos días, José preguntó **a quién**
Before some days José asked **who.ACC**
comentaron que el gerente **había despedido**
they-said that the manager **had fired**
por equivocación.
because-of mistake

"Some days ago, José asked who they said that the manager had fired by mistake."

b. ATTACHMENT AT VP2

José preguntó **a quién**, hace algunos días,
José asked **who.ACC** before some days
comentaron que el gerente **había despedido**
they-said that the manager **had fired**
por equivocación.
because-of mistake

"José asked who they said some days ago that the manager had fired by mistake."

c. ATTACHMENT AT VP3

José preguntó **a quién** comentaron que,
José asked **who.ACC** they-said that
hace algunos días, el gerente **había despedido**
before some days the manager **had fired**
por equivocación.
because-of mistake

"José asked who they said that the manager had fired some days ago by mistake."

2.1. Predictions

Predictions for the critical region (head verb) are summarized in **Table 2**. When the dependency length is increased (VP2 vs. VP1 and VP3 vs. VP1), DLT predicts increased processing effort, that is, locality-effects. In contrast, the expectation account predicts facilitation at the head verb, that is, antilocality effects. The activation-based model predicts, similar to DLT, increased processing effort for both VP2 and VP3 due to the decay of the wh-element. However, in contrast to DLT, the activation-based model also predicts that in VP3 this difficulty should be counteracted by the preactivation of the VP that contains the head verb. According to the activation account, while VP2 should display locality effects, the effect displayed by VP3 should depend on which underlying process is stronger: activation or decay (which in turn should depend on WMC).

It should be noted that while for self-paced reading experiments stronger locality effects imply longer reading times (Gibson, 2000; Grodner and Gibson, 2005; Bartek et al., 2011) and stronger antilocality effects imply shorter ones (Konieczny, 2000; Vasishth and Lewis, 2006; Levy, 2008), for eye-tracking studies these effects have been associated with different measures. Locality has been associated with the increase in the duration of first pass reading times in Staub (2010), total reading times and second pass reading in Bartek et al. (2011) and Levy and Keller (2013), and higher re-reading probabilities in Vasishth and Drenhaus (2011); and antilocality with the reduction of the duration of total reading times and second pass reading in Levy and Keller (2013), regression-path durations in Konieczny and Döring (2003), and lower first-pass regression probabilities in Vasishth and Drenhaus (2011).

Since the processing efforts of DLT and the activation account are associated with working memory constraints, according to these memory-based theories, participants with different WMC should show differential effects: the parse of the critical region will require more processing effort for low-WMC readers than for high-WMC. Thus, DLT predicts that as WMC increases, locality effects should decrease; and for high WMC (compared to low WMC) there should be the smallest difference between long and short conditions (see **Figure 1A**). For the expectation account, it is not clear whether WMC plays a role at all. If WMC is not relevant, there should not be a differential effect depending on the WMC of the readers (as in **Figure 1C**). It may be the case, however, that readers with more WMC are able to predict upcoming material better, then they should also display stronger antilocality effects (till a certain limit: either a minimal duration of the fixations or reading times or virtually no re-reading, as it is seen in **Figure 1D**). Regarding the activation-based account, its prediction for condition VP2 should be the same as the one of DLT: as WMC increases, locality effects should decrease; however, for condition VP3 the locality effects should be counteracted with facilitation due to preactivation, and given enough WMC, readers should offset the processing efforts and display antilocality effects (**Figure 1B**).

However, expectation and memory-based theories are not mutually exclusive; recent research supports the idea that insights from both types of theories are needed (Staub, 2010; Vasishth and Drenhaus, 2011; Levy and Keller, 2013; Levy et al., 2013; Husain et al., 2014). If DLT acts together with the expectation account (either the type that does not depend on memory, see **Figure 1C**, or the one that does depend on memory, see **Figure 1D**), locality effects should decrease as WMC increases until they become increasing antilocality effects, but, as before, the facilitation should not exceed a certain lower limit (see **Figure 1E**). As it is the case with each of these two accounts independently, the combination of DLT with the expectation account does not predict any difference between VP2 and VP3. If the activation-based model acts together with the expectation account, locality effects should also decrease together with an increase of WMC till they become increasing antilocality effects. However, processing efforts should be weaker and facilitation stronger for VP3 in contrast to VP2, since the facilitation of VP3 has two sources: expectations and preactivation, while the source of facilitation in VP2 is only expectations (see **Figure 1F**).

TABLE 1 | Summary of the conditions.

Cond.	Constituent modified by AdvP	Dependency length
VP1	Main VP (head: asked)	Short
VP2	Intermediate VP (head: said)	Long
VP3	VP where the dep. is completed (head: had fired)	Long

TABLE 2 | Summary of the conditions and predictions for the head of the dependency.

Cond.	Expectation account	Memory-based accounts	
		DLT	Activation
VP1	Baseline	Baseline	Baseline
VP2	Facilitation	Difficulty	Difficulty
VP3	Facilitation	Difficulty	Difficulty and Facilitation

2.2. General Procedure

Participants were tested individually using a PC computer. They got an overview of the whole experiment and then completed three tasks at their own pace: First, they performed a rapid automatized naming task; second, an operation span task; and

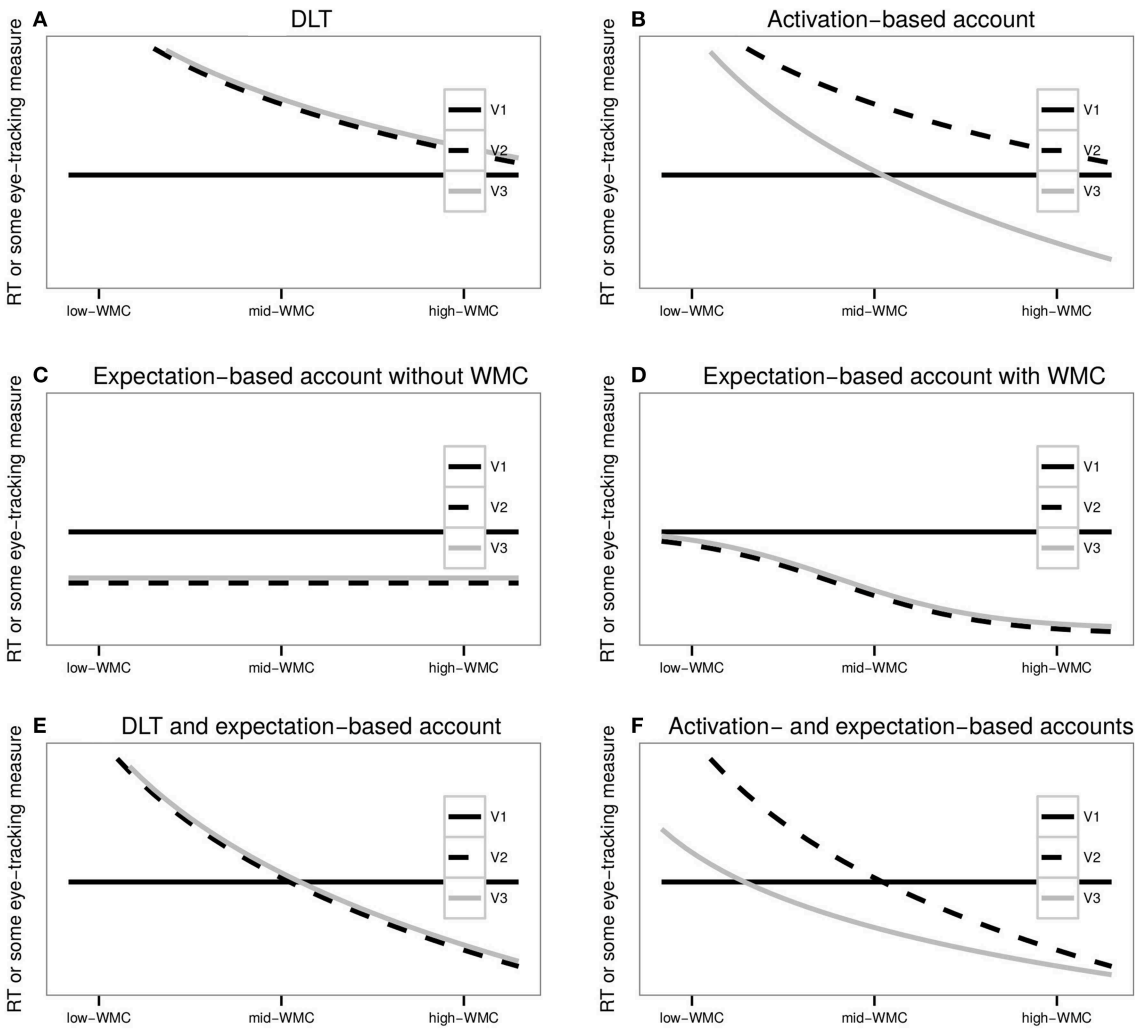


FIGURE 1 | The figure depicts the predictions of (A) DLT, (B) activation-based account, (C) expectation-based account unaffected by WMC, (D) expectation-based account affected by WMC, (E) the combination of the predictions of DLT and the expectation-based account, and (F) the combination of the predictions of the activation- and the expectation-based accounts.

finally, subjects performed an eye-tracking experiment in Experiment 1, and a self-paced reading task in Experiment 2.

2.2.1. Operation Span Task

Participants took part in the operation span task (Turner and Engle, 1989) using a software developed by von der Malsburg (https://github.com/tmalsburg/py-span-task) and used in von der Malsburg and Vasishth (2013) following the recommendations given in Conway et al. (2005). Even though variants of the reading span task by Daneman and Carpenter (1980) have been used in many psycholinguistic studies, it is likely that the reading span task measures verbal ability or reading experience as well as WMC (MacDonald and Christiansen, 2002; Conway et al., 2005). Since reading experience is also a good candidate for explaining differential effects in sentence processing, a solution is to include a nonverbal task to examine the domain-general aspects of cognition that may contribute to the individual differences (Swets et al., 2007). Since the operation span task probably measures mathematical ability as well as working memory (but not reading skills), if higher scores of the operation span task predict facilitation between experimental conditions, it would be unlikely that the result could be explained by the effect of reading experience alone.

The procedure of the operation span task test was similar to the one employed by von der Malsburg and Vasishth (2013) with some minor modifications: First, participants had to verify the correctness of 25 simple equations. At this stage, the reaction time of the Equations 10 to 25 was measured; the average reaction time plus two standard deviations was used as a time-out at the second stage. Calculating a time-out for every participant ensures that participants that are fast will not have time left to rehearse the items at the following stage of the test. Afterwards, participants had to carry out a dual task: check equations and memorize letters that were shown between the equations for

800 ms. After a group of equation-letter successions, participants were instructed to type in order the letters that had appeared before.

Before participating in the actual test, subjects practised with four trials of equation-letter successions. In the main test, successions of equation-letter had between three and seven elements, and there were eight sets for each size resulting in 32 trials. Presentation order of the sets was randomized and no feedback regarding the correctness of the judgments of equations or recalled items was given.

In all parts of the test, participants had to read the equations and letters aloud in order to prevent vocal rehearsal strategies. Only consonants were used as memory items to prevent participants from forming "words" with vowels and consonants, or "sentences," if words had been used.

Partial-credit unit scores, which indicate the mean proportion of correctly recalled items within the sets (Conway et al., 2005), were used as a numeric score of individual working memory.

2.2.2. Rapid Automatized Naming Tasks

Working memory-capacity correlates with other reader characteristics, which may in turn account for the variance in participants' reading behavior as well as or better than working memory capacity (Traxler et al., 2012). To determine whether working memory capacity correlates with reading times independently of reading skills, it is important to assess the effects of working memory capacity in the presence of some measure of reading skills.

Even though there are different ways to measure reading skills (among others: speeded naming abilities, oral language ability, vocabulary, attention), Kuperman and Van Dyke (2011) analyzed which tests from a broad battery of individual difference measures were predictive of eye-movement patterns associated with reading ability. They showed that rapid automatized naming was a robust predictor across the entire eye-movement record.

Participants with longer rapid automatized naming times tend to have lower reading comprehension scores, slower reading rates and their initial landing position when fixating tends to be further to the left (among others: Howe et al., 2006; Arnell et al., 2009; Kuperman and Van Dyke, 2011). Moreover, rapid automatized naming tasks seem to recruit a network of neural structures also involved in more complex reading tasks (Misra et al., 2004). In normal reading, readers must be able to disengage from one stimulus and move to another, rapidly programming saccades as the eyes move. Since this task involves speeded serial visual inspection and subsequent naming of items, the oculomotor component of this task is very similar to that required in natural reading.

In order to measure rapid automatized naming times, the first author developed a software that automatizes the test (https://github.com/bnicenboim/py-ran-task). In this task, participants saw a grid containing items (either letters or digits), and they were instructed to name them as fast as possible.

Each subject read a series of screens with 50 items; the items were the same set of letters or numbers that were used in Denckla and Rudel (1976): {o, a, s, d, p} and {2, 6, 9, 4, 7}. The first eight trials were composed of letters and the following eight had numbers. The items were displayed in five rows of ten columns and were listed in random order with the constraint that adjacent items were not the same. Before every trial, a screen with underscores instead of the items was displayed.

The participants were instructed to read aloud as fast as possible, and in case they misread, they were instructed to reread only the misread item. The test started with two practice trials to familiarize the participants with the task. Each trial started and ended with the spacebar: participants were instructed to start reading immediately after pressing the spacebar, and to press it again immediately after finishing reading aloud the last item.

Since the total reading times for letters and for numbers were highly correlated ($r = 0.88$ for Experiment 1 and $r = 0.87$ for Experiment 2), both were averaged together. The inverse of this averaged reading time was used as the *reading skills measure*; this way the measure furnishes an intuitive value associated with speed: a higher value represents a more skilled reader.

2.2.3. Data Analysis

The data analysis was conducted in the R programming environment (R Core Team, 2013), using either linear mixed-effects model (LMM; Pinheiro and Bates, 2000) or generalized linear mixed-effects models with a binomial link function to the response data (GLMM). Both are regression models that include both fixed effects (such as predictors) and random effects, and they are available in the package *lme4* (Bates et al., 2015). Since LMMs minimize the false positives when they include the maximal random effects structure justified by the design (Schielzeth and Forstmeier, 2009; Barr et al., 2013), both LMMs and GLMMs were fit following this guideline. However, the random effects structure was simplified by removing the correlations, since the models either did not converge or the correlation between variance components could not be estimated.

For large samples, the t distribution approximates the normal distribution and an absolute value of t larger than 2 indicates a significant effect at $\alpha = 0.05$. For all the models presented in the study, covariates such as WMC and reading skills were scaled and centered.

3. Experiment 1

3.1. Method

3.1.1. Participants

Seventy-six subjects aged between 17–42 years old (mean 24.1 years) participated in this experiment in Buenos Aires, Argentina. All participants were native speakers of Spanish and were naïve as to the purpose of the study. Five participants were excluded from the analysis: two participants had reading glasses that prevented an adequate calibration of the eye-tracker, two performed poorly in the mathematical task of the operation span test (with less than 70% accuracy), and another subject reported that she consciously re-read every sentence.

Partial-credit unit scores (Conway et al., 2005) for the operation span test measuring WMC of the remaining 71 participants ranged between 0.232–0.801 with an average of 0.543 (*SE*: 0.013).

Average character speed for the rapid automatized naming task for measuring reading skills ranged between 1.44–3.72 characters/second with an average of 2.54 (*SE*: 0.06) characters/second.

3.1.2. Stimuli

The stimuli for this experiment consisted of 48 items with three conditions (place of attachment) similar to example (5). Each participant read the 48 items together with 120 unrelated sentences (72 were experimental items of two unrelated experiments and 48 sentences were filler sentences) in an individually randomized order. The 144 experimental sentences (48 items in three conditions each) were presented in Latin square design. In order to ensure that participants had paid attention to the sentences, a true-or-false comprehension task was presented after half of all trials in the experiment, including fillers. Half of these statements were true and half false. For the sentences in (5), for example, the statement was false and it was the following: *El gerente fue despedido por equivocación.* "The manager was fired by mistake." The statements following other experimental sentences focused on different aspects of the stimuli: the participants (such as "Jose fired someone."), the action ("The manager hired someone."), the setting of the action (such as "Someone was fired on purpose."), etc. We chose to use true-or-false statements instead of yes-no questions in order to avoid long and unnatural questions.

3.1.3. Procedure

Participants performed the eye-tracking task after having completed a rapid automatized naming task and an operation span task. Before the eye-tracking experiment began, each participant was instructed to read for comprehension in a normal manner and had a practice session of seven sentences. Eye movements were recorded using an EyeLink 1000 eye-tracker, interfaced with a PC. Stimuli were displayed on an 21" monitor. Subjects were seated 65 cm from the computer screen. Viewing was binocular, but only the right eye was recorded. All sentences were displayed on a single line and were presented in twelve points Arial font. At the beginning of each trial, a dot appeared at the left edge of the screen and after participants fixated on this dot, the sentence appeared. Participants had to look at the bottom right corner of the screen to indicate they had finished reading. True-or-false statements appeared randomly for half of the stimuli at this point. No feedback was given as to whether the response was correct or not. After reading half of the sentences, participants took a 10-min break. A calibration procedure was performed at the beginning of the eye-tracking experiment, at the end of the break, and between trials as needed.

3.1.4. Data Analysis

Detection of saccades and fixations was done using a modification of the saccades package developed by von der Malsburg (https://github.com/tmalsburg/saccades), and eye-tracking measures were computed using *em2* package (Logačev and Vasishth, 2013). The appropriate transformation of the dependent variable was determined using the Box-Cox method (Box and Cox, 1964; Kliegl et al., 2010) with the boxcox function in the MASS package (Venables and Ripley, 2002). The log transformation was suggested as the most appropriate transformation.

3.2. Results

3.2.1. Comprehension Accuracy

Participants answered correctly on average 80% (*SE*: 1) comprehension probes of all trials, and 82% (*SE*: 1) of the trials belonging to the experiment. The comprehension accuracy for the experimental trials ranged between 58 and 100%, while the 25th, 50th, and 75th quartiles were 75, 83, and 90% respectively. In addition, a GLMM showed that WMC was a significant predictor of accuracy (higher capacity led to greater accuracy); Coef $= 0.21$, $SE = 0.10$, $z = 1.98$, $p = 0.048$.

3.2.2. Eye-Tracking Measures

Reading times were inspected at three regions of interest: the first critical region (auxiliary verb "había"), second critical region (participle form of the verb), and spillover region (P). We used successive differences contrast coding to test the predictions of the different accounts: VP2 (coded as 1) against VP1 (coded as -1) and VP3 (coded as 1) against VP2 (coded as -1). As in Vasishth and Drenhaus (2011), we found effects in the critical regions only in dependent measures related to re-reading; in the spillover region, we found effects only for total fixation time, consistent with Levy and Keller (2013). We provide the analysis of regions of interest for first-pass regression probability, re-reading probability and total fixation time. As defined in Vasishth and Drenhaus (2011), first-pass regression probability at a word is the probability of the eye moving leftwards after this word was fixated at least once; re-reading probability for a word is the probability of revisiting that word after having having made a first-pass.

After inspecting each LMM with total fixation time as dependent variable, we removed 0.12% of the data in order to keep the residuals normally distributed; the results of the model were virtually the same without this removal. Below we report only statistically significant effects.

3.2.2.1. First critical region (auxiliary verb "había")

We found a WMC and VP2-VP1 interaction for first-pass regression probabilities (Coef $= -0.38$, $SE = 0.17$, $z = -2.17$, $p = 0.03$) showing that as WMC increases, the probability of a regression at the auxiliary verb decreases for condition VP2 in comparison with VP1 (as shown in **Figure 2**).

Since we did not find evidence of more facilitation in VP3 in comparison with VP2, we also fitted a separate model that included the VP3-VP1 contrast. We found a decrease in re-reading probability for VP3 in comparison with VP1 (Coef $= -0.28$, $SE = 0.12$, $z = -2.40$, $p = 0.016$).

3.2.2.2. Second critical region (participle form)

As in the first critical region, we found a decrease in re-reading probabilities for VP3 condition in comparison with VP1 (Coef $= -0.20$, $SE = 0.10$, $z = -1.99$, $p = 0.047$).

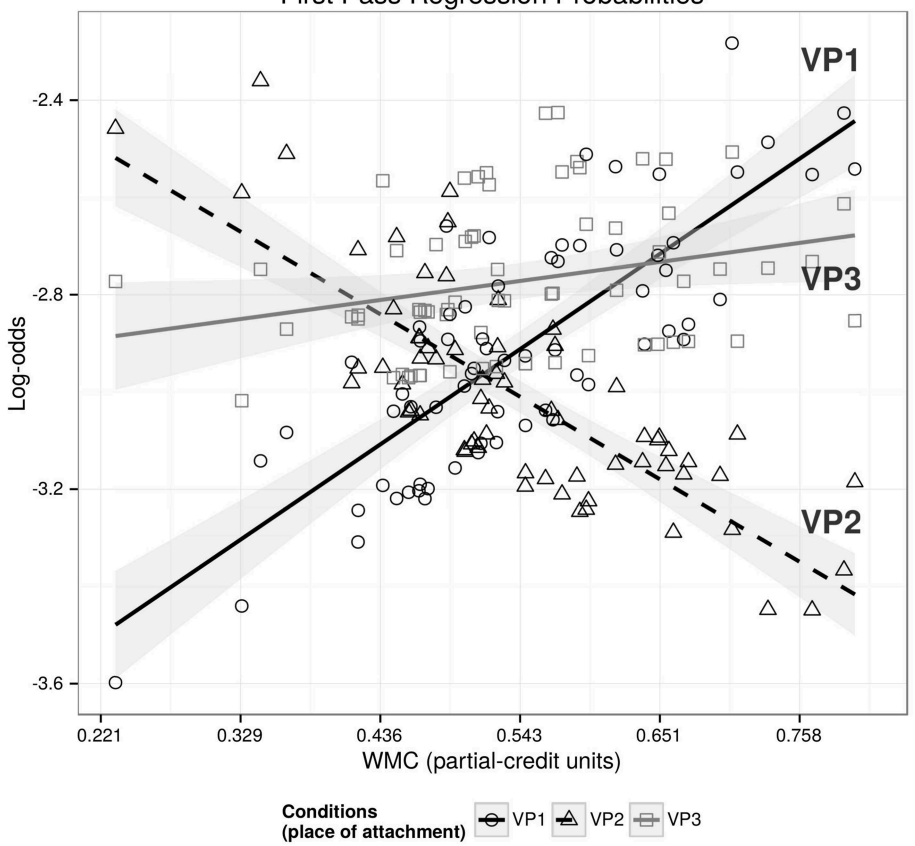


FIGURE 2 | The figure depicts the partial effects on first pass regression probabilities in log-odds scale for the contributing factors condition, WMC, and their interaction; random factors variance and effects due to reading skills were removed from the dependent variable using the remef function (Hohenstein and Kliegl, 2013).

3.2.2.3. Spillover (preposition)

We found a significant speedup for VP2 in comparison with VP1 for total reading time (Coef $= -0.06$, $SE = 0.03$, $t = -2.07$), and an unpredicted interaction between reading skills and VP2-VP1 (Coef $= 0.09$, $SE = 0.03$, $t = 2.86$) showing that as reading skills increases, total reading times at the spillover for condition VP2 increase in comparison with condition VP1.

3.3. Discussion

The central finding in the eye-tracking study is that individual differences associated with working memory have an impact in parsing sentences with long-distance dependencies. When the extra material modifies the intermediate VP (VP2), results for first pass regression probabilities for the critical region are consistent with the idea that expectations play a dominant role when the individual capacity of the participants is large enough to overcome the memory-driven locality effects (see **Figure 2**). That is, locality effects may become antilocality effects when WMC is large enough. This pattern can be explained by a memory account acting together with the expectation account. However, from this pattern alone it is not clear whether DLT or the activation-based model best explain the data. The predictions of DLT are based solely on dependency length, entailing that VP2 should be fully aligned with VP3 (see **Figure 1E**). The activation-based model predicts facilitation when the extra material is attached to the head verb, that is, facilitation for VP3 in comparison with VP2 (while sharing the same lower asymptote for extremely high WMC; see **Figure 1F**). At least for first pass regression probabilities for the critical region, it is unclear where VP3 condition stands: there is no significant facilitation in comparison with VP1 as all the described accounts would predict.

However, the study does provide some evidence for a differential effect that depends on *where* the extra material is attached, and not just on the linear distance of the dependency (as DLT and expectation account would predict). When the extra material is part of the same VP as the subcategorizing head verb (VP3), re-reading probabilities show facilitation compatible both with expectations and with the preactivation of the subcategorizing verb and similar to the evidence from SOV languages (Konieczny, 2000; Konieczny and Döring, 2003; Vasishth, 2003; Vasishth and Lewis, 2006; Levy and Keller, 2013). The fact that

facilitation occurs only for VP3 condition in comparison with the short dependency condition VP1, and not when the extra material modifies the intermediate VP (VP2), provides some indirect evidence indicating differential facilitation between VP2 and VP3 as predicted by the activation account.

As mentioned before, one of the main differences between the predictions depicted in **Figure 1** and our results is the status of VP3 condition: The facilitation of VP3 in comparison with the baseline VP1 appears in a different measure (re-reading instead of regression probabilities) than the facilitation of VP2 condition (in comparison with VP1), and it "spilled over" to the second critical region. In addition, and in contrast with VP2, the facilitation did not depend on the WMC of the participants.

Regarding the differences in the eye-tracking measures and spillover, the effect of adding preverbal material may have been more complex than hypothesized. The preverbal material may have added a new retrieval process at the head and thus overshadowed any facilitation caused by increased expectations. Furthermore, the appearance of the facilitation in different measures can be accounted for by assuming that facilitation due to preactivation, and facilitation due to increased expectations depend on different underlying mechanisms resulting in qualitatively different behavioral consequences in reading (Staub, 2010).

We can speculate that the difference in processing difficulty between VP3 and VP1 did not depend on WMC in our results because at VP3 condition, the facilitation has already reached a bottom asymptote (the minimum re-reading probability given the complexity of the stimuli; see **Figure 2F**). This lack of an effect of WMC on the facilitation might presumable be because of our relatively homogeneous pool of participants, who did not display a big enough variance in their WMC.

4. Experiment 2

This experiment is a replication of Experiment 1 using self-paced reading methodology. Even though eye-tracking experiments provide a more natural setting than self-paced reading, eye-tracking allows participants reading strategies that are absent in self-paced reading, such as skipping words and re-reading. Moreover, since it is possible to calculate many different eye-tracking measures, the chance of getting a false positive (a Type I error) goes up due to the multiple testing problem. Thus, one important motivation for the self-paced reading experiment was to determine whether the previous results were robust. A second motivation was to attempt a replication of the eye-tracking result using a different method. The absence of replication has been recognized as a major problem in psychology and related areas (Asendorpf et al., 2013).

4.1. Method

4.1.1. Participants

Eighty subjects aged between 18–44 years (mean age 25 years) participated in a self-paced reading experiment in Argentina. The first 34 subjects participated in Buenos Aires and the rest in Mendoza. All participants reported to be native speakers of Spanish and were naïve to the purpose of the study. Only one participant was excluded from the analysis, since s/he reported, after the experiment had been completed, that s/he suffered from a mental disorder related to memory.

Partial-credit unit scores for the operation span test measuring WMC of the remaining 79 participants ranged between 0.373–0.882 with an average of 0.631 (*SE*: 0.015). Average character speed for the rapid automatized naming task for measuring reading skills ranged between 1.60–3.45 characters/second with an average of 2.40 (*SE*: 0.05) characters/second.

4.1.2. Stimuli

The stimuli for this experiment consisted of 36 items similar to the items of Experiment 1, but with an extended spillover region. This extra region was included in case the self-paced reading task may delay the effects seen in the eye-tracking experiment.

Similarly to Experiment 1, each participant read the 36 items together with 176 unrelated sentences (120 were experimental items of three unrelated experiments and 56 sentences were filler sentences) in an individually randomized order after six practice trials; and the stimuli were presented in a Latin square design. A true-or-false comprehension task was presented after 65% of all trials in the experiment, including fillers. As in the previous experiment, the statements focused on various aspects of the stimuli, and the proportion of true and false statements was balanced.

4.1.3. Procedure

Subjects were tested individually using a PC. Participants completed the three tasks at their own pace: First, they performed a rapid automatized naming task, second, an operation span task, and finally, a self-paced reading task (Just et al., 1982).

Before the self-paced reading task began, each participant was instructed to read for comprehension in a normal manner and had a practice session of six sentences. All sentences were displayed on a single line and were presented in 18 pt Arial font using Linger software (http://tedlab.mit.edu/~dr/Linger/). In order to read each word of a sentence successively in a moving window display, participants had to press the space bar; then the word seen previously was masked and the next word was shown. At the end of some of the sentences, participants had to answer whether a certain statement related to the experimental item was true or false. No feedback was given as to whether the response was correct or not. Twice during the self-paced reading task, a screen announced the number of sentences read so far and invited the participants to take a break.

4.1.4. Data Analysis

The appropriate transformation of the dependent variable according to the Box-Cox method (Box and Cox, 1964) was the inverse transformation. We used $(-10^5/RT)$ to improve the readability of the models (a positive t-value for $-10^5/RT$ corresponds to a positive t-value of the untransformed measure RT).

4.2. Results

4.2.1. Comprehension Accuracy

Participants answered correctly on average 77% (SE: 1) comprehension probes of all trials, and 70% (SE: 1) of the trials belonging to the experiment. The comprehension accuracy for the experimental trials ranged between 46 and 88%, while the 25th, 50th,

and 75th quartiles were 62, 71, and 77% respectively. As in Experiment 1, a GLMM showed that WMC was a significant predictor of accuracy, with higher capacity leading to greater accuracy; Coef = 0.15, $SE = 0.07$, $z = 2.02$, $p = 0.043$.

4.2.2. Reading Times

We compared reading times at the same three regions of interest as in Experiment 1, using the same successive differences contrast coding. Since the effects appeared in the same regions as in Experiment 1, the added spillover regions were omitted from the analysis.

We removed 0.18% of the data in order to keep the residuals normally distributed; the results of the model were virtually the same without this removal.

4.2.2.1. First critical region (auxiliary verb "había")

For this region, including a quadratic term for WMC was justified according to a model comparison; an anova comparison of models based on a Chi-squared test yielded: $\chi^2_3 = 10.7$, $p = 0.013$.

The main results for this region are displayed in **Table 3**. Consistent with the indirect evidence in Experiment 1 (recall that for re-reading probabilities, we found significant facilitation in VP3 vs. VP1, but not in VP2 vs. VP1), we found a differential facilitation between VP2 and VP3: the critical region was read faster in VP3 in comparison with VP2. We also found a significant interaction between WMC^2 and VP2-VP1 showing an inverted U-shaped effect of WMC on reading times (see **Figure 3**), that is, shorter reading times in VP2 vs. VP1 for low and high-WMC than for mid-WMC. In other words, speedups were seen in low as well as high-capacity readers, but not in medium-capacity readers. An interaction between WMC and VP2-VP1, even though non-significant, suggests that the speedup may be stronger for high-WMC than for low-WMC. We also found significant interactions between WMC and VP3-VP2, and between WMC^2 and VP3-VP2. Due to these findings, we also fitted a separate model that included the VP3-VP1 contrast. This new model revealed that the effect of WMC was only relevant in relation to VP2 (as can be seen in **Figure 3**).

As expected, subjects with higher reading skills scores tended to have shorter reading times, but we also found an unpredicted interaction of reading skills with VP3-VP2 showing that as the reading skill score increases, reading times at the critical region get increasingly shorter for VP3 in comparison with VP2.

4.2.2.2. Second critical region

For these regions a quadratic term for WMC was not justified, so we report the main findings for the model including only linear terms for WMC and reading skills. As in the previous region, there was a speedup for VP3 in comparison VP2, which was independent of WMC (Coef = −7.17, $SE = 3.97$, $t = -1.81$). The results showed reading skills to be significant as well: subjects with a higher score tended to have shorter reading times (Coef = −30.11, $SE = 7.46$, $t = -4.03$).

TABLE 3 | Summary of the fixed effects in the LMM with a quadratic term of WMC for reading times at first critical region in Experiment 2.

Predictor	Coef	*SE*	*t*
VP2 vs. VP1	7.0	5.7	1.2
VP3 vs. VP2	−18.0	5.8	−3.1*
WMC	−2.5	6.6	−0.4
WMC^2	−5.6	6.3	−0.9
RS	−25.6	6.6	−3.9*
VP2 vs. VP1 : WMC	−7.3	4.2	−1.7
VP3 vs. VP2 : WMC	9.3	4.2	2.2*
VP2 vs. VP1 : WMC^2	−11.2	4.0	−2.8*
VP3 vs. VP2 : WMC^2	10.3	4.0	2.6*
VP2 vs. VP1 : RS	6.5	4.1	1.6
VP3 vs. VP2 : RS	−9.4	4.1	−2.3*

* *indicates a significant effect at* $\alpha = 0.05$.

4.3. Discussion

The main results of the self-paced reading study are an inverted U-shaped effect of WMC on reading times for the first critical region for the condition where the extra material modified the VP (VP2) in comparison with the condition with the short dependency (VP1), and a speedup at the two critical regions when the extra material modified the VP that contained the subcategorizing verb (VP3) in comparison with when it modified the intermediate VP (VP2).

The study thus shows that individual differences associated with working memory have an impact in reading strategies for processes associated with build-up of expectations and retrieval. Moreover, this study provides more evidence for a differential effect that depends on whether the VP that contains the head of the dependency is modified, as predicted by the activation-based model, but not by DLT and the expectation account.

We found that when the extra material modifies the VP where the dependency is completed (VP3), participants showed a speedup in comparison with the condition where the extra material modifies the intermediate verb (VP2). Since the dependencies in both conditions had the same length, this experiment provides further evidence for facilitation because of preactivation of the subcategorizing verb as predicted by the activation-based account (Vasishth and Lewis, 2006; and consistent with **Figure 1F**).

The data also showed a surprising inverted U-shaped interaction between WMC and VP2-VP1 conditions. An analogy to exam-taking may explain how two different underlying causes may lead to a process finishing early: students leave an examination hall early either because they do not have the resources (knowledge, skills, etc.) to complete the exam (i.e., they effectively give up), or because they have the resources in excess and can complete the exam quickly. Similarly, there may be two different reasons for the shorter RTs: Low-WMC subjects may read fast because they have done a shallow parse due to not having enough computational resources (probably using a good-enough parsing heuristic see: Ferreira et al., 2002; Ferreira and Patson, 2007), while high-WMC participants may read fast because they did a complete parse and still had enough resources to take advantage of the build-up of expectations (see the right part of **Figures 1E,F**). Medium-WMC participants, however, may have built a complete parse but either did not have enough resources

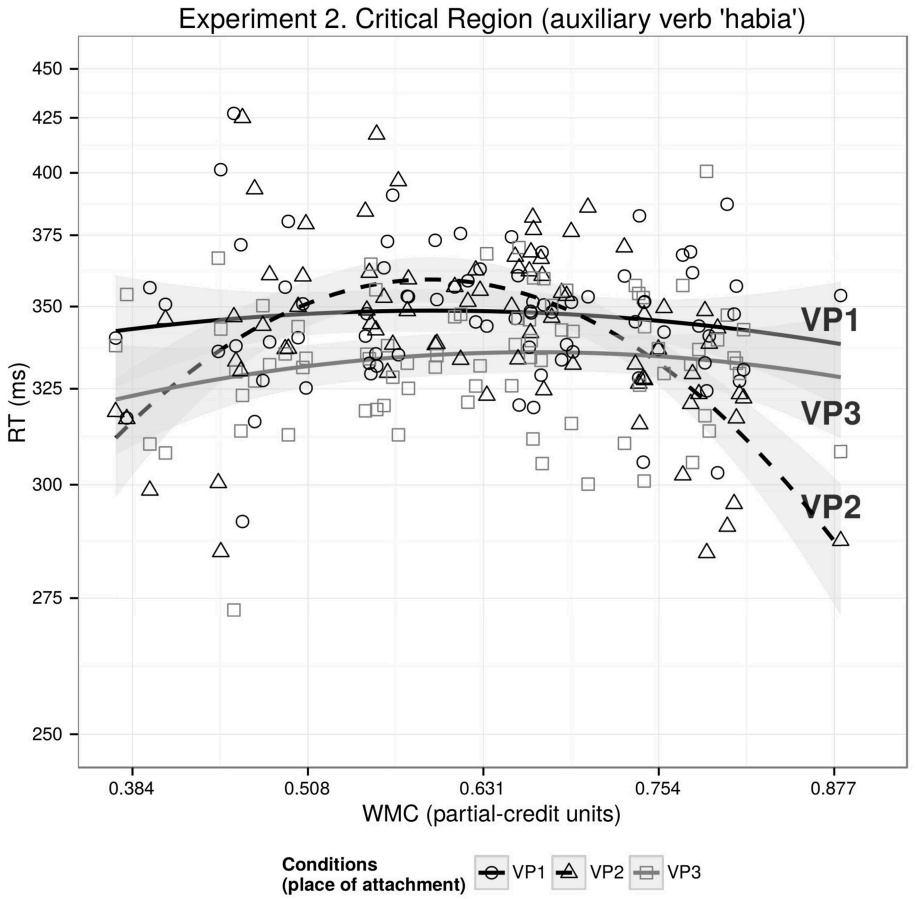


FIGURE 3 | The figure depicts the partial effects on the transformed -1/RT scale for the contributing factors condition, WMC, WMC^2 and their interaction; random factors variance and effects due to reading skills were removed from the dependent variable using the remef function (Hohenstein and Kliegl, 2013).

available for the build-up of predictions of the upcoming head, or the memory-driven locality effect offset the facilitation due to expectations. The difference between this study and the eye-tracking study may be due to the increased task demands of self-paced reading and the impossibility of making regressive saccades. This difference is also evident from the lower comprehension accuracy in self-paced reading in comparison with eye-tracking (70 vs. 82%).

As in the previous experiment, the speedup at the critical region depends only on WMC when the dependent-head distance is increased without a modification of the VP that contains the head (VP2-VP1), while the speedup is independent of WMC when distance is increased by a modification of the VP that contains the head (VP3-VP1). As it was shown in **Figures 1B,F**, it is expected that a facilitation that depends on WMC will have a bottom asymptote since the duration of the reading times cannot be zero and presumably there is a minimum time needed (for recognizing the word, pressing the space bar, etc). Since the activation-based model predicts stronger facilitation for VP3 in contrast to VP2, it also predicts that the effect of WMC on VP3 will reach the bottom asymptote earlier than on VP2 (and thus showing a "flat" WMC effect if all the participants have a relatively high WMC). It should be noted that for the extremely high values of WMC, however, the speedup of VP2 is stronger than of VP3, which is not predicted by the activation-based model (and neither by the expectation account or DLT). However, this is true for a few subjects, and it may be due to the lack of data for the extreme values of WMC.

In addition, the results showed that the facilitation due to preactivation (VP3 vs. VP2) "lasts longer." This is in some way parallel to the findings of Experiment 1, where the facilitation at VP3 condition (this time in comparison with VP1) appeared both in a different measure (re-reading instead of first pass regression probabilities) and it spilled over to the second critical region.

5. General Discussion

A major contribution of this paper is the finding that participants' WMC affects the processes involved in the dependency resolution. Even though recent research has shown that in some cases the relevant measure of individual difference to explain reading strategies is related to experience with language

rather than memory (vocabulary size in Prat, 2011; reading speed in Traxler et al., 2012), by taking into account the results of a rapid automatized naming task, which reflects experience with language, the current study showed WMC as measured by the operation span task to be a fruitful index of individual differences (at least for dependency resolution). Even though long-distance dependency completion is widely assumed to depend on the available working memory (but see Waters and Caplan's approach to working memory: Waters and Caplan, 1996; Caplan and Waters, 1999; Waters and Caplan, 2001), this is, to our knowledge, the first study showing that WMC modulates the reading times and regressions at the head of long-distance dependencies, as predicted by both DLT (Gibson, 2000) and the activation-based model (Vasishth and Lewis, 2006). The findings are consistent with the recent work of Caplan and Waters (2013). In this work, the authors argue that working memory supports retrieval in points of high processing load, which are identified by regressive saccades and longer self-paced reading times that enable better comprehension. In addition, our results show the added value of analyses that take individual variation into account instead of averaging over the data of participants (among others: Underwood, 1975; Brown and Heathcote, 2003; Traxler et al., 2005; and more recently Kliegl et al., 2011; Traxler et al., 2012; Payne et al., 2014).

The results of Experiments 1 and 2 together suggest that increasing the distance of the dependency affects the parsing of the head of the dependency in different ways, depending on whether the intervening material modifies the upcoming head or not. As predicted by the activation-based model (Vasishth and Lewis, 2006) but not by DLT or the expectation account, the facilitation is stronger when the intervening material modifies the upcoming head even when the length of the dependency is the same.

The increase of expectation-based facilitation at the subcategorizing head depends on adding lexical material that helps to sharpen predictions on the location of the upcoming head. However, the increase of lexical material also has its cost in memory processes, so expectation-driven facilitation seems to be noticeable as a speedup or as the decrease of regressions for participants with enough resources to overcome the difficulties caused by adding the extra lexical material (at least when the added facilitation due to the preactivation of the subcategorizing VP is absent). This predicts a monotonic effect of WMC, namely, when distance is increased, the difficulty for low-WMC is reduced as WMC increases, which turns into facilitation for high-WMC. While that was the case for our eye-tracking study (Experiment 1), this interaction was more complex than predicted for the self-paced reading task (Experiment 2).

Expectation-driven facilitation reduced the probability of regressions depending on the WMC of the participants of our eye-tracking study (Experiment 1), so that locality effects decreased as WMC increased until they became increasing antilocality effects. However, for the participants of the self-paced reading task (Experiment 2), the effect of WMC had an inverted-U shape, showing speedups in comparison with the short dependency condition for both high- and low-capacity readers. Since WMC predicted better comprehension accuracy, we assume that there are different underlying processes behind these two speedups, and only high-WMC readers are assumed to speed up because their WMC allowed them to parse the sentence and predict the upcoming lexical material. Since locality effects are assumed to be a response to either a memory overload (Gibson, 1998), the use of more computational resources (Gibson, 2000), or higher retrieval costs (Vasishth and Lewis, 2006), theories that predict locality effects would not predict that low-WMC participants would speed up in comparison with mid-WMC readers when the distance between head and argument is increased. In fact, there is ample evidence that proposes that individual differences in WMC reflect limitations in attention allocation for goals, especially in the face of interference or distraction (for a review see Engle, 2002).

There is independent evidence that high working memory load may lead to faster processing; this comes from the self-paced reading studies of Van Dyke and McElree (2006), who found that when subjects were presented with a memory load (a series of words to recall later) prior to reading a sentence, reading times were shorter and comprehension accuracy was lower in comparison with the conditions without the memory load. It seems that when the comprehender is parsing material while being engaged in processes that tax memory, a possible reading strategy is to disengage from the memory load sooner by reading faster. These results are in line with good-enough parsing (Ferreira et al., 2002; for a review: Ferreira and Patson, 2007), which states that the parser is not necessarily trying to achieve a fully specified representation of the sentence and that it might accept a partial or inconsistent representation. Furthermore, the findings converge with studies showing that low-WMC subjects may take less time when ambiguities are present (but they had worst accuracy) than high-WMCs (MacDonald et al., 1992; Pearlmutter and MacDonald, 1995; von der Malsburg and Vasishth, 2013), that they can read superficially enough to draw contradicting conclusions from a text (Oberauer et al., 2006); and that older adults' increase their reliance on heuristic-like good-enough processing to compensate for age-related deficits in WMC (Christianson et al., 2006).

Since this speedup for low-WMC readers is hypothesized to be a response to an incomplete parse of the more memory demanding condition, the speedup should appear together with a trade-off in the accuracy of the dependency completion. However, the true-or-false statements used for testing the participants' comprehension accuracy included many aspects of the stimuli in order to verify that they paid attention to the sentences, but they did not target exclusively whether the dependency was understood. Participants could in principle know whether the statement after the stimulus sentence was true or false, even without a complete understanding of the previous probe sentence. In addition, they could answer wrongly because they misunderstood other aspects of the sentences. The reason for this shortcoming is twofold: First, since most of the previous studies on locality effects examined only on RTs (except for McElree et al., 2003), the design of the experiment was not meant to explore the comprehension accuracy. Second, the nature of the stimuli

made it almost impossible to make short comprehension questions that could test only the dependency; this is so because the comprehension questions would ideally need to test whether sentences such as "it was commented that someone had fired who" are correct. Even though neglecting a deeper analysis of the sentence comprehension task is the normal state of affairs in psycholinguistics, it is a long-standing shortcoming in psycholinguistic research (but see: Christianson et al., 2001; Ferreira et al., 2001).

In sum, we have presented evidence that locality/antilocality effects are modulated by the participants' WMC. However, the exact relationship between WMC and expectations remains elusive. Two possible explanations are: (i) the prediction processes benefit from more WMC being available, as illustrated by **Figure 1D**, such that high-capacity readers may have a more precise expectation of the upcoming material, or they may be able to maintain the predictions for a head generated by the displaced argument (the wh-element in the experiments) for a longer time; or (ii) the prediction processes by themselves are unaffected by WMC (**Figure 1C**), while the stronger facilitation for high-WMC takes place due to the prediction processes being less affected by memory-driven locality effects.

Funding

The work was supported by Minerva Foundation, Potsdam Graduate School, and the University of Potsdam. Mariano Sigman is sponsored by CONICET and the James McDonnell Foundation 21st Century Science Initiative in Understanding Human Cognition—Scholar Award.

Acknowledgments

Thanks to Juan Kamienkowski and Diego Shalom for their assistance in the Integrative Neuroscience Lab. Thanks to Pavel Logačev for the student's exam simile and valuable feedback. Special thanks to the reviewers (Kiel Christianson and Leticia Pablos) for their valuable comments and suggestions.

References

Arnell, K. M., Joanisse, M. F., Klein, R. M., Busseri, M. A., and Tannock, R. (2009). Decomposing the relation between rapid automatized naming (RAN) and reading ability. *Can. J. Exp. Psychol.* 63, 173–184. doi: 10.1037/a0015721

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J., Fiedler, K., et al. (2013). Recommendations for increasing replicability in psychology. *Eur. J. Pers.* 27, 108–119. doi: 10.1002/per.1919

Barr, D. J., Levy, R. P., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: keep it maximal. *J. Mem. Lang.* 68, 255–278. doi: 10.1016/j.jml.2012.11.001

Bartek, B., Lewis, R. L., Vasishth, S., and Smith, M. R. (2011). In search of on-line locality effects in sentence comprehension. *J. Exp. Psychol. Learn. Mem. Cogn.* 37, 1178–1198. doi: 10.1037/a0024194

Bates, D., Mächler, M., Bolker, B., and Walker, S. G. (2015). Fitting linear mixed-effects models using lme4. *ArXiv e-prints* arXiv:1406.5823v1

Box, G. E. P., and Cox, D. R. (1964). An analysis of transformations. *J. R. Stat. Soc. Ser. B* 26, 211–252.

Brown, S. D., and Heathcote, A. (2003). Averaging learning curves across and within participants. *Behav. Res. Methods Instrum. Comput.* 35, 11–21. doi: 10.3758/BF03195493

Caplan, D., and Waters, G. S. (1999). Verbal working memory and sentence comprehension. *Behav. Brain Sci.* 22, 77–94. doi: 10.1017/S0140525X99001788

Caplan, D., and Waters, G. S. (2013). Memory mechanisms supporting syntactic comprehension. *Psychon. Bull. Rev.* 20, 243–268. doi: 10.3758/s13423-012-0369-9

Christianson, K., Hollingworth, A., Halliwell, J. F., and Ferreira, F. (2001). Thematic roles assigned along the garden path linger. *Cogn. Psychol.* 42, 368–407. doi: 10.1006/cogp.2001.0752

Christianson, K., Williams, C. C., Zacks, R. T., and Ferreira, F. (2006). Younger and older adults' "good-enough" interpretations of garden-path sentences. *Discourse Process.* 42, 205–238. doi: 10.1207/s15326950dp4202/6

Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., and Engle, R. W. (2005). Working memory span tasks: a methodological review and user's guide. *Psychon. Bull. Rev.* 12, 769–786. doi: 10.3758/BF03196772

Daily, L. Z., Lovett, M. C., and Reder, L. M. (2001). Modeling individual differences in working memory performance: a source activation account. *Cogn. Sci.* 25, 315–353. doi: 10.1207/s15516709cog2503/1

Daneman, M., and Carpenter, P. A. (1980). Individual differences in working memory and reading. *J. Verbal Learn. Verbal Behav.* 19, 450–466. doi: 10.1016/S0022-5371(80)90312-6

Daneman, M., and Carpenter, P. A. (1983). Individual differences in integrating information between and within sentences. *J. Exp. Psychol. Learn. Mem. Cogn.* 9, 561–583. doi: 10.1037/0278-7393.9.4.561

Daneman, M., and Merikle, P. (1996). Working memory and language comprehension: a meta-analysis. *Psychon. Bull. Rev.* 3, 422–433. doi: 10.3758/BF03214546

Demberg, V., and Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition* 109, 193–210. doi: 10.1016/j.cognition.2008.07.008

Denckla, M. B., and Rudel, R. G. (1976). Rapid 'automatized' Naming (R.A.N): dyslexia differentiated from other learning disabilities. *Neuropsychologia* 14, 471–479. doi: 10.1016/0028-3932(76)90075-0

Engle, R. W. (2002). Working memory capacity as executive attention. *Curr. Dir. Psychol. Sci.* 11, 19–23. doi: 10.1111/1467-8721.00160

Ericsson, K. A., and Kintsch, W. (1995). Long-term working memory. *Psychol. Rev.* 102, 211–245. doi: 10.1037/0033-295X.102.2.211

Fedorenko, E., Gibson, E., and Rohde, D. (2006). The nature of working memory capacity in sentence comprehension: evidence against domain-specific working memory resources. *J. Mem. Lang.* 54, 541–553. doi: 10.1016/j.jml.2005.12.006

Fedorenko, E., Woodbury, R., and Gibson, E. (2013). Direct evidence of memory retrieval as a source of difficulty in non-local dependencies in language. *Cogn. Sci.* 37, 378–394. doi: 10.1111/cogs.12021

Ferreira, F., Bailey, K. G. D., and Ferraro, V. (2002). Good-enough representations in language comprehension. *Curr. Dir. Psychol. Sci.* 11, 11–15. doi: 10.1111/1467-8721.00158

Ferreira, F., Christianson, K., and Hollingworth, A. (2001). Misinterpretations of garden-path sentences: implications for models of sentence processing and reanalysis. *J. Psycholinguist. Res.* 30, 3–20. doi: 10.1023/A:1005290706460

Ferreira, F., and Henderson, J. M. (1993). Reading processes during syntactic analysis and reanalysis. *Can. J. Exp. Psychol.* 47, 247–275. doi: 10.1037/h0078819

Ferreira, F., and Patson, N. D. (2007). The 'good enough' approach to language comprehension. *Lang. Linguist. Compass* 1, 71–83. doi: 10.1111/j.1749-818X.2007.00007.x

Fiebach, C. J., Schlesewsky, M., and Friederici, A. D. (2002). Separating syntactic memory costs and syntactic integration costs during parsing: the processing of German WH-questions. *J. Mem. Lang.* 47, 250–272. doi: 10.1016/S0749-596X(02)00004-9

Fodor, J. D. (1978). Parsing strategies and constraints on transformations. *Linguist. Inq.* 9, 427–473.

Gibson, E. (1998). Linguistic complexity: locality of syntactic dependencies. *Cognition* 68, 1–76. doi: 10.1016/S0010-0277(98)00034-1

Gibson, E. (2000). "The dependency locality theory: a distance-based theory of linguistic complexity," in *Image, Language, Brain: Papers from the First Mind Articulation Project Symposium* (Cambridge, MA), 95–126.

Gibson, E., Desmet, T., Grodner, D., Watson, D., and Ko, K. (2005). Reading relative clauses in english. *Cogn. Linguist.* 16, 313–353. doi: 10.1515/cogl.2005.16.2.313

Gibson, E., and Thomas, J. (1999). Memory limitations and structural forgetting: the perception of complex ungrammatical sentences as grammatical. *Lang. Cogn. Process.* 14, 225–248. doi: 10.1080/016909699386293

Grodner, D., and Gibson, E. (2005). Consequences of the serial nature of linguistic input for sentenial complexity. *Cogn. Sci.* 29, 261–290. doi: 10.1207/s15516709cog0000/7

Hale, J. T. (2001). "A probabilistic earley parser as a psycholinguistic model," in *Proceedings of NAACL 2001* (Morristown, NJ: Association for Computational Linguistics), 1–8.

Hasher, L., and Zacks, R. T. (1988). Working memory, comprehension, and aging: a review and a new view. *Psychol. Learn. Motiv.* 22, 193–225. doi: 10.1016/S0079-7421(08)60041-9

Hohenstein, S., and Kliegl, R. (2013). Semantic preview benefit during reading. *J. Exp. Psychol. Learn. Mem. Cogn.* 40, 166–190. doi: 10.1037/a0033670

Howe, A. E., Arnell, K. M., Klein, R. M., Joanisse, M. F., and Tannock, R. (2006). The ABCs of computerized naming: equivalency, reliability, and predictive validity of a computerized rapid automatized naming (RAN) task. *J. Neurosci. Methods* 151, 30–37. doi: 10.1016/j.jneumeth.2005.07.014

Husain, S., Vasishth, S., and Srinivasan, N. (2014). Strong expectations cancel locality effects: evidence from hindi. *PLoS ONE* 9:e100986. doi: 10.1371/journal.pone.0100986

Janssen, P., and Shadlen, M. N. (2005). A representation of the hazard rate of elapsed time in macaque area LIP. *Nat. Neurosci.* 8, 234–241. doi: 10.1038/nn1386

Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cogn. Sci.* 20, 137–194. doi: 10.1207/s15516709cog2002/1

Just, M. A., and Carpenter, P. A. (1992). A capacity theory of comprehension: individual differences in working memory. *Psychol. Rev.* 99, 122–149. doi: 10.1037/0033-295X.99.1.122

Just, M. A., Carpenter, P. A., and Woolley, J. D. (1982). Paradigms and processes in reading comprehension. *J. Exp. Psychol. Gen.* 111, 228–238. doi: 10.1037/0096-3445.111.2.228

King, J., and Just, M. A. (1991). Individual differences in syntactic processing: the role of working memory. *J. Mem. Lang.* 30, 580–602. doi: 10.1016/0749-596X(91)90027-H

Kliegl, R., Masson, M. E. J., and Richter, E. M. (2010). A linear mixed model analysis of masked repetition priming. *Vis. Cogn.* 18, 655–681. doi: 10.1080/13506280902986058

Kliegl, R., Wei, P., Dambacher, M., Yan, M., and Zhou, X. (2011). Experimental effects and individual differences in linear mixed models: estimating the relationship between spatial, object, and attraction effects in visual attention. *Front. Psychol.* 1:238. doi: 10.3389/fpsyg.2010.00238

Kluender, R., and Kutas, M. (1993). Bridging the gap: evidence from ERPs on the processing of unbounded dependencies. *J. Cogn. Neurosci.* 5, 196–214. doi: 10.1162/jocn.1993.5.2.196

Konieczny, L. (2000). Locality and parsing complexity. *J. Psycholinguist. Res.* 29, 627–645. doi: 10.1023/A:1026528912821

Konieczny, L., and Döring, P. (2003). "Anticipation of clause-final heads: evidence from eye-tracking and SRNs," *Proceedings of ICCS/ASCS* (Sydney, NSW).

Kuperman, V., and Van Dyke, J. A. (2011). Effects of individual differences in verbal skills on eye-movement patterns during sentence reading. *J. Mem. Lang.* 65, 42–73. doi: 10.1016/j.jml.2011.03.002

Kutas, M., DeLong, K. A., and Smith, N. J. (2011). "A Look around at what lies ahead: prediction and predictability in language processing," in *Predictions in the Brain: Using Our Past to Generate a Future*, ed M. Bar (Oxford, UK: Oxford University Press), 190–207.

Levy, R. P. (2008). Expectation-based syntactic comprehension. *Cognition* 106, 1126–1177. doi: 10.1016/j.cognition.2007.05.006

Levy, R. P., Fedorenko, E., and Gibson, E. (2013). The syntactic complexity of Russian relative clauses. *J. Mem. Lang.* 69, 461–495. doi: 10.1016/j.jml.2012.10.005

Levy, R. P., and Keller, F. (2013). Expectation and locality effects in German verb-final structures. *J. Mem. Lang.* 68, 199–222. doi: 10.1016/j.jml.2012.02.005

Lewis, R. L., and Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cogn. Sci.* 29, 375–419. doi: 10.1207/s15516709cog0000/25

Lewis, R. L., Vasishth, S., and Van Dyke, J. A. (2006). Computational principles of working memory in sentence comprehension. *Trends Cogn. Sci.* 10, 447–454. doi: 10.1016/j.tics.2006.08.007

Logačev, P., and Vasishth, S. (2013). *em2: A Package for Computing Reading time Measures for Psycholinguistics*. R package version 0.9.

MacDonald, M. C., and Christiansen, M. H. (2002). Reassessing working memory: comment on Just and Carpenter (1992) and Waters and Caplan (1996). *Psychol. Rev.* 109, 35–54. doi: 10.1037/0033-295X.109.1.35

MacDonald, M. C., Just, M. A., and Carpenter, P. A. (1992). Working memory constraints on the processing of syntactic ambiguity. *Cogn. Psychol.* 24, 56–98. doi: 10.1016/0010-0285(92)90003-K

McElree, B. (2000). Sentence comprehension is mediated by content-addressable memory structures. *J. Psycholinguist. Res.* 29, 111–123. doi: 10.1023/A:1005184709695

McElree, B., Foraker, S., and Dyer, L. (2003). Memory structures that subserve sentence comprehension. *J. Mem. Lang.* 48, 67–91. doi: 10.1016/S0749-596X(02)00515-6

McVay, J. C., and Kane, M. J. (2011). Why does working memory capacity predict variation in reading comprehension? On the influence of mind wandering and executive attention. *J. Exp. Psychol. Gen.* 141, 302–320. doi: 10.1037/a0025250

Misra, M., Katzir, T., Wolf, M., and Poldrack, R. A. (2004). Neural Systems for Rapid automatized naming in skilled readers: unraveling RAN-reading relationship. *Sci. Stud. Reading* 8, 241–256. doi: 10.1207/s1532799xssr0803/4

Oberauer, K., Weidenfeld, A., and Hörnig, R. (2006). Working memory capacity and the construction of spatial mental models in comprehension and deductive reasoning. *Q. J. Exp. Psychol.* 59, 426–447. doi: 10.1080/17470210500151717

Otten, M., and Van Berkum, J. J. A. (2009). Does working memory capacity affect the ability to predict upcoming words in discourse? *Brain Res.* 1291, 92–101. doi: 10.1016/j.brainres.2009.07.042

Pablos, L. (2006). *Pre-Verbal Structure Building in Romance Languages and Basque*. Ph.D. thesis, University of Maryland, College Park.

Payne, B. R., Grison, S., Gao, X., Christianson, K., Morrow, D. G., and Stine-Morrow, E. A. L. (2014). Aging and individual differences in binding during sentence understanding: evidence from temporary and global syntactic attachment ambiguities. *Cognition* 130, 157–173. doi: 10.1016/j.cognition.2013.10.005

Pearlmutter, N., and MacDonald, M. C. (1995). Individual differences and probabilistic constraints in syntactic ambiguity resolution. *J. Mem. Lang.* 34, 521–542. doi: 10.1006/jmla.1995.1024

Peterson, M. S., Kramer, A. F., Wang, R. F., Irwin, D. E., and McCarley, J. S. (2001). Visual search has memory. *Psychol. Sci.* 12, 287–292. doi: 10.1111/1467-9280.00353

Phillips, C., Kazanina, N., and Abada, S. H. (2005). ERP effects of the processing of syntactic long-distance dependencies. *Cogn. Brain Res.* 22, 407–428. doi: 10.1016/j.cogbrainres.2004.09.012

Pinheiro, J. C., and Bates, D. M. (2000). "Linear mixed-effects models: basic concepts and examples," in *Mixed-Effects Models in S and S-PLUS, Statistics and Computing*, eds J. Chambers, W. Eddy, W. Härdle, S. Sheather, and L. Tierney (New York, NY: Springer), 3–56.

Prat, C. S. (2011). The brain basis of individual differences in language comprehension abilities. *Lang. Linguist. Compass* 5, 635–649. doi: 10.1111/j.1749-818X.2011.00303.x

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R version 3.0.2. R Foundation for Statistical Computing.

Schielzeth, H., and Forstmeier, W. (2009). Conclusions beyond support: overconfident estimates in mixed models. *Behav. Ecol.* 20, 416–420. doi: 10.1093/beheco/arn145

Staub, A. (2010). Eye movements and processing difficulty in object relative clauses. *Cognition* 116, 71–86. doi: 10.1016/j.cognition.2010.04.002

Swets, B., Desmet, T., Hambrick, D. Z., and Ferreira, F. (2007). The role of working memory in syntactic ambiguity resolution: a psychometric approach. *J. Exp. Psychol. Gen.* 136:64. doi: 10.1037/0096-3445.136.1.64

Traxler, M. J., Long, D. L., Johns, C. L., Zirnstein, M., Tooley, K. M., and Jonathan, E. (2012). Individual differences in eye-movements during reading: working memory and speed-of-processing effects. *J. Eye Mov. Res.* 5, 1–16. Available online at: http://www.jemr.org/online/5/1/5

Traxler, M. J., Williams, R. S., Blozis, S. A., and Morris, R. K. (2005). Working memory, animacy, and verb class in the processing of relative clauses. *J. Mem. Lang.* 53, 204–224. doi: 10.1016/j.jml.2005.02.010

Trueswell, J. C., Tanenhaus, M. K., and Kello, C. (1993). Verb-specific constraints in sentence processing: separating effects of lexical preference from garden-paths. *J. Exp. Psychol. Learn. Mem. Cogn.* 19, 528–553. doi: 10.1037/0278-7393.19.3.528

Turner, M. T., and Engle, R. W. (1989). Is working memory capacity task dependent? *J. Mem. Lang.* 28, 127–154. doi: 10.1016/0749-596X(89)90040-5

Underwood, B. J. (1975). Individual differences as a crucible in theory construction. *Am. Psychol.* 30, 128–134. doi: 10.1037/h0076759

Unsworth, N., and Engle, R. W. (2007). The nature of individual differences in working memory capacity: active maintenance in primary memory and controlled search from secondary memory. *Psychol. Rev.* 114, 104–132. doi: 10.1037/0033-295X.114.1.104

Van Dyke, J. A., Johns, C. L., and Kukona, A. (2014). Low working memory capacity is only spuriously related to poor reading comprehension. *Cognition* 131, 373–403. doi: 10.1016/j.cognition.2014.01.007

Van Dyke, J. A., and Lewis, R. L. (2003). Distinguishing effects of structure and decay on attachment and repair: a cue-based parsing account of recovery from misanalyzed ambiguities. *J. Mem. Lang.* 49, 285–316. doi: 10.1016/S0749-596X(03)00081-0

Van Dyke, J. A., and McElree, B. (2006). Retrieval interference in sentence comprehension. *J. Mem. Lang.* 55, 157–166. doi: 10.1016/j.jml.2006.03.007

van Rij, J., van Rijn, H., and Hendriks, P. (2013). How WM load influences linguistic processing in adults: a computational model of pronoun interpretation in discourse. *Top. Cogn. Sci.* 5, 564–580. doi: 10.1111/tops.12029

Vasishth, S. (2003). *Working Memory in Sentence Comprehension: Processing Hindi Center Embeddings*. New York, NY: Routledge.

Vasishth, S., and Drenhaus, H. (2011). Locality in german. *Dialogue Discourse* 2, 59–82. doi: 10.5087/dad.2011.104

Vasishth, S., and Lewis, R. L. (2006). Argument-head distance and processing complexity: explaining both locality and antilocality effects. *Language* 82, 767–794. doi: 10.1353/lan.2006.0236

Venables, W. N., and Ripley, B. D. (2002). *Modern Applied Statistics with S, 4th Edn*. New York, NY: Springer.

von der Malsburg, T., and Vasishth, S. (2013). Scanpaths reveal syntactic underspecification and reanalysis strategies. *Lang. Cogn. Process.* 28, 1545–1578. doi: 10.1080/01690965.2012.728232

Vos, S. H., Gunter, T. C., Schriefers, H., and Friederici, A. D. (2001). Syntactic parsing and working memory: the effects of syntactic complexity, reading span, and concurrent load. *Lang. Cogn. Process.* 16, 65–103. doi: 10.1080/01690960042000085

Waters, G. S., and Caplan, D. (1996). The capacity theory of sentence comprehension: critique of Just and Carpenter (1992). *Psychol. Rev.* 103, 761–772. doi: 10.1037/0033-295X.103.4.761

Waters, G. S., and Caplan, D. (2001). Age, working memory, and on-line syntactic processing in sentence comprehension. *Psychol. Aging* 16, 128–144. doi: 10.1037/0882-7974.16.1.128

Wells, J. B., Christiansen, M. H., Race, D. S., Acheson, D. J., and MacDonald, M. C. (2009). Experience and sentence processing: Statistical learning and relative clause comprehension. *Cogn. Psychol.* 58, 250–271. doi: 10.1016/j.cogpsych.2008.08.002

Yuill, N., Oakhill, J., and Parkin, A. (1989). Working memory, comprehension ability and the resolution of text anomaly. *Br. J. Psychol.* 80, 351–361. doi: 10.1111/j.2044-8295.1989.tb02325.x

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2015 Nicenboim, Vasishth, Gattei, Sigman and Kliegl. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 08 March 2016
doi: 10.3389/fpsyg.2016.00280

When High-Capacity Readers Slow Down and Low-Capacity Readers Speed Up: Working Memory and Locality Effects

Bruno Nicenboim[1]*, *Pavel Logačev*[1], *Carolina Gattei*[2] and *Shravan Vasishth*[1]

[1] Department of Linguistics, University of Potsdam, Potsdam, Germany, [2] Grupo de Lingüística y Neurobiología Experimental del Lenguaje, INCIHUSA, CONICET, Mendoza, Argentina

We examined the effects of argument-head distance in SVO and SOV languages (Spanish and German), while taking into account readers' working memory capacity and controlling for expectation (Levy, 2008) and other factors. We predicted only *locality effects*, that is, a slowdown produced by increased dependency distance (Gibson, 2000; Lewis and Vasishth, 2005). Furthermore, we expected stronger locality effects for readers with low working memory capacity. Contrary to our predictions, low-capacity readers showed faster reading with increased distance, while high-capacity readers showed locality effects. We suggest that while the locality effects are compatible with memory-based explanations, the speedup of low-capacity readers can be explained by an increased probability of retrieval failure. We present a computational model based on ACT-R built under the previous assumptions, which is able to give a qualitative account for the present data and can be tested in future research. Our results suggest that in some cases, interpreting longer RTs as indexing increased processing difficulty and shorter RTs as facilitation may be too simplistic: The same increase in processing difficulty may lead to slowdowns in high-capacity readers and speedups in low-capacity ones. Ignoring individual level capacity differences when investigating locality effects may lead to misleading conclusions.

Keywords: locality, working memory capacity, individual differences, Spanish, German, ACT-R

OPEN ACCESS

Edited by:
Matthew Wagers, University of California, Santa Cruz, USA

Reviewed by:
Clinton L. Johns, Haskins Laboratories, USA
Michael Shvartsman, Princeton Neuroscience Institute, USA

***Correspondence:**
Bruno Nicenboim
bruno.nicenboim@uni-potsdam.de

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: *03 May 2015*
Accepted: *12 February 2016*
Published: *08 March 2016*

Citation:
Nicenboim B, Logačev P, Gattei C and Vasishth S (2016) When High-Capacity Readers Slow Down and Low-Capacity Readers Speed Up: Working Memory and Locality Effects. Front. Psychol. 7:280. doi: 10.3389/fpsyg.2016.00280

1. INTRODUCTION

When a reader or hearer is faced with a sentence containing a non-local dependency, (also called long-distance, filler-gap, or unbounded dependency) such as (1), the interpretation of the dependent (*what*) has to be delayed until the reader parses the head of the dependency (*did*). It has been argued that the delay taxes memory processes, and that processing difficulty increases with increasing distance (among others Gibson, 2000; Grodner and Gibson, 2005; Lewis and Vasishth, 2005; Vasishth and Lewis, 2006; Bartek et al., 2011; Husain et al., 2015). This increase in processing difficulty, which is reflected in longer reading times (RTs) at the head of the dependency, is known as a *locality effect* (Gibson, 2000; Lewis and Vasishth, 2005).

(1) Someone asked what the man did last summer.

While the underlying memory processes are subject to debate, theories that predict locality effects are based on the deterioration in *some* memory processes: either an increase in integration and storage costs in Dependency Locality Theory (DLT: Gibson, 2000); or decay and interference in the case of the activation-based theory (Vasishth and Lewis, 2006). Even though there has been evidence against online language processes drawing resources from a common working memory system (Waters and Caplan, 1996; Caplan and Waters, 1999; Waters and Caplan, 2001), in recent work, Caplan and Waters (2013) argue that working memory may support retrievals in points of high processing load. Locality effects may happen in these points of high processing load, which are identified by regressive saccades and longer self-paced reading times that enable better comprehension. The interaction between individual differences in working memory capacity (WMC) and dependency resolution can shed further light on memory-based explanations of locality effects: Differential effects for different capacities can support the assumption that locality-related processing difficulty may in fact be memory based. This is not explicitly stated in DLT, but it is implied since the upper limits on storage and integration cost (Gibson and Thomas, 1999; Gibson, 2000) should depend on WMC. Furthermore, Fedorenko et al. (2006, 2013) found a reduction in performance during long-distance dependency resolution and memory dual tasks, which they interpret as the integration of non-local dependents taxing memory resources.

The relationship between WMC and retrieval processes is more explicit in the activation-based model of sentence processing (Lewis and Vasishth, 2005; Vasishth and Lewis, 2006), which is based on the Adaptive Character of Thought-Rational framework (ACT-R; see for example Anderson et al., 2004). It is assumed that a head verb triggers the retrieval from memory of its non-local dependents using cues such as number, animacy, being a wh-element, and so forth. There is no assumption of serial search in memory, but there is instead a race between the stored items (i.e., the different encoded phrases), with the most highly activated item arriving to the threshold faster and being retrieved. The latency of a retrieval thus depends on the item's level of activation. While the activation of an item decreases with a certain decay rate from the moment of its encoding, retrieval cues are used to improve the chances to identify the "right item" from memory: matching cues boost the activation of an item (while mismatching cues are penalized).

WMC can be integrated into the activation-based model by assuming that it affects the activation of items in memory differentially. One possibility is that WMC affects the decay of information from memory. This has been modeled, for example, in Just and Carpenter's (1992) CAPS, by Byrne and Bovair (1997) to explain errors after an activity that has been completed (such as forgetting the credit card in an ATM); and it has been assumed in sentence processing by, for example, Cunnings and Felser (2013) to explain the differential processing of reflexives. However, it has long been believed that it is not mainly because time passes that information in memory erodes (for a recent example, see Berman et al., 2009). Some of the findings usually associated with decay can be accommodated within *interference-based decay* (Lustig et al., 2009), which is based on the idea that the passage of time increases the likelihood that the features of an item in memory will overlap with those of a noise distribution, making them increasingly difficult to distinguish (see also Oberauer and Kliegl, 2006).

Another possibility is that WMC differentially affects spreading activation, that is, the boost of activation due to matching cues. There are at least two ways in which this could happen. One way could be because WMC modulates the total amount of activation which is shared between matching cues (see for example Cantor and Engle, 1993 for the implementation in a predecessor of ACT-R, and Daily et al., 2001; van Rij et al., 2013 for the implementation in ACT-R of number recall and pronoun resolution respectively). Another way in which WMC could affect spreading activation was suggested by Bunting et al. (2004); in their view, WMC represents susceptibility to interference. Bunting et al.'s experiment showed that individual differences are better represented if low-capacity participants activate more irrelevant cues than high-capacity participants (recall that there is a total amount of activation that is shared between the cues).

If we assume, as ACT-R does, that decay and interference both play a role (and they may be functionally related, see, e.g., Altmann and Gray, 2002), we can schematize locality effects as follows: when a dependent is parsed, it is stored in memory (together with every other phrase parsed so far in the sentence). As the distance between dependent and head increases, the representation of the dependent decays, which translates to a reduction of its level of activation. Since more recent phrases will have a higher level of activation, the correct retrieval of a non-local dependent is possible by using retrieval cues that are derived from the word eliciting the retrieval (the head), together with context and grammatical knowledge (Lewis et al., 2006). Crucially, when the amount of activation available for boosting matching cues decreases or when this activation is shared between more cues, the role of decay due to the increased distance will dominate. This would entail that the role of decay will be more pronounced for low-capacity readers.

Thus, if the source of locality effects is memory based processes (such as the ones predicted by the activation-based model or implicit in DLT), low-capacity readers should show a stronger slowdown than high-capacity ones when dependent-head distance is increased. This prediction is also supported by the following findings: When faced with difficult sentences, the disadvantage of low-WMC readers seems to increase in comparison to high-WMC ones (for garden-path vs. non-garden path sentences: Christianson et al., 2006; for comprehension reaction times in subject- vs. object-relative clauses: King and Just, 1991; Vos et al., 2001). This is also supported by evidence showing that: (a) WMC influences the probabilities of success in integrating information over a distance in a text (Daneman and Carpenter, 1980); (b) WMC is associated with the ability to maintain on-task thoughts (McVay and Kane, 2011); and (c) there is a reduction in performance during long-distance dependency resolution and memory dual-tasks (Fedorenko et al., 2006, 2013). However, this prediction is also based

on the implicit assumption that RTs can be straightforwardly interpreted as indexing difficulty. We will argue that this is the case only when the retrieval of the dependent is successful. We will return to this topic and discuss the specifics of the role of WMC in the general discussion and modeling section.

Increasing dependent-head distance does not always have the same effect. Memory-driven explanations of locality effects are complicated by findings of so-called *antilocality* effects, that is evidence showing that increased distance can result in *faster* reading. For example, several studies on SOV structures (in Hindi: Vasishth, 2003; Vasishth and Lewis, 2006 and in German: Konieczny, 2000; Konieczny and Döring, 2003; Levy and Keller, 2013) showed that increasing the dependent-head distance can produce facilitation at the head of the dependency. However, such facilitation can be explained by increased expectations of the head (Levy, 2008; Levy et al., 2013; but for a memory-based explanation of facilitation see: Vasishth and Lewis, 2006; Nicenboim et al., 2015b). According to the expectation-based account, the primary source of difficulty incurred in processing a word is determined by the surprisal (negative log of the conditional probability) of a word given its context (Hale, 2001). Crucially for current purposes, this account suggests that when the distance of the dependency is increased, the appearance of the predicted head is delayed. As a consequence, the expectation of finding the head that will complete the dependency will increase monotonically. Thus, as the head is more expected, it will be processed more easily when it is encountered.

Importantly, memory- and expectation-based processes are theoretically not incompatible, and recent research (Staub, 2010; Vasishth and Drenhaus, 2011; Levy and Keller, 2013; Levy et al., 2013; Husain et al., 2014; Nicenboim et al., 2015b) shows that they may coexist. However, many of the experimental results in the literature are not easily interpretable, since increasing the distance by adding material between dependent and head systematically changes the sentences, resulting in confounding effects due to the different sentence structures engendered by the distance manipulation.

One aspect of the systematic difference between the sentences manipulated for dependency distance is the change in the linear position of the head. This is especially critical when the design argues for a speedup, since readers tend to speed up as the number of words increases (Ferreira and Henderson, 1993; Boston et al., 2008; Demberg and Keller, 2008); in (2), for example, distance is always confounded with position.

(2) a. SHORT Someone asked **what** the man **did** last summer.

b. LONG Someone asked **what** the man [words that should belong somehow to the sentence] **did** last summer.

The confound between word position and distance has been addressed (see for example: Vasishth and Drenhaus, 2011; Levy and Keller, 2013) by adding the same or similar *words that should belong somehow to the sentence* before the dependency in the short version; compare now (3a) with (3b).

(3) a. SHORT Someone [words that should belong somehow to the sentence] asked **what** the man **did** last summer.

b. LONG Someone asked **what** the man [words that should belong somehow to the sentence] **did** last summer.

Even though the word position confound is controlled, the new problem that arises is that the sentence structure is still consistently changed beyond the distance manipulation. If a difference is found at the head of the dependency *did* in (3), we cannot be sure whether it is a consequence of the distance manipulation or the change in the structure of the sentence. A slowdown (or a speedup) at the verb *did* in (3b) in comparison with (3a) could, in principle, have different alternative explanations. When lexical material is attached to a dependent to increase the dependent-head distance, the dependent that is retrieved in the longer version has also a richer semantic content that may produce a speedup at the verb (Hofmeister, 2007; Hofmeister and Sag, 2010; Hofmeister and Vasishth, 2014). This would be the case if *words that should belong somehow to the sentence* were, for example, a relative clause or a prepositional phrase in (3), so that the extra material is attached to *the man* in the long condition (and to *someone* in the short one). This is also exemplified in (4) from Grodner and Gibson (2005): when the distance is increased, the semantic content of the dependent also changes, namely, *the nurse from the clinic* is retrieved at the verb instead of just *the nurse*. Even though Grodner and Gibson did find locality effects, it does not rule out that the memory-driven locality effects were partially reduced by facilitation due to richer semantic content (and because of increased word position). But alternatively, the slowdown at the verb may have had independent reasons: When the dependent is more complex, it may include several nouns (*nurse* and *clinic* in the Experiment 4) that could cause encoding (Oberauer and Kliegl, 2006) and/or retrieval interference (Van Dyke and McElree, 2006), producing a slowdown at the head verb as well.

(4) Embedded verb conditions from Grodner and Gibson's (2005) experiment 2:

a. **The administrator** who the nurse **supervised** scolded the medic while ...

b. **The administrator** who the nurse from the clinic **supervised** scolded the medic while ...

In addition, there is evidence that preverbal material in the verbal phrase (VP) may cause a speedup at the verb, since the interposed material can help to strengthen the representation of the upcoming head by activating it through modification (as proposed by Vasishth and Lewis, 2006, and more recently Nicenboim et al., 2015b). This would be the case if *words that should belong somehow to the sentence* in (3) were an adverb such as *secretly*, so that the VP that contains the head in the long distance condition is *secretly did*, while it is only *did* in the short one (since *secretly* is attached to *asked* in the short condition). Furthermore, when the distance is increased by any manipulation, expectations may play a role (Hale, 2001; Levy,

2008): Once the reader starts parsing the embedded sentence at *what*, he or she will also start building expectations for the embedded verb; and these expectations will be different for the long and short conditions. In Levy's (2008) study, this is explained by assuming that the reader has knowledge about the grammar of the sentence, that is, he or she knows that the embedded sentence has some verb, but does not know when it will appear. The more constituents within the embedded sentence that have been integrated, the fewer possible choices there are for subsequent constituents. This means that the reader's expectation for the verb should increase as the number of integrated constituents increases. Thus, since the verb *did* in (3b) is assumed to be more expected than in (3a), it is also predicted to be processed faster.

One way to avoid many of the potential confounds and control for the differences in sentence structure is to compare each of the two experimental conditions, such as (5a) and (5b), to baseline conditions without an unbounded dependency, such as (5c) and (5d). Critically, in both short (5c) and long (5d) baseline conditions, a dependent of the verb (e.g., *something*) appears locally in the VP after the verb and remains at the same distance from the verb replacing the wh-element of the unbounded dependency conditions (5a) and (5b). In this experimental design, locality effects appear as an interaction between dependency type (unbounded vs. local, i.e., baseline), and the length of the material added immediately before the head verb (short vs. long). The sentences with local dependencies would act as baselines canceling out other effects that do not depend on the unbounded dependency. For example, if the extra material is attached to the subject of the embedded clause (*the man*), both long (unbounded and local dependency) conditions will have an argument with a richer semantic content that would require more encoding and trigger more expectations for a head verb (since the clause that starts either at the *what* or *that* is longer) than both short (unbounded and local dependency) conditions. Thus, locality effects at the critical region (*did*) would manifest as the difference between long-unbounded and short-unbounded (5b) − (5a) being larger than the difference between long-baseline and short-baseline (5d) − (5c) conditions.

(5) a. SHORT - UNBOUNDED DEPENDENCY
Someone [words that should belong somehow to the sentence] asked **what** the man **did** last summer.

b. LONG - UNBOUNDED DEPENDENCY
Someone asked **what** the man [words that should belong somehow to the sentence] **did** last summer.

c. SHORT - BASELINE (LOCAL DEPENDENCY)
Someone [words that should belong somehow to the sentence] asked if the man **did** *something* last summer.

d. LONG - BASELINE (LOCAL DEPENDENCY)
Someone asked if the man [words that should belong somehow to the sentence] **did** *something* last summer.

In the following experiments, we used this experimental design together with tasks that measure WMC and reading fluency in order to disentangle locality effects from potential confounds, and to find out whether locality interacts with individual differences. We used the operation span task (Turner and Engle, 1989; Conway et al., 2005) to obtain a reliable measure of WMC of our participants. We expected locality effects to be the strongest for readers with the lowest WMC readers, and we expected their magnitude to decrease with increasing WMC. One of the strengths of this type of design is that we can investigate locality effects without a priori commitments about the effect of the systematic change in the syntactic structure, that is, whether the long conditions will show a slowdown or a speedup at the critical region in comparison with the the short ones when we disregard the dependency manipulation.

It has been argued that differences in WMC may reflect differences in language experience or language skills, and not necessarily intrinsic capacity differences (MacDonald and Christiansen, 2002; Wells et al., 2009; Traxler et al., 2012), since WMC tends to correlate with many other reader characteristics.

In fact, while Traxler et al. (2005) found that WMC and syntactic complexity interacted in an eye-tracking experiment, a re-analysis of the data (Traxler et al., 2012) showed that reading speed accounted for more variation in individuals' responses than WMC. According to Traxler et al. (2012), fast readers, who read more often than slow readers, will have greater experience with language; this would in turn make them more sensitive to semantic cues in the syntactic analysis.

In order to obtain an independent measure of reading speed, we included an additional task called rapid automatized naming task (RAN: Denckla and Rudel, 1976). RAN has been shown to capture important variance associated with the processing of rapidly occurring serial information and it has been shown to predict reading speed, comprehension, and other characteristics associated with fluent reading (among others: Kuperman and Van Dyke, 2011; Araújo et al., 2015).

Norton and Wolf (2012) recently reviewed an extensive body of research that led them to consider RAN tasks "as one of the best, perhaps universal, predictors of reading fluency across all known orthographies" (p. 430). Norton and Wolf's view is that this task and reading are seen to require many of the same processes, such as eye saccade control, and the connecting of orthographic and phonological representations. By reading fluency, Norton and Wolf (2012) mean "fluent comprehension" (Wolf and Katzir-Cohen, 2001), that is, "a manner of reading in which all sublexical units, words, and connected text and all the perceptual, linguistic, and cognitive processes involved in each level are processed accurately and automatically so that sufficient time and resources can be allocated to comprehension and deeper thought" (Norton and Wolf, 2012, p. 215). Even though RAN tasks are usually used to study reading development and dyslexia, a few studies have shown that RAN is also predictive of some characteristics of reading fluency for non-college bound participants aged between 16 and 24 (Kuperman and Van Dyke, 2011), for undergrad students (Al Dahhan et al., 2014; Kuperman et al., in press), and for adults aged between 36 and 65 (van den Bos et al., 2002). In addition, some imaging studies performed in young adults have also shown that RAN and reading activate

similar neural networks of neural structures (Misra et al., 2004; Cummine et al., 2015). Even though RAN has been shown to be predictive of online processes associated with word recognition, a recent study (Kuperman et al., in press) argued that RAN may not be predictive of comprehension accuracy, at least for highly proficient population such as college students. However, it may be the case that in situations of high cognitive load, more fluent readers could show an advantage in comparison with less fluent readers. The inclusion of RAN can thus help us to determine whether some participants by virtue of being fluent readers have enough resources for a more efficient use of the retrieval cues and thus overcome more easily locality effects than less fluent readers.

Since most of the evidence from locality effects and most of the evidence from antilocality effects come from SVO and SOV structures respectively, our experiments also verify whether the same account has cross-linguistic validity.

2. EXPERIMENT 1

2.1. Methods

2.1.1. Participants

Seventy-nine subjects aged between 18 and 44 years old (mean 25.2 years) participated in the experiment in Argentina. All participants were native speakers of Spanish and were naïve to the purpose of the study. One additional participant was excluded from the analysis, since s/he reported that s/he suffered from a mental disorder related to memory after the experiment was conducted. Data from this experiment were collected in the same run as the self-paced reading experiment in Nicenboim et al. (2015b): the stimuli from one experiment served as filler sentences for the other experiment.

2.1.2. Stimuli

The stimuli for this experiment consisted of 48 items in Spanish with four conditions following the same logic as in (5) in a two-by-two design: embedded subject length × dependency, as illustrated in (6). The embedded subject length manipulation was created by converting the proper noun of the short condition into a PP that is attached to another NP: (6a vs. 6b, and 6c vs. 6d). The dependency manipulation was created by comparing conditions with an unbounded dependency vs. local dependency (baseline) conditions, so that only the conditions with the unbounded dependencies have shorter or longer dependencies, and the baseline conditions (6c-6d) have similar structures (shorter or longer subjects) but no unbounded dependencies.

(6) a. SHORT - UNBOUNDED DEPENDENCY

La hermana menor de Sofía preguntó **a quién**
The younger sister of Sofia asked **who.ACC**
fue que María **había saludado** en la puerta del
was that María **had greeted** at the door of the
colegio ayer a la tarde.
school yesterday at the afternoon

b. LONG - UNBOUNDED DEPENDENCY

Sofía preguntó **a quién** fue que
Sofia asked **who.ACC** was that
la hermana menor de María **había saludado** en la
the younger sister of María **had greeted** at the
puerta del colegio ayer a la tarde.
door of the school yesterday at the afternoon

c. SHORT - BASELINE

La hermana menor de Sofía preguntó si María
The younger sister of Sofia asked if Maria
había saludado a la prima de Paula en la puerta
had greeted to the cousin of Paula at the door
del colegio ayer a la tarde.
of the school yesterday at the afternoon

d. LONG - BASELINE

Sofía preguntó si la hermana menor de María
Sofia asked if the younger sister of Maria
había saludado a la prima de Paula en la puerta
had greeted to the cousin of Paula at the door
del colegio ayer a la tarde.
of the school yesterday at the afternoon

The 48 experimental items of the current experiment were presented together with 108 experimental items for other experiments and 56 filler sentences. The sentences presented included (i) 36 items with embedded object questions and adverbs in different positions from Nicenboim et al. (2015b); (ii) 48 items with embedded object questions from an unpublished study; (iii) 24 items with object and subject experiencer psychological verbs and different word order (SVO-OVS) from an unpublished study; and (iv) 56 filler sentences with a variety of saying verbs and embedded sentences.

2.1.3. Procedure

Subjects were tested individually using a PC. Participants completed three tasks at their own pace: tests to assess the individual differences in WMC (operation span task: Turner and Engle, 1989; Conway et al., 2005) and in reading fluency (rapid automatized naming: Denckla and Rudel, 1976), and a moving window self-paced reading task (Just et al., 1982).

2.1.3.1. Operation span

Participants took part in an operation span task (Turner and Engle, 1989) using a software developed by von der Malsburg (2015) and used previously in von der Malsburg and Vasishth (2013). Even though variants of the reading (or listening) span task by Daneman and Carpenter (1980) have been used in many psycholinguistic studies, we chose to use the *operation* instead of the *reading* span task, since the latter is likely to measure verbal ability or reading experience as well as working memory capacity (MacDonald and Christiansen, 2002; Conway et al., 2005). We elaborate on this point below.

Even though both reading span and operation span have been defined as measures of verbal working memory (Conway et al., 2005), we think that using the operation span task presents a methodological advantage. The reading span task measures

participants' abilities to do language-processing tasks, such as maintaining the phonological activation for the words in the face of competing demands from sentence processing, and thus it is not surprising that the reading span may be predictive of sentence processing phenomena (MacDonald and Christiansen, 2002). In contrast, the operation span task (described below) is further from language related tasks. And in fact, Turner and Engle (1989) motivation for the use of the operation span was that "A measure of WM should successfully transcend task dependence in its prediction of higher level cognitive functioning. That is, the memory span task could be embedded in a concurrent processing task that is unrelated to any particular skills measure and still predict success in the higher level task" (Turner and Engle, 1989, p. 129). Furthermore, a study of McVay and Kane (2011) showed some critical differences between reading and operation span task. McVay and Kane used among other measures of individual differences three complex span tasks, namely, operation, reading, and spatial span tasks. Even though the three tasks were highly correlated, the reading span task correlated with more reading comprehension tasks (and more strongly) than the operation span.

The procedure of the operation span task test was the following: At a first stage, participants had to judge the correctness of 25 simple equations. During this practice, the reaction time of Equations 10–25 was measured; the average reaction time plus two standard deviations was used as a time-out at the second stage. Having a time-out for every participant ensures that participants that are fast will not have time left to rehearse the items at the next stage of the test. At the second stage, participants had to verify equations and memorize letters (always consonants) that were shown between the equations. After each equation, a consonant was shown for 800 ms; and after a group from three to seven equation-letter successions, participants were instructed to type the letters that had appeared before in their order of presentation. During both parts of the test, participants had to read the equations and letters aloud in order to prevent vocal rehearsal strategies.

As a numeric score of individual working memory, we computed partial-credit unit scores, which indicate the mean proportion of correctly recalled items within the sets (Conway et al., 2005).

2.1.3.2. Rapid automatized naming

Participants' reading fluency was operationalized using rapid automatized naming speed. Subjects that perform this task faster tend to have better reading comprehension scores, faster reading rates and their initial landing position when fixating tends to be closer to the center (among others: Howe et al., 2006; Arnell et al., 2009; Kuperman and Van Dyke, 2011; Araújo et al., 2015). Rapid automatized naming times were measured using a software developed by the first author (https://github.com/bnicenboim/py-ran-task). The procedure of the test was the following: Each subject was instructed to read a series of trials with 50 items; the items were the same set of letters or numbers that were used in Denckla and Rudel (1976): {o, a, s, d, p} and {2, 6, 9, 4, 7}. The first eight trials were composed of letters and the following eight ones of numbers. The items were displayed in five rows of ten columns and were listed in random order. Participants were instructed to start reading aloud as fast as possible immediately after pressing the spacebar, and to press it again immediately after finishing reading aloud the last item. In case they misread, they were instructed to reread only the misread item. The test started with two practice trials to familiarize the participants with the task.

2.1.3.3. Self-paced reading

For the self-paced reading task all sentences were displayed in a single line and were presented in 18 pt Arial font using Linger software (http://tedlab.mit.edu/~dr/Linger/). A true-or-false comprehension task was presented after 65% of all trials in the experiment including fillers to ensure that participants had paid attention to the sentences. The statements focused on various aspects of the stimuli, and the proportion of true and false statements was balanced. For the sentences in the previous example (6) the statement was: *La hermana menor de Sofía preguntó algo.* "The younger sister of Sofía asked something," which was true for the short conditions but false for the long ones. The statements following other experimental sentences focused on different aspects of the stimuli: the participants, the action, the setting of the action, etc. As in Nicenboim et al. (2015b), we chose to use true-or-false statements instead of yes-no questions in order to avoid long and unnatural questions.

2.2. Results

2.2.1. Data Analysis

The data analysis was conducted in the R programming environment (R Core Team, 2015), using hierarchical models (also known as mixed effects or multilevel models) in Stan (Stan Development Team, 2015b) with the R package RStan (Stan Development Team, 2015a). We fit Bayesian rather than frequentist models, which are generally fit with lme4 (Bates et al., 2014; we provide, however, the results of the frequentist models in the Supplementary Material for comparison purposes). First, hierarchical models minimize false positives when they include the maximal random effects structure justified by the design (Schielzeth and Forstmeier, 2009; Barr et al., 2013). However, such maximal frequentist models did not converge for our data and therefore had to be simplified. In contrast, their Bayesian counterpart could be fit in Stan, by using appropriate weakly informative priors for the correlation matrices (so-called LKJ priors). Third, Bayesian hierarchical models solve the multiple comparisons problem since all relevant research questions can be represented as parameters in one coherent hierarchical model (Gelman et al., 2012). This puts more burden on the hierarchical models and shifts point estimates and their corresponding intervals toward each other via "shrinkage" or "partial pooling" (see Gelman et al., 2012, for more details). Fourth, Bayesian procedures provide credible intervals rather than confidence intervals. A 95% credible interval demarcates the range within which we can be certain with probability 0.95 that the true value of a parameter lies (given the data at hand). By contrast, a frequentist confidence interval (CI) is a property of the statistical procedure and not of the parameter. The CI indicates that when the procedure is used repeatedly across a series of hypothetical

data sets (i.e., the sample space), the procedure will yield intervals which contain the true parameter value in 95% of the cases (Hoekstra et al., 2014 and see Morey et al., 2016 for an extreme example of the difference between confidence and credible intervals). Thus, the frequentist CI cannot be used for inference because it tells us nothing about the uncertainty regarding the parameter's value. By contrast, the Bayesian credible interval expresses uncertainty about the parameter.

Another reason for using Bayesian models is that Bayesian procedures allow us to fit virtually any kind of distribution in a straightforward way. Residual RTs in self-paced reading are usually not normally distributed: They are limited on the left by some amount of time (i.e., the shift of the distribution), and they are highly right skewed. RTs can be reciprocal or log-transformed, but these transformations still assume that RTs are defined by their scale (mean) and shape (standard deviation), and they are unshifted (or have a shift of 0 ms). Rouder (2005) raises the concern that restricting the shift to be zero is unreasonable for response times. Unshifted distributions for *reading* times in SPR may also be unreasonable, since they do not take into account that there is a minimal amount of time that takes to read a word and press a button on the keyboard, typically around 150–250 ms. Evidence from distributional models similar to the shifted lognormal shows that shifts are nonzero and vary across participants (see, for example, Logan, 1992; Rouder et al., 2005). If distributions are shifted and analyzed as unshifted lognormal, with increasing shift, estimates of the mean artificially increase, and estimates of the standard deviation artificially decrease; and these artifacts may influence conclusions (Rouder, 2005). We decided to fit models with shifted lognormal distributions not only to avoid anti-conservative conclusions, but also to get more accurate estimates by fitting our data with a model that resembles the process that generates the data. Furthermore, when we compared the shifted lognormal distribution with unshifted distributions such as a reciprocal or a log transformation on the normal distribution, a model ranking according to the Watanabe-Akaike information criterion (or Widely applicable information criterion or WAIC; Watanabe, 2010; Vehtari and Gelman, 2014) favored the model with the shifted lognormal distribution. This may not be the only way to achieve a realistic fit to RTs; however, the shifted lognormal distribution has two key characteristics that are desirable of a RT distribution (Rouder et al., 2008): (i) it has a shift (which is absent in, for example, the ex-Gaussian distribution) and (ii) its error variance increases with mean RT (Wagenmakers and Brown, 2007). In addition, lognormal distributions are ubiquitous in nature, are well understood (Limpert et al., 2001), and are already used in psycholinguistics. We acknowledge that deeper research is needed to evaluate the advantages and disadvantages of different distributions in RTs in self-paced reading (similar to what was done for visual search by Palmer et al., 2011).

Thus we fitted a hierarchical model with a shifted lognormal distribution, allowing the shift to vary by participant. We present the posterior probability of the coefficients being positive given the data and its 95% credible interval. For all the models presented in the experiments, the predictors were sum coded (-1 and 1 for baseline and long dependency, and −1 and 1 for short and long), and covariates WMC and reading fluency were scaled and centered. In order to be able to compare the results across experiments (and regions), we report the estimates of the parameters $\hat{\delta_k}$ that quantify the effect size of each given coefficient k of the mixed model $\hat{\beta_k}$, such that $\hat{\delta_k} = \hat{\beta_k}/\hat{\sigma}$, where $\hat{\sigma}$ is the estimated standard error of the model (as recommended by Rouder et al., 2012). Effect sizes are a dimensionless quantity (Wagenmakers et al., 2010) and depend less on the methodology (self-paced reading, eye-tracking, EEG), the language, the type of participants (students or general population), etc, than the estimates. (We provide the code of the model in the Supplementary Material.)

We checked the convergence of the models after fitting them with eight chains and 2000 iterations, half of which were the burn-in or warm-up phase. In order to assess convergence, we verified that the $\hat{R}$s were close to one, and we also visually inspected the chains (Gelman et al., 2014).

2.2.2. Results of the Individual Differences Measures

2.2.2.1. Operation span

Partial-credit unit scores for the operation span test measuring WMC of the 79 participants had an average of 0.63 ($SE = 0.01$; range 0.37–0.88).

2.2.2.2. Rapid automatized naming

Average character speed for the rapid automatized naming task for measuring reading fluency ranged between 1.60 and 3.45 characters/second with an average of 2.40 ($SE = 0.05$) characters/second. The reciprocal of the averaged reading time was used as the reading fluency measure; this way a higher value represents a more skilled reader.

These two measures were not correlated for the participants of the experiment; $r = -0.04$, CrI (Credible Interval) $= [-0.26, 0.18]$. However, both were moderately correlated with the general accuracy for all the items; WMC: $r = 0.21$, $CrI = [0.00, 0.42]$; reading fluency: $r = 0.29$, $CrI = [0.05, 0.50]$. It should be noted that even though these two measures were not correlated for our subjects, who were mostly university students, it does not mean they are not correlated in the general population. The lack of correlation may be due to the so-called Berkson's paradox (Berkson, 1946), which arises when a specific part of the population is absent (in this case we can assume that people with not enough reading fluency or WMC would not attend college). However, the lack of correlation is informative in that the two measures may be tapping different underlying capacities or skills.

2.2.3. Results of the Self-Paced Reading Experiment

2.2.3.1. Comprehension accuracy

Participants answered correctly on average 77% ($SE = 1$) comprehension probes of the trials belonging to the experiment.

2.2.3.2. Reading times

We fitted a single model for our four regions of interest using Helmert contrasts; see example (7). This type of coding ensures the interpretability of the effects of interest (length, dependency type, WMC, reading fluency, and their interaction) and allows

us to detect a change in the pattern of the effects across the regions. We defined four contrasts that compare each region with the average of the preceding ones: (i) The first critical region (the auxiliary verb "había") is first compared with the precritical region (always a proper noun), then (ii) the second critical region (a participle form of the verb), (iii) the first spillover (a preposition), and finally (iv) the second spillover (a determiner) are compared with the average of their respective preceeding regions; see **Table 1**. In order to account for the correlations between the regions in a single sentence, we included random effects by sentence besides by participants and items as it is usual. We included random intercepts for participants, item and sentences, and by-participants and items random slopes for length, dependency and their interaction (with their correlations).

TABLE 1 | Helmert contrasts used for both experiments.

Precritical	−1	−1	−1	−1
Critical 1	1	−1	−1	−1
Critical 2	0	2	−1	−1
Spillover 1	0	0	3	−1
Spillover 2	0	0	0	4

Figure 1 shows mean RTs for high- and low-WMC readers at each comparable region, while **Figure 2** shows only the locality effects × WMC interaction.

(7) ... preguntó {a quién fue que; si} (la hermana menor
... asked {who.ACC was that; if} (the younger sister

de) | María | había | saludado | {a; en} |
of) | María | had | greeted | {to; in} |
| precritical | critical 1 | critical 2 | spillover 1 |
la | ...
the | ...
spillover 2 |

Observations with RTs under 150 ms and above 5000 ms were removed from the data (3.84%) after checking the residuals of the model. Values below 150 ms are too fast to be reading times, and they are likely to be erroneous taps on the spacebar. If RTs that are too fast are included, the model cannot estimate the appropriate shifts in the distribution (Rouder et al., 2005).

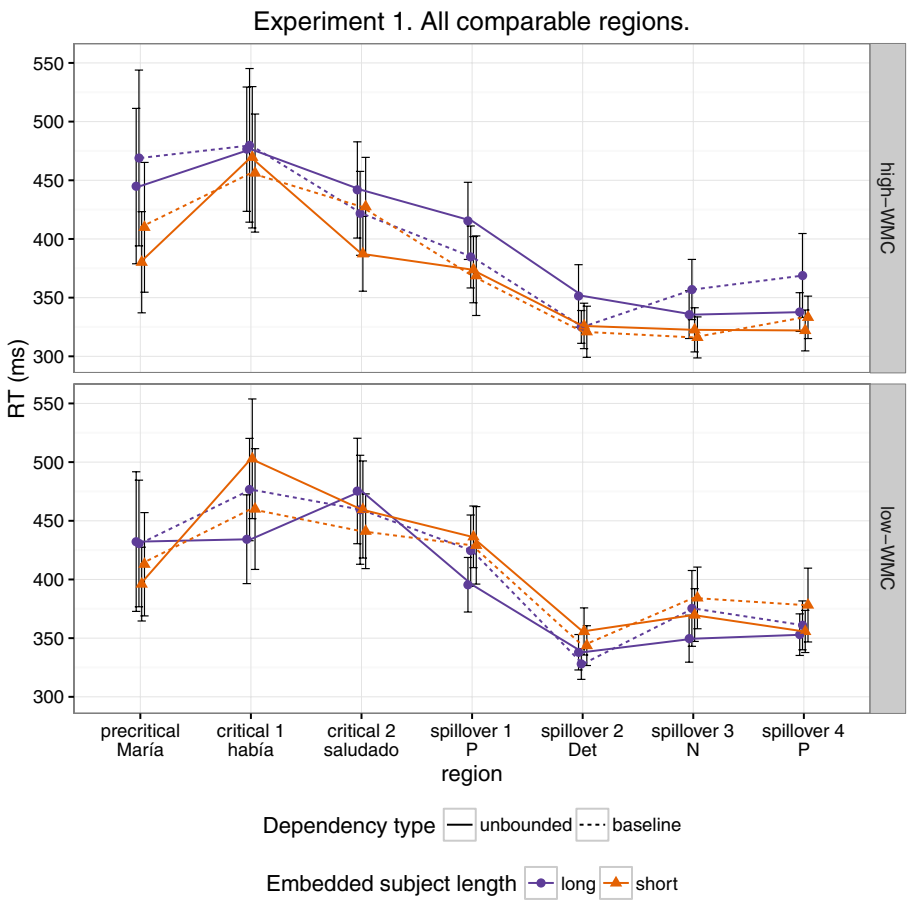


FIGURE 1 | Comparison of mean RTs for high- and low-WMC readers at each comparable region for every condition.

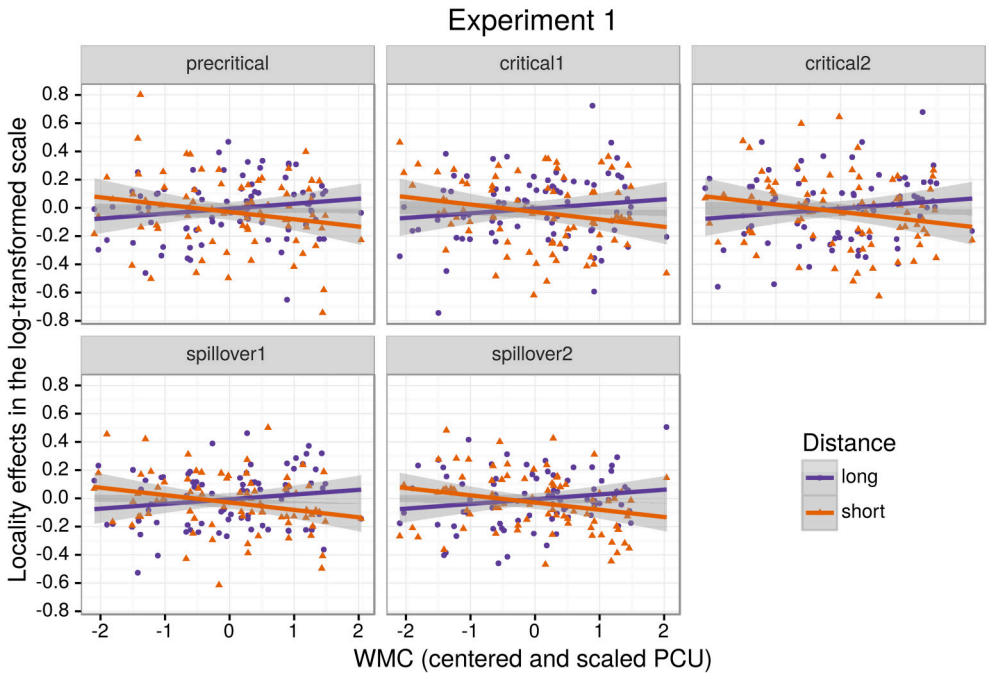


FIGURE 2 | The figure depicts the partial effect for the difference between unbounded conditions and baseline conditions, that is the (anti)locality effects, on the transformed scale of the analysis; random factors variance and effects due to reading skills were removed from the dependent variable (Hohenstein and Kliegl, 2013).

TABLE 2 | Main results for Experiment 1 (Spanish).

Predictor	$\hat{\delta}$	95% CrI		$P(\hat{\delta} > 0)$
Length	0.03	0	0.07	0.97
Dependency	−0.01	−0.04	0.02	0.17
WMC	−0.02	−0.2	0.17	0.42
RF	−0.16	−0.34	0.03	0.04
Length:dependency	0.01	−0.02	0.04	0.74
Length:dependency:WMC	0.04	0.01	0.06	1
Length:dependency:RF	−0.01	−0.04	0.01	0.15

WMC stands for working memory capacity and RF for reading fluency. The first column $\hat{\delta}$ shows the estimated effect size of the coefficients; the next two columns show the 2.5th and 97.5th percentiles of their posterior distribution, that is, where the effect size lies with 95% probability; and $P(\hat{\delta} > 0)$ indicates the posterior probability that each coefficient is positive.

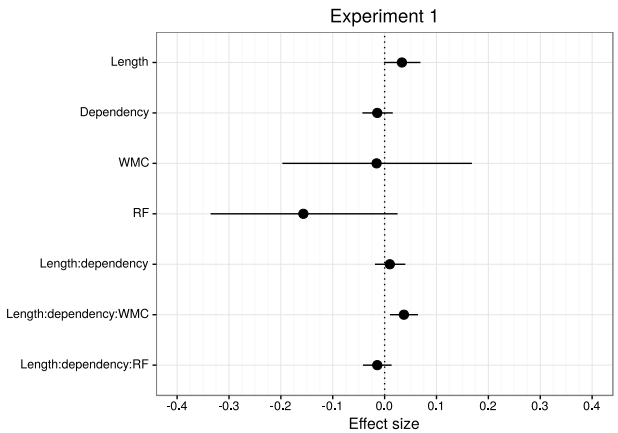


FIGURE 3 | Overview of mean and 95% credible intervals for the effect sizes of the parameters of interest for Experiment 1.

Table 2 and **Figure 3** summarize the main results of the model for the effects of reading fluency, WMC, locality (embedded subject length × dependency), and its interaction with reading fluency and WMC, including the data from all the regions of interest.

In contrast to Null Hypothesis Significance Testing (NHST), where a sharp binary decision is made between "significant" and "non significant" effects, a Bayesian analysis allows us to compute the probability that the coefficient is positive or negative given the data. The 95% Bayesian credible interval has the interpretation that researchers often ascribe mistakenly to frequentist confidence intervals (Morey et al., 2016): it gives the range over which we can be 95% certain, given the data, that the true value of the parameter lies. This statement cannot even be made in NHST, since the true parameter is a point value with no probability distribution. A common way (Kruschke et al., 2012) to interpret the 95% credible interval is to consider an effect to be strong if 0 lies outside the interval. If 0 is included within the interval, there might still be weak evidence for an effect if the probability of the parameter being less than (or greater than) 0 may still be quite large. An example may clarify this: if the probability of the parameter being less than 0 is

0.04, i.e., $P(\hat{\delta} < 0) = 0.04$, this means that there is a 0.96 probability, given the data, that the parameter is negative. Here, it would be odd to say that "there is no effect" given that the posterior probability of the parameter being negative is 0.96. Accordingly, we will interpret the results as follows: if 0 lies outside the 95% credible interval, we assume that the evidence is strong that there is an effect; if 0 is included within the interval but the probability of the parameter being less than or greater than 0 ($P(\hat{\delta} < 0)$ or $P(\hat{\delta} > 0)$, depending on the expected sign of the effect) is high, we will say that there is weak evidence of an effect; and if the probability $P(\hat{\delta} < 0)$ or $P(\hat{\delta} > 0)$ is low, we will conclude that there is no evidence of an effect. For a detailed tutorial on fitting and interpreting Bayesian linear mixed models, see Sorensen et al. (2015).

The model reveals three main findings: (i) As expected, subjects with higher reading fluency scores tended to have shorter RTs (notice that even though zero is included in the credible interval, the effect size is between four and ten times larger than the rest of the effects, and 96% of its posterior probability is below zero); (ii) we did not find the hypothesized locality effects, that is, an interaction between embedded subject length and dependency type regardless of WMC; and (iii) the model shows evidence for an interaction between locality effects and WMC (embedded subject length × dependency type × WMC): For the conditions with unbounded dependencies only, the low-WMC readers showed a slight advantage for the long condition, which was reduced as WMC increased until it became an advantage for the short condition. Even though the interaction between locality effects (embedded subject length × dependency type) and reading fluency showed the predicted direction (smaller locality effects as reading fluency increases), the model shows very weak to no evidence for the effect. We do not report the interactions with the different regions in **Table 1** since they show no evidence that the pattern of the effects varies across regions (including the precritical region as it can be seen in **Figures 1**, **2**). However, nested comparisons where the models were evaluated at the different regions show that the locality × WMC interaction was mainly driven by the precritical, first critical, and spillover regions; see **Table 3**.

It is also worth noting that the length of the embedded subject had an effect on the RTs at the regions of interest, irrespective of the dependency manipulation. This effect would have been confounded with locality in the absence of appropriate baselines. This raises the concern that some of the previous studies that reported a main effect of locality could in principle have been reporting the effect of increasing the complexity of the subject that appeared prior to the verb.

TABLE 3 | Main results for each region of Experiment 1 (Spanish).

Predictor	$\hat{\delta}$	95% CrI		$P(\hat{\delta} > 0)$
PRECRITICAL				
Length	0.04	−0.01	0.08	0.94
Dependency	−0.03	−0.07	0	0.03
WMC	0.01	−0.15	0.17	0.54
RF	−0.19	−0.34	−0.02	0.01
Length:dependency	0.03	−0.01	0.07	0.93
Length:dependency:WMC	0.04	0	0.07	0.99
Length:dependency:RF	−0.04	−0.07	0	0.02
CRITICAL1				
Length	0.04	0	0.08	0.99
Dependency	−0.02	−0.06	0.01	0.1
WMC	0.02	−0.15	0.2	0.6
RF	−0.1	−0.26	0.06	0.1
Length:dependency	0.01	−0.03	0.04	0.66
Length:dependency:WMC	0.05	0.01	0.08	1
Length:dependency:RF	−0.02	−0.05	0.02	0.18
CRITICAL2				
Length	0.03	0	0.07	0.97
Dependency	−0.02	−0.06	0.01	0.1
WMC	0	−0.15	0.15	0.49
RF	−0.2	−0.35	−0.04	0.01
Length:dependency	0.02	−0.02	0.05	0.8
Length:dependency:WMC	0.02	−0.02	0.05	0.82
Length:dependency:RF	−0.01	−0.05	0.02	0.25
SPILLOVER1				
Length	0.02	−0.01	0.06	0.88
Dependency	0	−0.03	0.04	0.6
WMC	−0.06	−0.18	0.07	0.19
RF	−0.14	−0.27	−0.01	0.02
Length:dependency	−0.01	−0.04	0.03	0.38
Length:dependency:WMC	0.04	0.01	0.07	0.99
Length:dependency:RF	0	−0.03	0.03	0.5
SPILLOVER2				
Length	0.01	−0.03	0.05	0.76
Dependency	0.01	−0.03	0.05	0.75
WMC	−0.04	−0.16	0.08	0.24
RF	−0.18	−0.3	−0.06	0
Length:dependency	−0.01	−0.04	0.03	0.36
Length:dependency:WMC	0.01	−0.02	0.05	0.72
Length:dependency:RF	0	−0.03	0.04	0.59

WMC stands for working memory capacity and RF for reading fluency. The first column $\hat{\delta}$ shows the estimated effect size of the coefficients; the next two columns show the 2.5th and 97.5th percentiles of their posterior distribution, that is, where the effect size lies with 95% probability; and $P(\hat{\delta} > 0)$ indicates the posterior probability that each coefficient is positive.

2.3. Discussion

For this experiment, even though we found an effect of embedded subject length, we did not find evidence of locality effects (an embedded subject length × dependency type interaction) across the board. Furthermore, even though an interaction between WMC and locality effects was expected, the interaction was predicted in the opposite direction. We predicted that low-capacity participants would show the strongest locality effects, while counter-intuitively, in our experiment it was the high-WMC participants that showed the strongest locality effects (the largest difference between *(long unbounded − long baseline)* and

(short unbounded − short baseline)), while low-WMC showed antilocality effects; as shown in **Figures 1**, **2**.

This interaction seems counterintuitive because theories that predict locality effects would not predict that high-WMC participants would show stronger locality effects. Locality effects are hypothesized to be a behavioral response to either the use of more computational resources (Gibson, 2000), or higher retrieval costs due to more interference and decay (Lewis and Vasishth, 2005) when the distance between head and argument is increased. However, the speedup of low-WMC readers can be accounted for by adding two intuitively plausible assumptions to memory-based explanations, namely, that low-capacity readers experience retrieval failures more frequently than high-capacity readers, thus leading to unresolved dependencies and an incomplete sentence representation compatible with good enough processing (Ferreira and Patson, 2007); and that retrieval failures are faster on average than complete retrievals. We provide further evidence supporting this claim in the next experiment and the modeling section.

Furthermore, reading fluency correlated with comprehension accuracy (as strongly as WMC) for this experiment, and participants with higher scores in reading fluency tended to read faster the regions of interest. However, we found very weak to no evidence favoring the hypothesis that fluent readers would overcome more easily locality effects than less fluent readers.

While the pattern showing stronger locality effects for high-WMC participants begins at the precritical region (a proper noun that is either the subject or the last part of it) before the verb, memory driven locality effects are predicted to appear no sooner than the verb. However, pre-verbal locality effects have been detected also in Vasishth and Drenhaus's (2011) study, and they also appeared in some degree in the next experiment. This phenomenon will be addressed in the general discussion.

3. EXPERIMENT 2

The second experiment attempts to replicate Experiment 1 using SOV structures in German, in contrast to the SVO structures in Spanish of the previous experiment. The main objective of the second experiment was to verify whether the same account for the findings of Experiment 1 is valid for an SOV language. This is important because SVO structures seem to trigger mostly locality effects at the head verb (among others Grodner and Gibson, 2005; Lewis and Vasishth, 2005; Vasishth and Lewis, 2006; Demberg and Keller, 2008; Bartek et al., 2011), while SOV structures seem to trigger either antilocality effects (Konieczny, 2000; Konieczny and Döring, 2003; Vasishth, 2003; Vasishth and Lewis, 2006; but see Safavi et al., Submitted) or both locality and antilocality (Vasishth and Drenhaus, 2011; Levy and Keller, 2013; Husain et al., 2014). It was therefore important to verify whether the same results can be obtained with the same manipulation irrespective of the OV/VO order.

3.1. Methods

3.1.1. Participants

Seventy-two subjects aged between 17 and 43 years old (mean 24.6 years) were recruited using ORSEE (Greiner, 2004) at the University of Potsdam, Germany. All participants reported to be native speakers of German and were naïve to the purpose of the study. Three other participants had to be removed from the data: one subject answered randomly at the operation span task, another subject answered the comprehension questions at chance level, and the data of a third participant was lost due to technical reasons.

3.1.2. Stimuli

Similarly to Experiment 1, the stimuli for this experiment consisted of 48 items in German with four conditions in a two-by-two design: embedded subject length × dependency (see Example 8).

For this experiment, the embedded subject length manipulation was created by changing the determiner (*die*) of the noun phrase of the short condition with a longer genitive phrase such as *Marias äußerst kaltschnäuzige*, "Mary's extremely uncaring": (8a vs. 8b, and 8c vs. 8d). The dependency manipulation was created as in Experiment 1 by comparing conditions with an unbounded dependency vs. local dependency (baseline) conditions. Thus, conditions (8a–8b) were compared with two baseline conditions (8c–8d) with similar structure, but that lacked the unbounded dependency: The dependent of the verb *jemanden* (someone.ACC) appeared at the same distance of the verb in both short and long baseline conditions.

(8) a. SHORT - UNBOUNDED DEPENDENCY

Marias äußerst kaltschnäuzige Lehrerin fragte,
Mary's extremely uncaring teacher asked
wen die Mutter gestern beim Treffen
who.ACC the mother yesterday at.the meeting
angeschrien hat mit schriller Stimme.
yelled had with shrill voice

b. LONG - UNBOUNDED DEPENDENCY

Die Lehrerin fragte, **wen**
The teacher asked **who.ACC**
Marias äußerst kaltschnäuzige Mutter gestern
Mary's extremely uncaring mother yesterday
beim Treffen **angeschrien hat** mit schriller
at.the meeting **yelled had** with shrill
Stimme.
voice

c. SHORT - BASELINE

Marias äußerst kaltschnäuzige Lehrerin fragte,
Mary's extremely uncaring asked teacher
ob die Mutter jemanden beim Treffen
if the mother someone.ACC at.the meeting
angeschrien hat mit schriller Stimme.
yelled had with shrill voice

d. LONG - BASELINE

Die Lehrerin fragte, ob
The teacher asked if
Marias äußerst kaltschnäuzige Mutter jemanden
Mary's extremely uncaring mother someone

beim Treffen **angeschrien hat** mit schriller
at.the meeting **yelled had** with shrill
Stimme.
voice

The 48 experimental items of the current experiment were presented together with 98 experimental items belonging to experiments from unpublished studies. The sentences presented included (i) 32 items with subject and object relative clauses attached to the subject or the object of sentences; (ii) 42 items with attachment ambiguity involving dative and genitive noun phrases; and (iii) 24 items that contrasted personal and demonstrative pronouns.

3.1.3. Procedure

The procedure was the same as the one used in Experiment 1, with the exception that comprehension questions appeared after every trial in the self-paced reading experiment.

3.2. Results

3.2.1. Results of the Individual Differences Measures

3.2.1.1. Operation span

Partial-credit unit scores for the operation span test measuring WMC of the 72 participants had an average of 0.63 ($SE = 0.02$; range 0.28–0.92).

3.2.1.2. Rapid automatized naming

Average character speed for the rapid automatized naming task for measuring reading fluency ranged between 1.43 and 3.61 characters/second with an average of 2.64 ($SE= 0.06$) characters/second. As in Experiment 1, the reciprocal of the averaged reading time was used as the reading fluency measure.

As in Experiment 1, these two measures were not correlated for the participants of the experiment; $r = 0.02$, $CrI = [-0.23, 0.27]$. In contrast with the previous experiment, only WMC was correlated with the general accuracy for all the items; WMC: $r = 0.42$, $CrI = [0.23, 0.60]$; reading fluency: $r = 0.01$, $CrI = [-0.24, 0.27]$.

3.2.2. Results of the Self-Paced Reading Experiment

3.2.2.1. Comprehension accuracy

Participants answered correctly on average 80% (SE=1) comprehension probes of the trials belonging to the experiment.

3.2.2.2. Reading times

As for Experiment 1, we fitted a single model for our four regions of interest (9) using Helmert contrasts. **Figure 4** shows mean RTs for high- and low-WMC readers at each comparable region, while **Figure 5** shows only the locality effects × WMC interaction.

(9) ... fragte {wen; ob} {die; Marias äußerst
... asked {who.ACC; if} {the; Maria's extremely

kaltschnäuzige} Mutter	gestern	beim	Treffen	
uncaring} mother	yesterday	at.the	meeting	
			precritical	
angeschrien	hat	mit	schriller	...
shouted	have	with	shrill	...
critical 1	critical 2	spillover 1	spillover 2	

As in Experiment 1, RTs under 150 ms and above 5000 ms were removed from the data (2.83% of the observations).

Table 4 and **Figure 6** summarize the main results of the model for the effect of reading fluency, WMC, locality effect (embedded subject length × dependency), and its interaction with reading fluency and WMC, including the data from all the regions of interest. We omitted the interactions with the different regions since the effects of interest had the same pattern in all the regions. **Table 5** summarizes the results from nested comparisons where the models were evaluated at the different regions.

The models reveal the following: As in Experiment 1, even though it is with less certainty, subjects with higher reading fluency scores tended to have shorter RTs.

In addition, and as in the previous experiment, we did not find the hypothesized locality effects in this experiment. The models, however, show evidence for an interaction between locality effects and WMC. This interaction has the same pattern in all regions of interest. The resulting effect is similar to the one of Experiment 1, even though the underlying pattern is different (see **Figure 4**): the effect was mainly driven by a speedup in long baseline conditions in comparison with short baseline conditions. This speedup was reduced as WMC decreased until it became an advantage for the short condition for low-WMC readers; compare the figures depicting the effects for high- and low-WMC in Experiment 2 (**Figure 5**) with Experiment 1 (**Figure 2**).

We also found some evidence for a three-way interaction between embedded subject length, dependency type, and reading fluency, with the same direction as in Experiment 1, that is, decreasing locality effects as the score of reading fluency increases. The interaction had the following pattern: For the unbounded dependency conditions, as reading fluency increased, RTs at the long condition decreased in comparison with the RTs at the short condition; while for the baseline conditions this pattern was reversed.

3.3. Discussion

We found a dependency type × embedded subject length × WMC interaction, which had the same sign as in the previous experiment. However, while in Experiment 1 the effect seemed to be caused by the difference between the unbounded dependency conditions, in Experiment 2, the effect was mainly caused by a difference between the baseline conditions. In contrast to the Spanish stimuli, the subject did not immediately precede the verb in the German stimuli and therefore had to be retrieved from memory. Since the long conditions appear together with a more informative and salient subject, and the encoding of the longer subjects seems to have not spilled over the head verb; it may be the case that the subject retrieval is faster (Hofmeister, 2007; Hofmeister and Vasishth, 2014), thus leading to a speedup in both long conditions (both unbounded dependency and baseline conditions).

But crucially, the dependency type × embedded subject length × WMC interaction had the same direction and similar magnitude as in Experiment 1, that is, high-WMC participants showed the largest difference between *long unbounded − long baseline* and *short unbounded − short baseline*, while this difference is inverted for low-WMC readers. This outcome allows

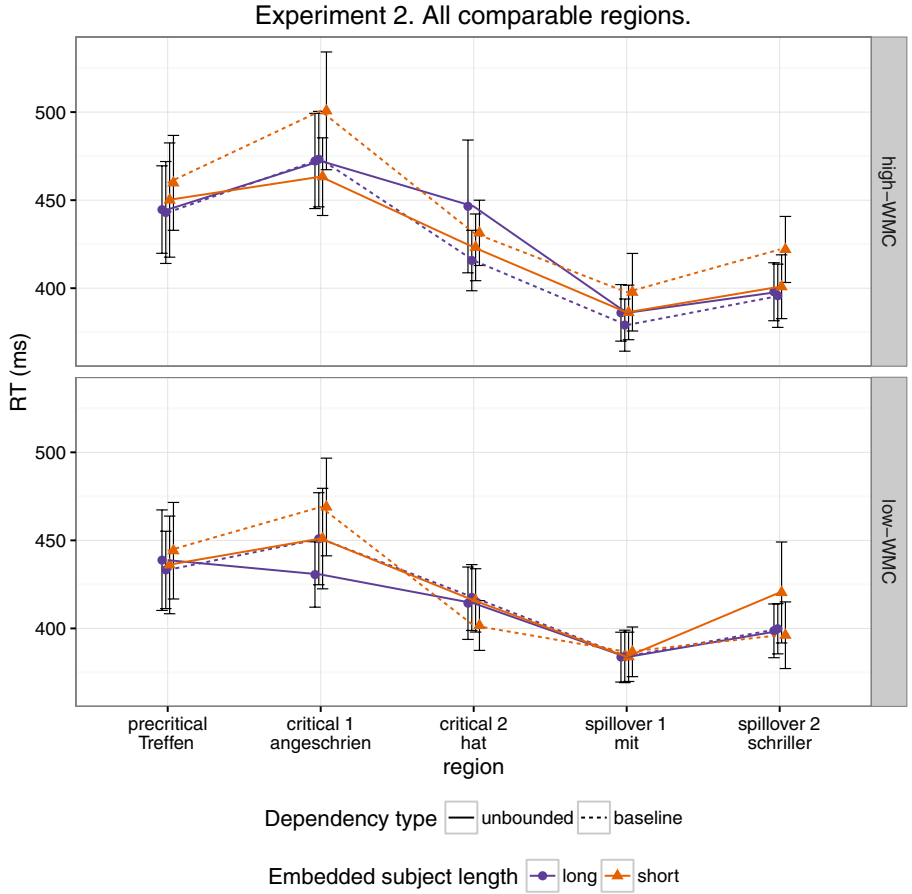


FIGURE 4 | Comparison of mean RTs for high- and low-WMC readers at each comparable region for every condition.

us to give the same interpretation to the results of the current experiment: high-WMC readers showed locality effects and low-WMC readers showed a speedup, which we argue that it is associated with a higher proportion of failure in retrieval in the long unbounded dependency condition.

In contrast with Experiment 1, reading fluency did not show a correlation with comprehension accuracy (while only WMC did). Similarly to the first experiment, however, participants with higher scores in reading fluency tended to read the critical region faster. In addition, we found somewhat stronger evidence favoring the hypothesis that fluent readers would overcome locality effects more easily than less fluent readers.

4. GENERAL DISCUSSION

We found no evidence for locality effects across the board in either experiment, that is, no evidence for an interaction between dependency type and embedded subject length independent of individual differences in WMC. However, we did find evidence for an interaction between locality effects and WMC (dependency × embedded subject length × WMC) for both Spanish and German experiments. Even though there were differences in how the three-way interaction was produced between the two experiments, this may be due to the differences in the overall structure of the sentences, namely, SVO and SOV structures (and see the previous discussion). More importantly, when the differences are controlled via baselines, we see an interaction with the same (counterintuitive) pattern in both experiments: high-WMC readers showed the strongest locality effects that were reduced with decreasing WMC and eventually changed direction, such that low-capacity readers showed a speedup effect.

The speedup of low-capacity readers is in line with independent evidence showing that in some cases high working memory load may lead to faster RTs: Van Dyke and McElree (2006) found that readers showed shorter RTs (together with lower comprehension accuracy) when a memory load was present in comparison with the conditions without the memory load. Furthermore, our findings are also compatible with studies showing that low-WMC subjects may take less time when ambiguities are present (at the expense of their accuracy) than high-WMCs (MacDonald et al., 1992; Pearlmutter and MacDonald, 1995).

It should be underscored that, unlike Just and Carpenter (1992), we do not argue that the effect of WMC is directly on mechanisms specific to language, such as parsing rules.

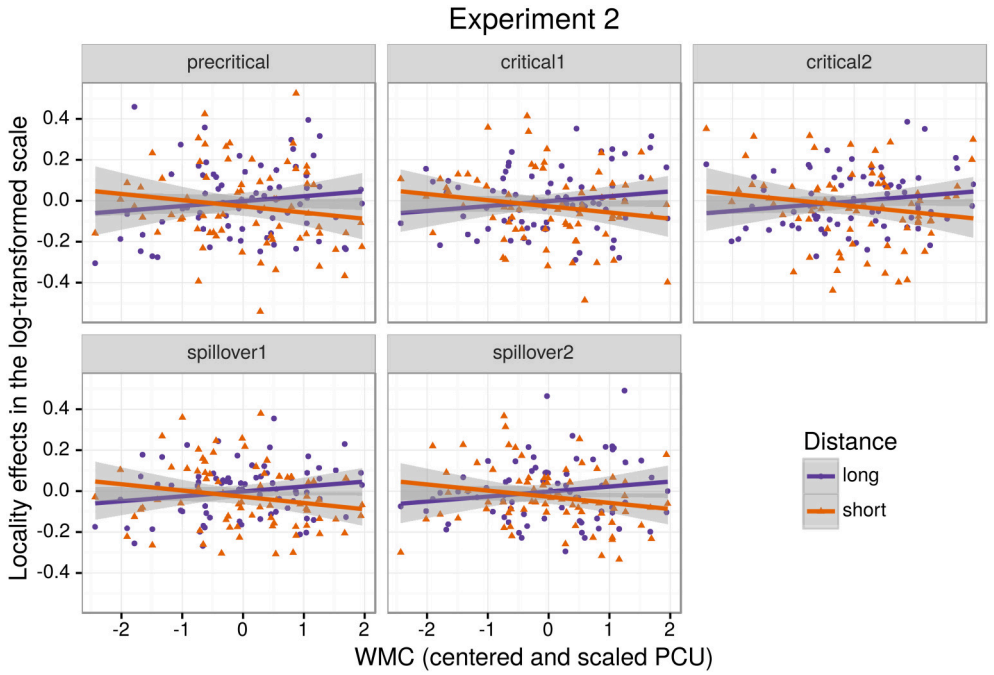


FIGURE 5 | The figure depicts the partial effect for the difference between unbounded conditions and baseline conditions, that is the (anti)locality effects, on the transformed scale of the analysis; random factors variance and effects due to reading skills were removed from the dependent variable (Hohenstein and Kliegl, 2013).

TABLE 4 | Main results for Experiment 2 (German).

Predictor	$\hat{\delta}$	95% CrI		$P(\hat{\delta} > 0)$
Length	−0.01	−0.04	0.02	0.2
Dependency	−0.02	−0.05	0.01	0.11
WMC	0.09	−0.1	0.28	0.83
RF	−0.14	−0.33	0.04	0.06
Length:dependency	0.02	−0.02	0.05	0.85
Length:dependency:WMC	0.03	0.01	0.06	0.99
Length:dependency:RF	−0.03	−0.05	0	0.04

The first column $\hat{\delta}$ shows the estimated effect size of the coefficients; the next two columns show the 2.5th and 97.5th percentiles of their posterior distribution, that is, where the effect size lies with 95% probability; and $P(\hat{\delta} > 0)$ indicates the posterior probability that each coefficient is positive.

We argue instead that the effect of WMC is on the retrieval of the dependents, which we assume is driven by the same cognitive mechanisms as retrieval outside sentence processing. There is a great deal of evidence suggesting that high-WMC participants tend to do better on tasks that involve retrieval in comparison with low-WMC ones, particularly under conditions of interference: for example, Conway and Engle (1994) found that high- and low-WMC individuals differed in retrieval efficiency only when items were associated with multiple cues (which caused more interference). In a study by Kane and Engle (2000), participants were shown a list of category exemplars followed by a distractor activity. After the distractor task, the participants were instructed to recall the category exemplars. Kane and Engle found that all participants recalled a similar number of words on the first trial but that low-WMC individuals recalled fewer items than high-WMC individuals as the task progressed. Kane and Engle concluded that low-capacity individuals were more susceptible to the buildup of proactive interference than were high-capacity ones. Conway et al. (2001) extended the investigation of the cocktail party phenomenon, the situation in which one can attend to only part of a noisy environment, but stimuli such as one's own name can suddenly capture attention. While previous investigations have shown that approximately 33% of the participants hear their name in an unattended, irrelevant message channel, Conway et al. found that 65% of low-WMC participants did detect their name in contrast with 20% of high-WMC ones. This result also suggests that low-WMC are also more susceptible to interference. Kane et al. (2001) reported similar differences in an antisaccade paradigm, which presents a conflict between task goals and visual cues. High-WMC participants made fewer errors, they recovered from these errors more rapidly, they initiated antisacades more quickly, and they identified targets more quickly than did low-WMC participants.

Besides ACT-R, two recent theories of WMC posit a role of individual differences in differential effects at retrieval: Unsworth et al. (Unsworth and Engle, 2007; Unsworth et al., 2009) have recently suggested a dual-component framework for interpreting individual differences in WMC. In this framework, WMC partially reflects differences in attention control abilities together with retrieval abilities in which information that could not be maintained in the focus of attention (due to

distraction and/or capacity constraints) is retrieved via a cue-dependent search process. In addition, Oberauer et al. (2012) have postulated a computational model, "serial order in a box - complex span" or SOB-CS (an extension of C-SOB; Farrell, 2006; Lewandowsky and Farrell, 2008, which originated as SOB; Farrell and Lewandowsky, 2002), where capacity is limited only by interference between representations. One of the individual differences that the model assumes is a parameter that determines the degree of discriminability between retrieval candidates.

In our view, non-local dependency resolution is a case where the individual differences in WMC may play a role: an argument that is no longer in the focus of attention has to be retrieved from memory, using information from the verb that is retrieved online, and after the parser has encoded a variable amount of lexical material that can produce interference together with either pure time-based decay or interference-based decay.

However, we must acknowledge that recent findings raise the concern that WMC may have limited value for explaining individual differences in linguistic contexts. A recent study by Van Dyke et al. (2014) replicated Van Dyke and McElree (2006) while including a battery of tests for measuring individual differences as well. This recent study showed that while high-span participants read more slowly in the conditions with high cognitive load and showed higher accuracy in comparison with low-span participants, the effect of WMC may be spurious. When receptive vocabulary was included in the analysis, it showed the same effects previously attributed to WMC, revealing that the participants with better scores in the vocabulary task were more affected by the interference during online reading. Similarly, a study of Traxler and Tooley (2007) investigating syntactic ambiguity showed that vocabulary size predicted the degree to which readers were disrupted by the syntactic misanalysis for several eye-tracking measures; while WMC was only a marginal predictor for total reading times. In addition, Long et al. (2008) study of recollection and familiarity of previously read sentences showed that only individual differences of readers' background knowledge was predictive of better performance but not WMC (but neither neither print exposure or vocabulary size). Long et al. (2008) argued that because retrieval cues were minimal, access to the text representation depended more on the reader's background knowledge than on the reading skills or WMC of the participants.

Our results do not rule out the possibility that retrieval processes in sentence processing are based on different mechanisms which are independent of WMC, and that the effect that we found is due to WMC being a proxy for other individual differences such as robustness of lexical representations (Traxler and Tooley, 2007; Van Dyke et al., 2014). This is a valid criticism, but it affects any experiment that includes individual differences. No matter how extensive the battery of tasks, there is always the possibility that a predictor is in fact a proxy for another unmeasured predictor.

In addition, the locality effects × WMC interaction in the two experiments should not be dismissed as a simple speed-accuracy trade-off. It is a well known phenomenon that accuracy deteriorates with increasing speed (see for example, Pachella, 1974, and more recently, Heitz, 2014). This general phenomenon, however, does not explain why low-WMC participants would decide to sacrifice accuracy for speed even to a rate that is higher than when there is a lower cognitive load (i.e., a shorter dependency). Furthermore, it also does not explain what mechanisms low-WMC participants may have used to identify the high-cognitive load conditions in order to speed up.

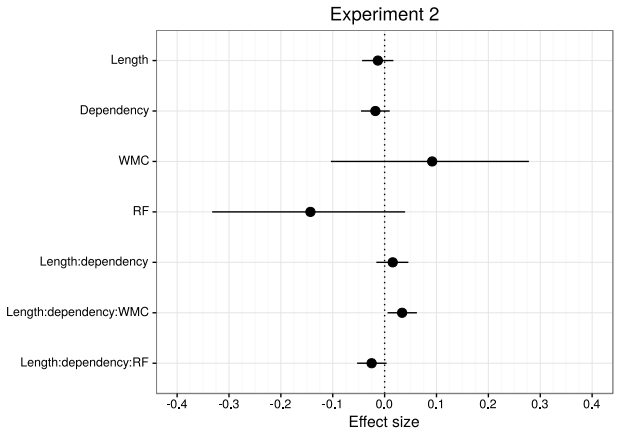


FIGURE 6 | Overview of mean and 95% credible intervals for the effect sizes of the parameters of interest for Experiment 2.

We suggest that the locality effects of the high-WMC readers and the speedup of the low-WMC readers can be explained by adding two assumptions to memory-based explanations, namely, (i) that failures of the retrieval of the dependent (the *wh*-element in this case) are more frequent in low-WMC participants than in high-WMC ones; and (ii) that retrieval failures are faster on average than complete retrievals.

The locality effects × WMC interaction in the two experiments may be related to some type of good-enough parsing strategy (Ferreira et al., 2002; Ferreira and Patson, 2007), where low-WMC readers failed to achieve a complete and fully specified representation of the sentence more often when faced with the long unbounded dependency condition. Without the possibility of re-reading, and since the comprehension questions were not targeting exclusively whether the dependency was understood, low-WMC readers may have failed in many cases to retrieve the dependent and continued reading.

In other words, we speculate that the average time T for the completion of a dependency is determined by:

$$T = T_{baseline} + P_{retrieval} \cdot T_{retrieval} + (1 - P_{retrieval}) \cdot T_{failure}$$

while the proportion of completed retrievals $P_{retrieval}$ is higher for high-WMC readers in comparison with low-WMC readers when the dependent-head distance is increased; and $T_{retrieval}$ at a long dependency is larger than $T_{retrieval}$ at a short dependency.

Notice, however, that without the proportion of completed retrievals ($P_{retrieval}$) for each case, the model previously presented is unidentifiable. The proportion of completed retrievals should be linked to the accuracy of the comprehension of the dependencies.

TABLE 5 | Main results for each region of Experiment 2 (German).

Predictor	$\hat{\delta}$	95% CrI		$P(\hat{\delta} > 0)$
PRECRITICAL				
Length	0.01	−0.03	0.05	0.63
Dependency	−0.03	−0.07	0.01	0.07
WMC	0.08	−0.1	0.26	0.82
RF	−0.14	−0.32	0.05	0.07
Length:dependency	0.04	0	0.07	0.97
Length:dependency:WMC	0.03	−0.01	0.07	0.95
Length:dependency:RF	−0.02	−0.06	0.02	0.14
CRITICAL1				
Length	−0.02	−0.06	0.02	0.14
Dependency	−0.06	−0.1	−0.03	0
WMC	0.14	−0.02	0.3	0.96
RF	−0.12	−0.29	0.03	0.06
Length:dependency	0.01	−0.03	0.05	0.74
Length:dependency:WMC	0.03	−0.01	0.06	0.92
Length:dependency:RF	−0.05	−0.08	−0.01	0
CRITICAL2				
Length	0	−0.04	0.03	0.42
Dependency	−0.01	−0.05	0.03	0.37
WMC	0.11	−0.04	0.26	0.93
RF	−0.11	−0.26	0.04	0.07
Length:dependency	0	−0.04	0.04	0.47
Length:dependency:WMC	0.05	0.01	0.08	0.99
Length:dependency:RF	−0.01	−0.04	0.03	0.37
SPILLOVER1				
Length	−0.01	−0.05	0.03	0.3
Dependency	0.02	−0.03	0.06	0.78
WMC	0.1	−0.06	0.27	0.9
RF	−0.17	−0.33	0	0.02
Length:dependency	0.01	−0.03	0.05	0.75
Length:dependency:WMC	0.01	−0.02	0.05	0.75
Length:dependency:RF	−0.02	−0.05	0.02	0.19
SPILLOVER2				
Length	−0.04	−0.07	0	0.02
Dependency	0	−0.04	0.03	0.42
WMC	0.07	−0.09	0.24	0.81
RF	−0.18	−0.35	−0.02	0.02
Length:dependency	0.01	−0.03	0.05	0.7
Length:dependency:WMC	0.04	0	0.07	0.98
Length:dependency:RF	−0.02	−0.05	0.02	0.15

WMC stands for working memory capacity and RF for reading fluency. The first column $\hat{\delta}$ shows the estimated effect size of the coefficients; the next two columns show the 2.5th and 97.5th percentiles of their posterior distribution, that is, where the effect size lies with 95% probability; and $P(\hat{\delta} > 0)$ indicates the posterior probability that each coefficient is positive.

There is some evidence that high-WMC outperformed low-WMC in general comprehension in this experiment; but we could not target the comprehension of the dependencies in the experimental stimuli. The true-or-false statements used in both Experiment 1 and 2 (as in Nicenboim et al., 2015b) included many aspects of the stimuli to verify that participants paid attention to the sentences, but they did not target exclusively whether the dependency was understood. Since the dependencies included a wh-argument, comprehension questions would ideally need to verify unnatural constructions, namely, whether it is true that, for example, "Maria greeted whom." However, preliminary data from our lab (Nicenboim et al., 2015a), where the stimuli allowed for more informative question-response accuracy, suggest that at least for interference effects in relative clauses, both low-WMC and high-interference conditions seem to provoke more retrieval failures.

In the following section we present simulations based on the ACT-R framework to illustrate in which situations and under which assumptions our hypothesis holds.

Regarding the effect of reading fluency on locality effects, the experiments presented some weak evidence favoring the hypothesis that fluent readers may overcome locality effects more easily than less fluent ones. The evidence is rather weak for the following reasons: Reading fluency predicted comprehension accuracy in Experiment 1, where it interacted very weakly with locality effects and with much uncertainty. In contrast, reading fluency did not predict comprehension accuracy in Experiment 2, while it interacted more strongly with locality effects and with less uncertainty. Given the similarity between the experiments, it is hard to explain the discrepancies.

To some degree in Experiment 1 and with more uncertainty in Experiment 2, the pattern showing stronger locality effects for high-WMC participants begins at the precritical region before the verb. Memory driven locality effects, however, are predicted to be triggered by a retrieval process that would start presumably no sooner than the verb. One possible explanation proposed by Vasishth and Drenhaus (2011) is that the verb phrase may have already been built when the proper noun preceding the verb is processed. This assumption is consistent with Levy's (2008) expectation-based account, because the parser can deduce that the verb will appear immediately afterwards and thus anticipate the retrieval process.

It should be noted that expectations were controlled only under the simplifying assumption that given that a clause has a finite length, the probability that the next word will be the subcategorizing verb rises as the number of words after finding the wh-element increases. In a way, this is similar to the increasing hazard function proposed for visual search by Peterson et al. (2001). A more formal verification could not be conducted, since the sentences used for our two experiments were too complex for a correct parsing of a probabilistic top-down parsing (Roark, 2001; Roark et al., 2009) trained with Spanish (Moreno et al., 2003) and German treebanks (Brants et al., 2004). Even after unlexicalizing the treebanks, the parser failed to identify the structure of the sentences used in our stimuli. However, given that the speedup occurred for low-span readers in sentences with long dependencies, assuming that increasing the length of the dependencies still caused an increase in expectations beyond the control of the baseline would require the implausible assumption that low-span readers are better at making predictions than high-span readers.

5. MODELING

Even though both the activation-based account and DLT would intuitively predict that increasing the distance between dependent and head should have produced a slowdown (once expectations are controlled), our results do not show a main effect of locality and only an interaction with WMC. Thus, we first verified that the activation-based account in fact predicts locality effects and stronger effects for low-span readers using the ACT-R framework (see for example Anderson et al., 2004). ACT-R is a general cognitive architecture used to model a vast variety of cognitive phenomena; for our purposes, however, the relevant aspect of the architecture is that it can model the retrieval of items stored in memory. In order to simplify our models, we used only the equations that determine the probability and latency of a retrieval and not the full framework. In this section, we tested different implementations of WMC with the "default" ACT-R equations and we show that, no matter what the parameter settings are, they fail to account qualitatively for the results. Therefore, we tentatively suggest that a basic assumption about the relationship between latencies and activation needs some reconsideration; we propose that items in memory with an activation below a certain threshold may show shorter latencies because of an early aborting of the retrieval process.

The exact predictions of the ACT-R implementation of the activation-based account will depend on the exact syntactic structure and the type of parser that is assumed together with the values of the ACT-R parameters. In addition, it cannot at present accommodate certain aspects that seem to have an uncontroversial effect in language, such as expectations (Hale, 2001; Levy, 2008). Thus, we focused on the explanation of (anti-)locality effects, (i.e., the interaction distance × dependency), which was the theoretical comparison of interest; and we did not investigate the underlying processes that generated the reading times for each condition (see Introduction).

In this framework, the latency of the retrieval of an item from memory is assumed to be a function of the item's activation value A:

$$Latency = F \cdot e^{-A} \tag{1}$$

where F (the latency factor) is a scaling constant.

After verifying that ACT-R did not predict that other noun phrases would be mistakenly retrieved, we focused only on the retrieval of the wh-element. At the moment of retrieval, the activation A is calculated as the sum of (i) a *base level activation* BA that depends on the previous use of the item (i.e., the number of previous retrievals and the time passed since those retrievals); (ii) *spreading activation* S that depends on a limited amount of source activation W that is shared between all other items with features that match the retrieval cues; (iii) a penalty component for mismatching features (that we omit from the following equation); and (iv) a random noise component ϵ (that follows a logistic distribution with a mean of zero and scale σ):

$$A = BA + S + \epsilon \tag{2}$$

Locality effects affect only the base level activation due to decay; in our specific case, the base level activation of the wh-element can be described as:

$$BA = log(t^{-d}) + \beta \tag{3}$$

where d is the decay rate, t is the time since the encoding of the wh-element, and β is the base-level constant.

The equation for the spreading activation S ensures that the wh-element would be retrieved due to the boost of activation produced by the unique matching features. For simplicity, we can assume that the wh-element has a unique feature that distinguishes it from the other four competitor NPs (in example 6: Sofía, the younger sister, the younger sister of Sofía, and María), namely being $+wh$, and two non-unique features ($+animate$ and $+NP$) that it does share with the other NPs. The spreading activation of the wh-element is a function of the source activation W, and the weighted sum of the strength of association of the cues. The source activation is usually set to one, but it can also vary by participants (Daily et al., 2001), and it is divided between the cues. In the present case, this can be simplified as:

$$\begin{aligned} S = \; & W \cdot \big[w_{wh} \cdot (MAS - \log(1)) + w_{anim} \cdot (MAS - \log(5)) \\ & + w_{NP} \cdot (MAS - \log(5)) \big] \end{aligned} \tag{4}$$

where MAS is the maximum associative strength; and w_{wh}, w_{anim}, and w_{NP} are the weights given to the cues $+wh$, $+animate$, and $+NP$, and must sum to one. The maximum associative strength is subtracted by the natural logarithm of the number of competing items in memory that match a given cue plus one. MAS is an arbitrary value, which is usually fixed since it trades off with F (Schneider and Anderson, 2012). We fixed this parameter to two since the difference between MAS and $log(matchingcues + 1)$ must be always positive in ACT-R. The three summands of the previous equation represent three features that match with three retrieval cues: The first summand represents the unique feature $+wh$, which ensures the highest value of S for the wh-element, and the next two summands represent the features $+animate$ and $+NP$, which are shared with four competitors (hence log(5), as there are five noun phrases in total). (The spreading activation equations of the competitor noun phrases would have only the last two summands; and their activation would be reduced further by a penalty component that is also subtracted from their total activation).

WMC has been assumed to either affect the decay rate or affect in some way the spreading activation, that is, the activation shared between the retrieval cues (see the Introduction section). We simulated these possibilities by using standard ACT-R parameters from sentence processing (Lewis and Vasishth, 2005; Vasishth and Lewis, 2006), except for MAS, the latency factor, and the base levels constant that were adjusted to achieve realistic latencies based on previous studies.

The first possibility is the *capacity-as-decay-rate* model, which assumes that higher-WMC should predict a lower decay rate d (e.g., Byrne and Bovair, 1997; Cunnings and Felser, 2013). Then high-WMC participants will be less affected by longer dependency distance (which entails longer time since encoding); see **Figure 7A**.

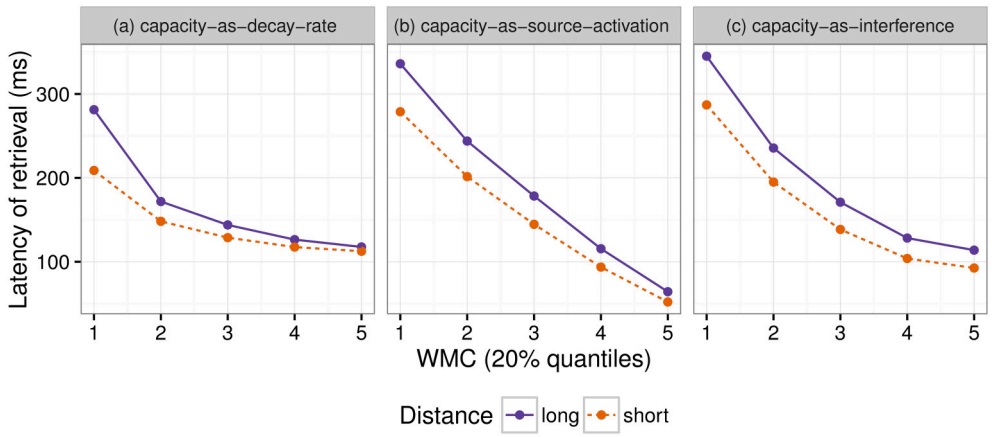


FIGURE 7 | The figures show simulated retrieval latencies at the critical region. The simulation is based on the default but simplified version of ACT-R. See **Table 6** for the parameters used.

TABLE 6 | Parameter values for the models with the default (simplified) ACT-R.

	cap-as-decay-rate	cap-as-source activation	cap-as interference
d	[0.06, 1.5]	0.5	0.5
W	1	[0.5, 3.5]	1
w_{wh}	1/3	1/3	[0.1, 1]
MAS	2	2	2
F	1.6	1.6	1.6
σ	0.25	0.25	0.25
β	−0.65	−1	−1
(Without threshold) τ	$-Inf$	$-Inf$	$-Inf$
(With threshold) τ	0	0	0

The decay time was calculated from the data of the German experiment; we used the mean reading time elapsed from the wh-element until the verb, which was 2614 ms for the short condition and 3922 ms for the long one.

If higher-WMC correlates with more spreading activation, there are two approaches: (i) The *capacity-as-source-activation* model assumes that the total amount of activation that is shared between matching cues (the source activation W) is a function of WMC (as in Cantor and Engle, 1993; Daily et al., 2001; van Rij et al., 2013); see **Figure 7B**. (ii) The *capacity-as-interference* model assumes that WMC represents susceptibility to interference (Bunting et al., 2004). Non-unique retrieval cues such as looking for a noun phrase or for the feature *+animate* cause the limited amount of source activation to be shared between competitor noun phrases, decreasing the total level of activation of the target (and also increasing the activation level of competitors). A way to model this susceptibility to interference is to change the weight given to unique cues and non-unique cues, so that as WMC increases, the weight given to a unique retrieval cue (such as being a wh-element) increases too; see **Figure 7C**.

These models predict mainly that an increase in WMC would increase the speed of the retrievals, as well as an interaction between WMC and dependency-head distance in raw RTs. The strength of the effect of WMC as well as the interaction will depend on the values of the parameters, and given that there is noise in the system (recall that the activation includes also a component ϵ), not every possible model will show these effects.

Given the relation between activation and latency, the models that assume that WMC affects the activation linearly (such as capacity-as-source-activation and capacity-as-interference) have two important implications: The first one is that if WMC affects the spreading activation S, locality effects *in raw latencies* should be modulated by WMC. The second implication is that for *log-transformed latencies*, the interaction should be exactly zero. The reason is the following: Locality effects are produced by the difference in the retrieval latencies, such that due to decay, the base level activation BA decreases as the distance between wh-element and head increases:

$$\begin{aligned} Locality &= Latency_{LongDep} - Latency_{ShortDep} \\ &= F \cdot (e^{-(BA_{low}+S)} - e^{-(BA_{high}+S)}) \\ &= F \cdot e^{-S} \cdot (e^{-BA_{low}} - e^{-BA_{high}}) \end{aligned} \quad (5)$$

If, as hypothesized, WMC only affects the spreading activation S, such that the S is higher for high-WMC than for low-WMC, then the interaction between locality effects and WMC would be defined as follows:

$$\begin{aligned} Locality \times WMC &= Locality_{LowWMC} - Locality_{HighWMC} \\ &= F \cdot (e^{-S_{low}} - e^{-S_{high}}) \cdot (e^{-BA_{low}} - e^{-BA_{high}}) \end{aligned} \quad (6)$$

However, log-transformed locality effects are independent of S:

$$\begin{aligned} log(Locality) &= log(F) - (BA_{low} + S) - [log(F) - (BA_{high} + S)] \\ &= -BA_{low} + BA_{high} \end{aligned} \quad (7)$$

and thus the difference between locality effects for high and low-WMC for log-transformed latencies would be simply zero.

But critically, no matter the values of the parameters, these two models cannot predict our findings, namely, a speedup for

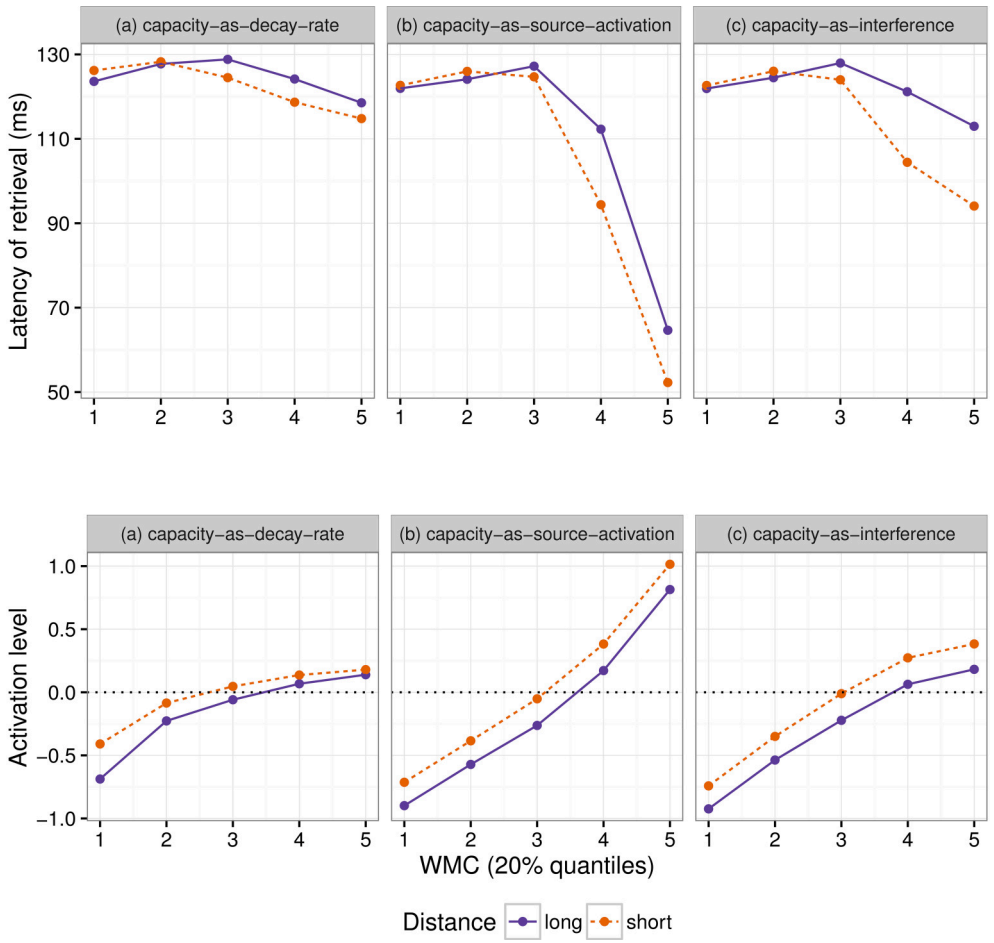


FIGURE 8 | The upper figures show simulated retrieval latencies and the lower figures the activation values that produced the latencies. The simulation is based on the default but simplified version of ACT-R, where the threshold τ is zero. See **Table 6** for the parameters used.

low-span participants. This is so because the baseline activation of the wh-element when it is retrieved after a longer time (due to more intervening material between itself and the head verb) can never be higher than the level of activation when the element is retrieved after shorter time; furthermore, the spreading activation can at most attenuate this effect and only as WMC increases.

It is further assumed in ACT-R models that there is a minimum level of activation τ that an item needs in order to be retrieved. This acts as a time-out: when an item has so low activation that it would take an unrealistic amount of time to be retrieved, the retrieval fails. The maximum amount of time is a function of this activation threshold τ such that:

$$max(Latency) = F \cdot e^{-\tau} \tag{8}$$

If τ plays a role in retrieval, because the activation level of the dependent does not always exceed this value, it will produce a ceiling effect. Under this view, if the activation level of the dependent for low-WMC failed more often than for high-WMC to surpass τ, it would entail a maximum possible time for both short and long conditions. This would produce a difference in retrieval probabilities between short and long conditions, since it is more likely that long conditions fail more often to surpass the value τ. However, this would also mean that with low-WMC, the difference between long and short conditions may disappear; see **Figure 8**. Our data cannot be accommodated in these models either, since the difference between long and short conditions was reversed for low-WMC.

The pattern that we found in our data, however, can only be accommodated in the models presented before by changing one assumption, namely, by assuming that in the cases where the activation does not reach the threshold τ, the retrieval would be aborted at any moment before the maximum amount of time. In this view, failed retrievals would take less time on average than the time needed to retrieve the item, with the activation influencing the retrieval probability and WMC in turn influencing the level of activation; see **Figure 9**. This would mean that τ would act as a critierion for aborting instead of a time-out. We simulated this by assuming that WMC affects the activation of the wh-element very weakly and, critically, that retrievals can fail at any time before the maximum retrieval latency (i.e., following a uniform distribution limited between zero and $max(Latency)$).

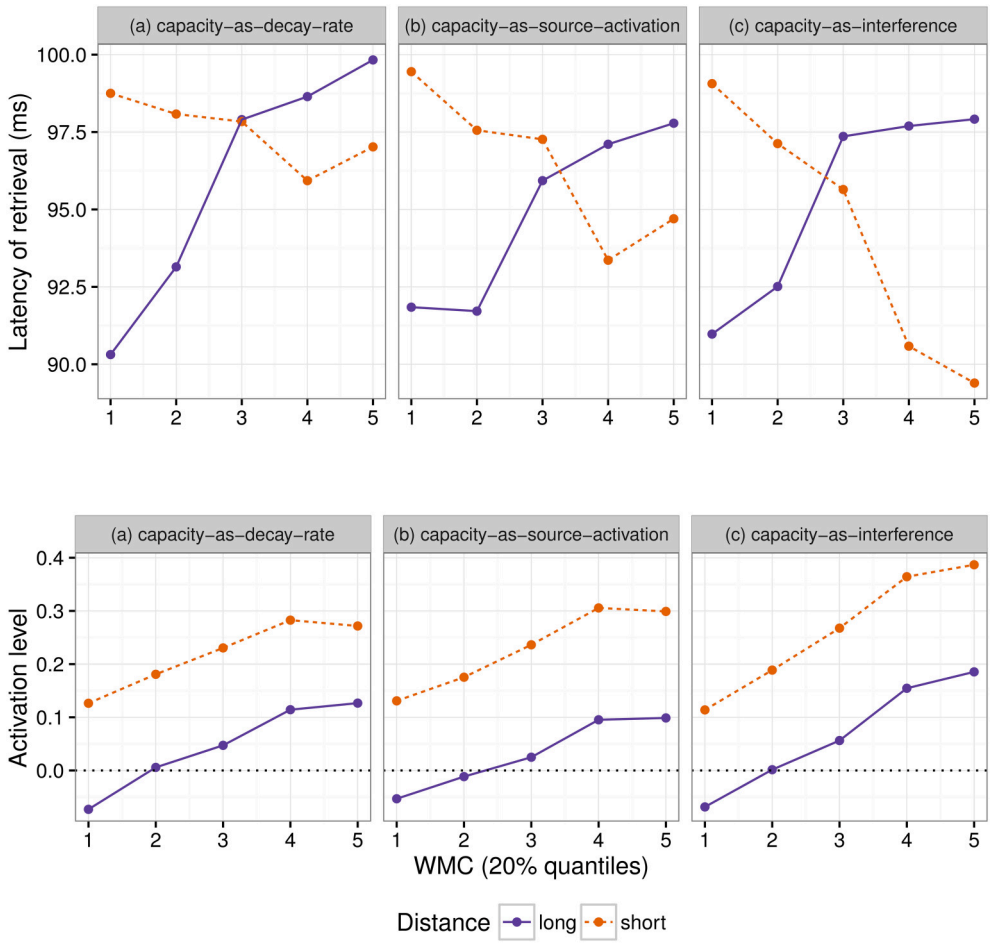


FIGURE 9 | The upper figures show simulated retrieval latencies and the lower figures the activation values that produced the latencies. The simulation is based on the modified version of ACT-R, where the threshold τ is zero. See **Table 7** for the parameters used.

There are, of course, other possibilities that will fit with the general pattern as well: Any distribution of latencies with a mean that is smaller than the average latency for a retrieval will show this pattern. Importantly, by relaxing the ACT-R assumption that too low activation must produce the longest possible latency, we are able to account qualitatively for the pattern in our data. This is so, because participants with lower-WMC would fail more often than high-WMC, and since they would complete retrievals relatively slowly, their failures would be on average faster. An interesting prediction from this modification is that in the small number of cases where a retrieval would fail for high-WMC participants, because high-WMC subjects should produce faster than average retrievals, they would still show slower failures in comparison to their retrievals.

There is some parallelism between fast failures in our experiment and fast errors in two-alternative forced choice tasks. Recent research in two-alternative forced choice tasks has shown that time-varying collapsing thresholds (e.g., Frazier and Yu, 2008; Drugowitsch et al., 2012; Thura et al., 2012) can explain wrong answers that are given too early, even though there is no apparent imposed deadline. Self-paced reading presents a paradigm, however, where the only possible choice at every point is to press the space bar to continue reading. In order to build a complete representation of the sentences,

TABLE 7 | Parameter values for the models with the modified (simplified) ACT-R.

	cap-as-decay-rate	cap-as-source-activation	cap-as-interference
d	[0.33, 0.62]	0.5	0.5
W	1	[0.9, 1.2]	1
w_{wh}	1/3	1/3	[0.2, 0.5]
MAS	2	2	2
F	1.6	1.6	1.6
σ	0.25	0.25	0.25
β	−0.3	−0.3	−0.3
(With threshold) τ	0	0	0

The decay time was calculated from the data of the German experiment; we used the mean reading time elapsed from the wh-element until the verb, which was 2614 ms for the short condition and 3922 ms for the long one.

participants reading the verb region should delay pressing the space bar, until they retrieve from memory the dependent and they complete the dependency. However, we have argued that, when the dependent does not have enough activation, retrieval processes are aborted early. Assuming time has a cost, Frazier and Yu (2008) argue that an optimal stopping rule for a process is to stop the first time that the expected cost of continuing exceeds that of stopping, and to continue only if it is going to improve the chances of success enough to offset the extra time. A stopping rule in self-paced reading would mean pressing the space bar and continue reading. When an item to be retrieved has enough activation, an optimal stopping rule could be to wait and continue reading only when the retrieval is finished. Alternatively, when an item has insufficient activation, the parser could evaluate that the activation would not be enough to finish the retrieval before a time out ($F \cdot e^{-\tau}$), abort the process, and continue reading, explaining the fast failures.

Further research with data that include RTs as well as some index of retrieval accuracy, which is as little contaminated as possible with general comprehension accuracy, other retrievals, and offline processes, could shed light on how and when exactly retrieval fails.

6. CONCLUSION

We presented two experiments showing that working memory affects locality effects. The results show that working memory affects retrieval times at unbounded dependency resolution, but in an unexpected manner: high-capacity readers showed the strongest locality effects that decreased with decreasing capacity and eventually changed direction, such that low-capacity readers showed antilocality effects.

We suggest that the results may not be simply due to a speed-accuracy trade-off and that they can be explained by adding two assumptions to memory-based explanations: (i) compared to high-capacity readers, low-capacity readers experience retrieval failures more frequently; and (ii) retrieval failures are on average faster than complete retrievals. We suggest that the retrieval failures end quickly because of insufficient activation, and this activation depends not only on dependent-head distance but also on the capacity of the readers.

All in all, both experiments show that translating longer RTs into processing difficulty and shorter RTs into facilitation may be too simplistic, especially when readers face long and complex sentences (which are not uncommon in psycholinguistic studies). Our results suggest that the same increase in processing difficulty may lead to slowdowns in high-capacity readers and speedups in low-capacity ones.

FUNDING

The work was supported by Minerva Foundation, Potsdam Graduate School, and the University of Potsdam. We acknowledge the support of the Deutsche Forschungsgemeinschaft and Open Access Publishing Fund of University of Potsdam.

ACKNOWLEDGMENTS

A preliminary version of this paper appeared in the Proceedings of TL/MAPLL 2014, Tokyo, Japan. Thanks to Richard McElreath for sharing his unpublished book, Statistical Rethinking, which influenced the statistical analysis. Thanks to Bob Carpenter and Stan's forum for help fitting the statistical models. Thanks to Lena Jäger for helpful comments on a draft and Felix Engelmann for help with ACT-R modeling.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpsyg.2016.00280

Data and complete code are available in the website of the first author.

REFERENCES

Al Dahhan, N., Georgiou, G. K., Hung, R., Munoz, D., Parrila, R., and Kirby, J. R. (2014). Eye movements of university students with and without reading difficulties during naming speed tasks. *Ann. Dyslexia* 64, 137–150. doi: 10.1007/s11881-013-0090-z

Altmann, E. M., and Gray, W. D. (2002). Forgetting to remember: the functional relationship of decay and interference. *Psychol. Sci.* 13, 27–33. doi: 10.1111/1467-9280.00405

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., and Qin, Y. (2004). An integrated theory of the mind. *Psychol. Rev.* 111, 1036–1060. doi: 10.1037/0033-295X.111.4.1036

Araújo, S., Reis, A., Petersson, K. M., and Faísca, L. (2015). Rapid automatized naming and reading performance: a meta-analysis. *J. Educ. Psychol.* 107, 868–883. doi: 10.1037/edu0000006

Arnell, K. M., Joanisse, M. F., Klein, R. M., Busseri, M. A., and Tannock, R. (2009). Decomposing the relation between Rapid Automatized Naming (RAN) and reading ability. *Can. J. Exp. Psychol.* 63, 173–184. doi: 10.1037/a0015721

Barr, D. J., Levy, R. P., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: keep it maximal. *J. Mem. Lang.* 68, 255–278. doi: 10.1016/j.jml.2012.11.001

Bartek, B., Lewis, R. L., Vasishth, S., and Smith, M. R. (2011). In search of on-line locality effects in sentence comprehension. *J. Exp. Psychol. Learn. Mem. Cogn.* 37, 1178–1198. doi: 10.1037/a0024194

Bates, D. M., Mächler, M., Bolker, B. M., and Walker, S. C. (2014). Fitting Linear Mixed-Effects Models using lme4. *J. Stat. Softw.* arXiv:1406.5823.

Berkson, J. (1946). Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bull.* 2, 47–53.

Berman, M. G., Jonides, J., and Lewis, R. L. (2009). In search of decay in verbal short-term memory. *J. Exp. Psychol. Learn. Mem. Cogn.* 35, 317. doi: 10.1037/a0014873

Boston, M. F., Hale, J., Kliegl, R., Patil, U., and Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: an evaluation using the potsdam sentence corpus. *J. Eye Mov. Res.* 2, 1–12.

Brants, S., Dipper, S., Eisenberg, P., Hansen-Schirra, S., König, E., Lezius, W., et al. (2004). Tiger: linguistic interpretation of a German corpus. *J. Lang. Comput.* 2, 597–620. doi: 10.1007/s11168-004-7431-3

Bunting, M. F., Conway, A. R., and Heitz, R. P. (2004). Individual differences in the fan effect and working memory capacity. *J. Mem. Lang.* 51, 604–622. doi: 10.1016/j.jml.2004.07.007

Byrne, M. D., and Bovair, S. (1997). A working memory model of a common procedural error. *Cogn. Sci.* 21, 31–61.

Cantor, J., and Engle, R. W. (1993). Working-memory capacity as long-term memory activation: an individual-differences approach. *J. Exp. Psychol. Learn. Mem. Cogn.* 19, 1101.

Caplan, D., and Waters, G. S. (1999). Verbal working memory and sentence comprehension. *Behav. Brain Sci.* 22, 77–94.

Caplan, D., and Waters, G. S. (2013). Memory mechanisms supporting syntactic comprehension. *Psychon. Bull. Rev.* 20, 243–268. doi: 10.3758/s13423-012-0369-9

Christianson, K., Williams, C. C., Zacks, R. T., and Ferreira, F. (2006). Younger and older adults' "good-enough" interpretations of garden-path sentences. *Discourse Process.* 42, 205–238. doi: 10.1207/s15326950dp4202_6

Conway, A. R., Cowan, N., and Bunting, M. F. (2001). The cocktail party phenomenon revisited: the importance of working memory capacity. *Psychon. Bull. Rev.* 8, 331–335. doi: 10.3758/BF03196169

Conway, A. R. A., and Engle, R. W. (1994). Working memory and retrieval: a resource-dependent inhibition model. *J. Exp. Psychol. Gen.* 123, 354–373. doi: 10.1037/0096-3445.123.4.354

Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., and Engle, R. W. (2005). Working memory span tasks: a methodological review and user's guide. *Psychon. Bull. Rev.* 12, 769–786. doi: 10.3758/BF03196772

Cummine, J., Chouinard, B., Szepesvari, E., and Georgiou, G. (2015). An examination of the rapid automatized naming–reading relationship using functional magnetic resonance imaging. *Neuroscience* 305, 49–66. doi: 10.1016/j.neuroscience.2015.07.071

Cunnings, I., and Felser, C. (2013). The role of working memory in the processing of reflexives. *Lang. Cogn. Process.* 28, 188–219. doi: 10.1080/01690965.2010.548391

Daily, L. Z., Lovett, M. C., and Reder, L. M. (2001). Modeling individual differences in working memory performance: a source activation account. *Cogn. Sci.* 25, 315–353. doi: 10.1207/s15516709cog2503_1

Daneman, M., and Carpenter, P. A. (1980). Individual differences in working memory and reading. *J. Verbal Learn. Verbal Behav.* 19, 450–466.

Demberg, V., and Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition* 109, 193–210. doi: 10.1016/j.cognition.2008.07.008

Denckla, M. B., and Rudel, R. G. (1976). Rapid 'Automatized' Naming (R.A.N): dyslexia differentiated from other learning disabilities. *Neuropsychologia* 14, 471–479. doi: 10.1016/0028-3932(76)90075-0

Drugowitsch, J., Moreno-Bote, R., Churchland, A. K., Shadlen, M. N., and Pouget, A. (2012). The cost of accumulating evidence in perceptual decision making. *J. Neurosci.* 32, 3612–3628. doi: 10.1523/JNEUROSCI.4010-11.2012

Farrell, S. (2006). Mixed-list phonological similarity effects in delayed serial recall. *J. Mem. Lang.* 55, 587–600. doi: 10.1016/j.jml.2006.06.002

Farrell, S., and Lewandowsky, S. (2002). An endogenous distributed model of ordering in serial recall. *Psychon. Bull. Rev.* 9, 59–79. doi: 10.3758/BF03196257

Fedorenko, E., Gibson, E., and Rohde, D. (2006). The nature of working memory capacity in sentence comprehension: Evidence against domain-specific working memory resources. *J. Mem. Lang.* 54, 541–553. doi: 10.1016/j.jml.2005.12.006

Fedorenko, E., Woodbury, R., and Gibson, E. (2013). Direct evidence of memory retrieval as a source of difficulty in non-local dependencies in language. *Cogn. Sci.* 37, 378–394. doi: 10.1111/cogs.12021

Ferreira, F., Bailey, K. G. D., and Ferraro, V. (2002). Good-enough representations in language comprehension. *Curr. Dir. Psychol. Sci.* 11, 11–15. doi: 10.1111/1467-8721.00158

Ferreira, F., and Henderson, J. M. (1993). Reading processes during syntactic analysis and reanalysis. *Can. J. Exp. Psychol.* 47, 247–275. doi: 10.1037/h0078819

Ferreira, F., and Patson, N. D. (2007). The 'good enough'approach to language comprehension. *Lang. Linguist. Compass* 1, 71–83. doi: 10.1111/j.1749-818X.2007.00007.x

Frazier, P., and Yu, A. J. (2008). "Sequential hypothesis testing under stochastic deadlines," in *Advances in Neural Information Processing Systems 20*, eds J. Platt, D. Koller, Y. Singer, and S. Roweis (Vancouver, BC: Curran Associates, Inc.), 465–472.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2014). *Bayesian Data Analysis*, *3rd Edn.* Boca Raton, FL: Taylor & Francis.

Gelman, A., Hill, J., and Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *J. Res. Educ. Eff.* 5, 189–211. doi: 10.1080/19345747.2011.618213

Gibson, E. (2000). "The dependency locality theory: a distance-based theory of linguistic complexity," in *Image, Language, Brain: Papers from the First Mind Articulation Project Symposium* (Cambridge, MA), 95–126.

Gibson, E., and Thomas, J. (1999). Memory limitations and structural forgetting: the perception of complex ungrammatical sentences as grammatical. *Lang. Cogn. Process.* 14, 225–248. doi: 10.1080/016909699386293

Greiner, B. (2004). "An online recruitment system for economic experiments," in *Forschung und wissenschaftliches Rechnen 2003. GWDG Bericht 63*, 79–93.

Grodner, D., and Gibson, E. (2005). Consequences of the serial nature of linguistic input for sentenial complexity. *Cogn. Sci.* 29, 261–290. doi: 10.1207/s15516709cog0000_7

Hale, J. T. (2001). "A probabilistic earley parser as a psycholinguistic model," in *Proceedings of NAACL 2001*, (Morristown, NJ: Association for Computational Linguistics), 1–8.

Heitz, R. P. (2014). The speed-accuracy tradeoff: history, physiology, methodology, and behavior. *Front. Neurosci.* 8:150. doi: 10.3389/fnins.2014.00150

Hoekstra, R., Morey, R. D., Rouder, J. N., and Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychon. Bull. Rev.* 21, 1157–1164. doi: 10.3758/s13423-013-0572-3

Hofmeister, P. (2007). "Retrievability and gradience in filler-gap dependencies," in *Proceedings from the Annual Meeting of the Chicago Linguistic Society*, Vol. 43 (Chicago, IL), 109–123.

Hofmeister, P., and Sag, I. A. (2010). Cognitive constraints and island effects. *Language* 86, 366–415. doi: 10.1353/lan.0.0223

Hofmeister, P., and Vasishth, S. (2014). Distinctiveness and encoding effects in online sentence comprehension. *Front. Psychol.* 5:1237. doi: 10.3389/fpsyg.2014.01237

Hohenstein, S., and Kliegl, R. (2013). Semantic preview benefit during reading. *J. Exp. Psychol. Learn. Mem. Cogn.* 40, 166–190. doi: 10.1037/a0033670

Howe, A. E., Arnell, K. M., Klein, R. M., Joanisse, M. F., and Tannock, R. (2006). The ABCs of computerized naming: equivalency, reliability, and predictive validity of a computerized rapid automatized naming (RAN) task. *J. Neurosci. Methods* 151, 30–37. doi: 10.1016/j.jneumeth.2005.07.014

Husain, S., Vasishth, S., and Narayanan, S. (2015). Integration and prediction difficulty in Hindi sentence comprehension: evidence from an eye-tracking corpus. *J. Eye Mov. Res.* 8, 1–12.

Husain, S., Vasishth, S., and Srinivasan, N. (2014). Strong expectations cancel locality effects: evidence from Hindi. *PLoS ONE* 9:e100986. doi: 10.1371/journal.pone.0100986

Just, M. A., and Carpenter, P. A. (1992). A capacity theory of comprehension: individual differences in working memory. *Psychol. Rev.* 99, 122.

Just, M. A., Carpenter, P. A., and Woolley, J. D. (1982). Paradigms and processes in reading comprehension. *J. Exp. Psychol. Gen.* 111, 228–238.

Kane, M. J., Bleckley, M. K., Conway, A. R., and Engle, R. W. (2001). A controlled-attention view of working-memory capacity. *J. Exp. Psychol. Gen.* 130, 169. doi: 10.1037/0096-3445.130.2.169

Kane, M. J., and Engle, R. W. (2000). Working-memory capacity, proactive interference, and divided attention: Limits on long-term memory retrieval. *J. Exp. Psychol. Learn. Mem. Cogn.* 26, 336–358. doi: 10.1037/0278-7393.26.2.336

King, J., and Just, M. A. (1991). Individual differences in syntactic processing: the role of working memory. *J. Mem. Lang.* 30, 580–602. doi: 10.1016/0749-596X(91)90027-H

Konieczny, L. (2000). Locality and parsing complexity. *J. Psycholinguist. Res.* 29, 627–645. doi: 10.1023/A:1026528912821

Konieczny, L., and Döring, P. (2003). "Anticipation of clause-final heads: evidence from eye-tracking and SRNs," in *Proceedings of ICCS/ASCS* (Sydney, NSW).

Kruschke, J. K., Aguinis, H., and Joo, H. (2012). The time has come: Bayesian methods for data analysis in the organizational sciences. *Organ. Res. Methods* 15, 722–752. doi: 10.1177/1094428112457829

Kuperman, V., and Van Dyke, J. A. (2011). Effects of individual differences in verbal skills on eye-movement patterns during sentence reading. *J. Mem. Lang.* 65, 42–73. doi: 10.1016/j.jml.2011.03.002

Kuperman, V., Van Dyke, J. A., and Henry, R. (in press). Eye-movement control in RAN and reading. *Sci. Stud. Read.* doi: 10.1080/10888438.2015.1128435

Levy, R. P. (2008). Expectation-based syntactic comprehension. *Cognition* 106, 1126–1177. doi: 10.1016/j.cognition.2007.05.006

Levy, R. P., Fedorenko, E., and Gibson, E. (2013). The syntactic complexity of Russian relative clauses. *J. Mem. Lang.* 69, 461–495. doi: 10.1016/j.jml.2012.10.005

Levy, R. P., and Keller, F. (2013). Expectation and locality effects in German verb-final structures. *J. Mem. Lang.* 68, 199–222. doi: 10.1016/j.jml.2012.02.005

Lewandowsky, S., and Farrell, S. (2008). Short-term memory: new data and a model. *Psychol. Learn. Motiv.* 49, 1–48. doi: 10.1016/S0079-7421(08)00001-7

Lewis, R. L., and Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cogn. Sci.* 29, 375–419. doi: 10.1207/s15516709cog0000_25

Lewis, R. L., Vasishth, S., and Van Dyke, J. A. (2006). Computational principles of working memory in sentence comprehension. *Trends Cogn. Sci.* 10, 447–454. doi: 10.1016/j.tics.2006.08.007

Limpert, E., Stahel, W. A., and Abbt, M. (2001). Log-normal distributions across the sciences: keys and clues on the charms of statistics, and how mechanical models resembling gambling machines offer a link to a handy way to characterize log-normal distributions, which can provide deeper insight into variability and probabilityâ€"normal or log-normal: that is the question. *BioScience* 51, 341–352. doi: 10.1641/0006-3568(2001)051[0341:LNDATS]2.0.CO;2

Logan, G. D. (1992). Shapes of reaction-time distributions and shapes of learning curves: A test of the instance theory of automaticity. *J. Exp. Psychol. Learn. Mem. Cogn.* 18, 883.

Long, D. L., Prat, C., Johns, C., Morris, P., and Jonathan, E. (2008). The importance of knowledge in vivid text memory: an individual-differences investigation of recollection and familiarity. *Psychon. Bull. Rev.* 15, 604–609. doi: 10.3758/PBR.15.3.604

Lustig, C. A., Berman, M. G., Nee, D. E., Lewis, R. L., Moore, K. S., and Jonides, J. (2009). "Psychological and neural mechanisms of short-term memory," in *Handbook of Neuroscience for the Behavioral Sciences*, eds G. G. Berntson and J. T. Cacioppo (Hoboken, NJ: John Wiley and Sons, Inc.), 567–585.

MacDonald, M. C., and Christiansen, M. H. (2002). Reassessing working memory: comment on Just and Carpenter (1992) and Waters and Caplan (1996). *Psychol. Rev.* 109, 35–54. doi: 10.1037/0033-295X.109.1.35

MacDonald, M. C., Just, M. A., and Carpenter, P. A. (1992). Working memory constraints on the processing of syntactic ambiguity. *Cogn. Psychol.* 24, 56–98. doi: 10.1016/0010-0285(92)90003-K

McVay, J. C., and Kane, M. J. (2011). Why does working memory capacity predict variation in reading comprehension? On the influence of mind wandering and executive attention. *J. Exp. Psychol. Gen.* 141, 302–320. doi: 10.1037/a0025250

Misra, M., Katzir, T., Wolf, M., and Poldrack, R. A. (2004). Neural systems for rapid automatized naming in skilled readers : unraveling RAN-reading relationship. *Sci. Stud. Read.* 8, 241–256. doi: 10.1207/s1532799xssr0803_4

Moreno, A., López, S., Sánchez, F., and Grishman, R. (2003). "Developing a syntactic annotation scheme and tools for a Spanish treebank," in *Treebanks*, ed A. Abeillé (Dordrecht: Springer), 149–163.

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., and Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychon. Bull. Rev.* 23, 103–123. doi: 10.3758/s13423-015-0947-8

Nicenboim, B., Suckow, K., and Vasishth, S. (2015a). "Fail fast or succeed slowly: good-enough processing can mask interference effects," in *Proceedings of the CUNY Sentence Processing Conference*, University of Southern California (Los Angeles, CA).

Nicenboim, B., Vasishth, S., Gattei, C., Sigman, M., and Kliegl, R. (2015b). Working memory differences in long-distance dependency resolution. *Front. Psychol.* 6:312. doi: 10.3389/fpsyg.2015.00312

Norton, E. S., and Wolf, M. (2012). Rapid automatized naming (RAN) and reading fluency: implications for understanding and treatment of reading disabilities. *Ann. Rev. Psychol.* 63, 427–452. doi: 10.1146/annurev-psych-120710-100431

Oberauer, K., and Kliegl, R. (2006). A formal model of capacity limits in working memory. 55, 601–626. doi: 10.1016/j.jml.2006.08.009

Oberauer, K., Lewandowsky, S., Farrell, S., Jarrold, C., and Greaves, M. (2012). Modeling working memory: an interference model of complex span. *Psychon. Bull. Rev.* 19, 779–819. doi: 10.3758/s13423-012-0272-4

Pachella, R. G. (1974). "The interpretation of reaction time in information processing research," in *Human Information Processing: Tutorials in Performance and Cognition*, ed B. Kantowitz (Hillsdale, NJ: Lawrence Erlbaum Associates), 41–82.

Palmer, E. M., Horowitz, T. S., Torralba, A., and Wolfe, J. M. (2011). What are the shapes of response time distributions in visual search? *J. Exp. Psychol. Hum. Percept. perform.* 37, 58. doi: 10.1037/a0020747

Pearlmutter, N., and MacDonald, M. C. (1995). Individual differences and probabilistic constraints in syntactic ambiguity resolution. *J. Mem. Lang.* 34, 521–542.

Peterson, M. S., Kramer, A. F., Wang, R. F., Irwin, D. E., and McCarley, J. S. (2001). Visual search has memory. *Psychol. Sci.* 12, 287–292. doi: 10.1111/1467-9280.00353

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Roark, B. (2001). Probabilistic top-down parsing and language modeling. *Comput. Linguist.* 27, 249–276. doi: 10.1162/089120101750300526

Roark, B., Bachrach, A., Cardenas, C., and Pallier, C. (2009). "Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 3*, (Stroudsburg, PA: Association for Computational Linguistics), 324–333.

Rouder, J. N. (2005). Are unshifted distributional models appropriate for response time? *Psychometrika* 70, 377–381. doi: 10.1007/s11336-005-1297-7

Rouder, J. N., Lu, J., Speckman, P. L., Sun, D., and Jiang, Y. I. (2005). A hierarchical model for estimating response time distributions. *Psychon. Bull. Rev.* 12, 195–223. doi: 10.3758/BF03257252

Rouder, J. N., Morey, R. D., Speckman, P. L., and Province, J. M. (2012). Default bayes factors for anova designs. *J. Math. Psychol.* 56, 356–374. doi: 10.1016/j.jmp.2012.08.001

Rouder, J. N., Tuerlinckx, F., Speckman, P., Lu, J., and Gomez, P. (2008). A hierarchical approach for fitting curves to response time measurements. *Psychon. Bull. Rev.* 15, 1201–1208. doi: 10.3758/PBR.15.6.1201

Schielzeth, H., and Forstmeier, W. (2009). Conclusions beyond support: overconfident estimates in mixed models. *Behav. Ecol.* 20, 416–420. doi: 10.1093/beheco/arn145

Schneider, D. W., and Anderson, J. R. (2012). Modeling fan effects on the time course of associative recognition. *Cogn. Psychol.* 64, 127–160. doi: 10.1016/j.cogpsych.2011.11.001

Sorensen, T., Hohenstein, S., and Vasishth, S. (2015). Bayesian linear mixed models using Stan: a tutorial for psychologists, linguists, and cognitive scientists. ArXiv e-print.

Stan Development Team (2015a). *Rstan: the R interface to Stan, version 2.7.0*.

Stan Development Team (2015b). *Stan: A C++ Library for Probability and Sampling, version 2.7.0*.

Staub, A. (2010). Eye movements and processing difficulty in object relative clauses. *Cognition* 116, 71–86. doi: 10.1016/j.cognition.2010.04.002

Thura, D., Beauregard-Racine, J., Fradet, C.-W., and Cisek, P. (2012). Decision making by urgency gating: theory and experimental support. *J. Neurophysiol.* 108, 2912–2930. doi: 10.1152/jn.01071.2011

Traxler, M. J., Long, D. L., Johns, C. L., Zirnstein, M., Tooley, K. M., and Jonathan, E. (2012). Individual differences in eye-Movements during reading : working memory and speed-of-processing effects. *J. Eye Mov. Res.* 5, 1–16.

Traxler, M. J., and Tooley, K. M. (2007). Lexical mediation and context effects in sentence processing. *Brain Res.* 1146, 59–74. doi: 10.1016/j.brainres.2006.10.010

Traxler, M. J., Williams, R. S., Blozis, S. A., and Morris, R. K. (2005). Working memory, animacy, and verb class in the processing of relative clauses. *J. Mem. Lang.* 53, 204–224. doi: 10.1016/j.jml.2005.02.010

Turner, M. T., and Engle, R. W. (1989). Is working memory capacity task dependent? *J. Mem. Lang.* 28, 127–154. doi: 10.1016/0749-596X(89)90040-5

Unsworth, N., and Engle, R. W. (2007). The nature of individual differences in working memory capacity: active maintenance in primary memory and controlled search from secondary memory. *Psychol. Rev.* 114, 104–132. doi: 10.1037/0033-295X.114.1.104

Unsworth, N., Spillers, G. J., and Brewer, G. A. (2009). Examining the relations among working memory capacity, attention control, and fluid intelligence from a dual-component framework. *Psychol. Sci. Q.* 51, 388–402.

van den Bos, K. P., Zijlstra, B. J., and lutje Spelberg, H. C. (2002). Life-span data on continuous-naming speeds of numbers, letters, colors, and pictured objects, and word-reading speed. *Sci. Stud. Read.* 6, 25–49. doi: 10.1207/S1532799XSSR0601_02

Van Dyke, J. A., Johns, C. L., and Kukona, A. (2014). Low working memory capacity is only spuriously related to poor reading comprehension. *Cognition* 131, 373–403. doi: 10.1016/j.cognition.2014.01.007

Van Dyke, J. A., and McElree, B. (2006). Retrieval interference in sentence comprehension. *J. Mem. Lang.* 55, 157–166. doi: 10.1016/j.jml.2006.03.007

van Rij, J., van Rijn, H., and Hendriks, P. (2013). How WM load influences linguistic processing in adults: a computational model of pronoun interpretation in discourse. *Top. Cogn. Sci.* 5, 564–580. doi: 10.1111/tops.12029

Vasishth, S. (2003). *Working Memory in Sentence Comprehension: Processing Hindi Center Embeddings*. New York, NY: Routledge.

Vasishth, S., and Drenhaus, H. (2011). Locality in German. *Dialogue Discourse* 2, 59–82. doi: 10.5087/dad.2011.104

Vasishth, S., and Lewis, R. L. (2006). Argument-head distance and processing complexity: explaining both locality and antilocality effects. *Language* 82, 767–794. doi: 10.1353/lan.2006.0236

Vehtari, A., and Gelman, A. (2014). *WAIC and Cross-Validation in Stan*. Available online at: http://www.stat.columbia.edu/~gelman/research/unpublished/waic_stan.pdf (Accessed April 27, 2015).

von der Malsburg, T. (2015). Py-Span-Task: A software for testing working memory span. *Zenodo*. doi: 10.5281/zenodo.18238

von der Malsburg, T. and Vasishth, S. (2013). Scanpaths reveal syntactic underspecification and reanalysis strategies. *Lang. Cogn. Process.* 28, 1545–1578. doi: 10.1080/01690965.2012.728232

Vos, S. H., Gunter, T. C., Schriefers, H., and Friederici, A. D. (2001). Syntactic parsing and working memory: the effects of syntactic complexity, reading span, and concurrent load. *Lang. Cogn. Process.* 16, 65–103. doi: 10.1080/01690960042000085

Wagenmakers, E.-J., and Brown, S. (2007). On the linear relation between the mean and the standard deviation of a response time distribution. *Psychol. Rev.* 114, 830. doi: 10.1037/0033-295X.114.3.830

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., and Grasman, R. (2010). Bayesian hypothesis testing for psychologists: a tutorial on the Savage-Dickey method. *Cogn. Psychol.* 60, 158–189. doi: 10.1016/j.cogpsych.2009.12.001

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *J. Mach. Learn. Res.* 11, 3571–3594.

Waters, G. S., and Caplan, D. (1996). The capacity theory of sentence comprehension: critique of Just and Carpenter (1992). *Psychol. Rev.* 103, 761–772.

Waters, G. S., and Caplan, D. (2001). Age, working memory, and on-line syntactic processing in sentence comprehension. *Psychol. Aging* 16, 128–144. doi: 10.1037/0882-7974.16.1.128

Wells, J. B., Christiansen, M. H., Race, D. S., Acheson, D. J., and MacDonald, M. C. (2009). Experience and sentence processing: statistical learning and relative clause comprehension. *Cogn. Psychol.* 58, 250–271. doi: 10.1016/j.cogpsych.2008.08.002

Wolf, M., and Katzir-Cohen, T. (2001). Reading fluency and its intervention. *Sci. Stud. Read.* 5, 211–239. doi: 10.1207/S1532799XSSR0503_2

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2016 Nicenboim, Logačev, Gattei and Vasishth. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 10 April 2015
doi: 10.3389/fpsyg.2015.00384

Hyper-active gap filling

Akira Omaki[1], Ellen F. Lau[2], Imogen Davidson White[2], Myles L. Dakan[2], Aaron Apple[1] and Colin Phillips[2]*

[1] Department of Cognitive Science, Johns Hopkins University, Baltimore, MD, USA, [2] Department of Linguistics, University of Maryland, College Park, MD, USA

Much work has demonstrated that speakers of verb-final languages are able to construct rich syntactic representations in advance of verb information. This may reflect general architectural properties of the language processor, or it may only reflect a language-specific adaptation to the demands of verb-finality. The present study addresses this issue by examining whether speakers of a verb-medial language (English) wait to consult verb transitivity information before constructing filler-gap dependencies, where internal arguments are fronted and hence precede the verb. This configuration makes it possible to investigate whether the parser actively makes representational commitments on the gap position before verb transitivity information becomes available. A key prediction of the view that rich pre-verbal structure building is a general architectural property is that speakers of verb-medial languages should predictively construct dependencies in advance of verb transitivity information, and therefore that disruption should be observed when the verb has intransitive subcategorization frames that are incompatible with the predicted structure. In three reading experiments (self-paced and eye-tracking) that manipulated verb transitivity, we found evidence for reading disruption when the verb was intransitive, although no such reading difficulty was observed when the critical verb was embedded inside a syntactic island structure, which blocks filler-gap dependency completion. These results are consistent with the hypothesis that in English, as in verb-final languages, information from preverbal noun phrases is sufficient to trigger active dependency completion without having access to verb transitivity information.

Keywords: filler-gap dependency, active gap filling, prediction, verb transitivity, island, plausibility mismatch effects, eye-tracking

OPEN ACCESS

Edited by:
Claudia Felser, University of Potsdam, Germany

Reviewed by:
Arild Hestvik, University of Delaware, USA
Oliver Boxell, University of Potsdam, Germany

****Correspondence:***
Akira Omaki, Department of Cognitive Science, Johns Hopkins University, 3400 North Charles Street, 237 Krieger, Baltimore, MD 21218, USA
omaki@jhu.edu

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: *22 December 2014*
Accepted: *18 March 2015*
Published: *10 April 2015*

Citation:
Omaki A, Lau EF, Davidson White I, Dakan ML, Apple A and Phillips C (2015) Hyper-active gap filling. Front. Psychol. 6:384. doi: 10.3389/fpsyg.2015.00384

Introduction

A leading goal of sentence processing research is to understand how the parser adapts to a multitude of linguistic differences across languages to enable successful comprehension. In this regard, comparisons of verb-medial and verb-final languages have provided a valuable source of evidence (Mazuka and Lust, 1990; Inoue and Fodor, 1995). The main verb contains rich information such as subcategorization and thematic role information that is critical for constructing structural analyses and interpretations (e.g., Chomsky, 1965; Grimshaw, 1990; Pollard and Sag, 1994; Levin and Rappaport Hovav, 1995). Much experimental evidence shows that the verb is a valuable source of information for parsing (e.g., Ford et al., 1982; Tanenhaus and Carlson, 1989; Boland et al., 1990; MacDonald et al., 1994; Spivey-Knowlton and Sedivy, 1995; Garnsey et al., 1997; Mauner and Koenig, 2000; Traxler et al., 2002; Blodgett and Boland, 2004; Snedeker and Trueswell, 2004).

The importance of the information from the verb head has engendered theoretical claims that structure building processes do not even start until the parser encounters the head of a phrase (e.g., verbal head) to be constructed, even in verb-final languages where this would be significantly delayed (Abney, 1989; Pritchett, 1992).

However, subsequent empirical research on verb-final languages like Japanese or German has generated evidence against such head-driven parsing theories in their strongest form, demonstrating that the parser uses various morphological and syntactic cues to incrementally build structures and interpretations in verb-final languages (Bader and Lasser, 1994; Koh, 1997; Clahsen and Featherston, 1999; Kamide and Mitchell, 1999; Konieczny, 2000; Bornkessel et al., 2002; Felser et al., 2003; Kamide et al., 2003; Aoshima et al., 2009; Yoshida, unpublished doctoral dissertation). Thus, although verb information strongly influences parsing decisions when available, speakers of verb-final languages often begin building syntactic and semantic structure in advance of the verb.

These findings raise the question of whether pre-verbal structure building reflects a language-specific adaptation to the processing demands of verb-finality, or rather a property of a general parsing architecture that speakers of all languages use. For example, consider less frequent cases in verb-medial languages where multiple arguments precede the verb. A classic example of this comes from processing of 'filler-gap' dependencies as illustrated by the relative clause construction shown in (1), where the object noun phrase (NP) *the city* (called the *filler*) is dislocated from the post-verbal thematic position (called the *gap*[1]), and the parser needs to associate the filler and the gap in order to assign a thematic interpretation.

(1) The city that the author visited _____ was named for an explorer.

It has been reported that speakers of verb-final languages complete filler-gap dependencies in advance of verb information, associating the filler with the earliest structural position where a thematic role could be assigned (pre-verbal object gap creation: Nakano et al., 2002; Aoshima et al., 2004). The current study examines whether this may also be the case in a verb-medial language like English, and whether pre-verbal gap creation is a language-general parsing procedure rather than an adaptation specific to verb-final languages. Under this hypothesis, we predict that English speakers should posit a gap irrespective of whether the verb ultimately licenses a direct object gap position, and that signs of reading disruption should be observed in cases where the verb does not accommodate a direct object.

We report the results of three on-line reading experiments in English that tested this prediction by examining the effect of verb transitivity on reading times in filler-gap configurations. The results are consistent with the hypothesis that the parser actively associates the filler with the verb in advance of the verb across languages, regardless of differences in verb positions. These results suggest that the procedure for filler-gap dependency completion may be uniform across languages, and are consistent with the view that the parser predictively constructs rich representations at the earliest possible moment in advance of critical bottom–up evidence.

[1]In this paper we use the 'gap-filling' or 'gap-creation' terminology in a theory-neutral way, as is typical in the psycholinguistic literature. This terminology should not be taken as indicating a commitment to representations that include gaps or traces; all of the processing theories we discuss here could be specified in terms of representations that do not include empty categories.

Background on Active Filler-Gap Dependency Processing

Past research on filler-gap dependency processing has established that the parser postulates a gap before there is sufficient bottom–up evidence to confirm that analysis (*Active gap filling:* Fodor, 1978; Crain and Fodor, 1985; Stowe, 1986; Frazier and Flores d'Arcais, 1989). For example, Stowe (1986) observed the so-called *Filled gap effect* in (2), i.e., slower reading times at the direct object position *us* in the wh-fronting condition (2a) than in a control condition that did not involve wh-fronting (2b). This pattern of reading times suggests that the parser had already posited a gap following the transitive verb, before checking whether the direct object position was occupied.

(2) a. My brother wanted to know who Ruth will bring us home to _____ at Christmas.
b. My brother wanted to know if Ruth will bring us home to Mom at Christmas.

Converging evidence comes from an eye-tracking experiment by Traxler and Pickering (1996), who manipulated the thematic fit between the filler and the potential verb host, as in (3).

(3) We like the city/book that the author *wrote* unceasingly and with great dedication about _____ while waiting for a contract.

Traxler and Pickering found a plausibility mismatch effect at the critical verb in (3), i.e., the first fixation time at the optionally transitive verb *wrote* increased when the filler was an implausible object of the verb (i.e., *the city*), compared to when the filler was a plausible object of the verb (i.e., *the book*). This suggests that at least as early as the verb position, the parser postulates a gap and analyzes the filler as the object of the verb, even when the filler is a poor semantic fit to that role. In fact, there is ample time course evidence for active object gap creation, using a variety of dependent measures such as reading time and gaze duration measures (Crain and Fodor, 1985; Frazier, 1987; Frazier and Clifton, 1989; de Vincenzi, 1991; Pickering and Traxler, 2001, 2003; Aoshima et al., 2004; Phillips, 2006; Wagers and Phillips, 2009), cross-modal priming (Nicol and Swinney, 1989; Nicol, 1993; Nakano et al., 2002), visual world eye-tracking (Sussman and Sedivy, 2003) as well as event-related potentials (Garnsey et al., 1989; Featherston et al., 2000; Kaan et al., 2000; Felser et al., 2003; Phillips et al., 2005; Gouvea et al., 2010).

The work summarized above may suggest that filler-gap dependency completion is triggered only after the parser gains access to the verb and confirms that the verb is transitive and is able to syntactically accommodate an object. However, evidence

that active dependency completion does not depend on verb information has been presented by studies that investigated (i) subject gap creation in English, as well as (ii) object gap creation in verb-final languages. For example, Lee (2004) used sentences like (4) to reveal a filled gap effect in the subject NP position.

(4) a. That is the laboratory which, on two different occasions, Irene used a courier to deliver the samples to ___.
b. That is the laboratory to which, on two different occasions, Irene used a courier to deliver the samples ___.

Here, the content of the wh-filler is manipulated in such a way that the wh-filler can plausibly be a subject (4a) or not (4b). The results showed a longer reading time at the subject NP *Irene* in (4a) than in (4b), suggesting that the parser had postulated a subject gap before encountering the actual subject NP. Although this interpretation has been challenged (Staub, 2010), it would in any case not be surprising that the parser actively creates a subject gap without having access to verb information, given that a subject is present in any sentence, regardless of verb properties. In this sense, if verb information were to play a role in the parser's attempt to posit a gap, the critical empirical evidence should come from dependency completion at the object position, where the presence or absence of an object gap relies on properties of the verb.

Evidence for pre-verbal object gap creation has been reported for verb-final languages like Japanese in which the object gap position linearly precedes the verb. For example, Aoshima et al. (2004) examined processing of scrambling sentences in which a dative object NP was dislocated to the sentence initial position, and found a filled gap effect at a pre-verbal dative object position for the first verb phrase (VP) in the sentence (see also Omaki et al., 2014). Using similar sentences, Nakano et al. (2002) reported evidence for an antecedent priming effect for the scrambled NP at a pre-verbal gap position, although the priming effect was only found in the high working memory span group. These data indicate that the parser can in principle complete filler-gap dependencies before accessing verb information.

In verb-medial languages, no such evidence for pre-verbal object gap creation has been reported to date. This may reflect a real difference between languages in processing strategy, and pre-verbal object gap creation in verb-final languages may reflect the parser's adaptation to the demands of processing these languages. Maintaining a structurally unintegrated filler in memory has been argued to impose a burden on working memory (King and Just, 1991; Gibson, 1998; Gordon et al., 2002; Haarmann and Cameron, 2005). Alternatively, the parser may be architecturally constrained to assign a thematic interpretation to the filler as soon as possible (Pickering and Barry, 1991; Aoshima et al., 2004). On this view, the parser should prioritize integrating the filler into the first grammatically permissible structural position that can potentially receive a thematic role. Given that filler-gap dependencies are potentially unbounded, waiting for the verb before constructing the ultimate object gap position could impose a large processing burden on speakers of verb-final languages.

In verb-medial languages like English, verbs become available relatively earlier in the sentence, such that the average working memory cost of waiting for the verb would be less than in verb-final languages. The advantage of waiting for the verb information is that the parser can reduce the likelihood of making risky commitments, because the verb may turn out to be intransitive and disallow an object NP analysis for the filler. In English, therefore, the parser may create an object gap position only after the verb is confirmed to be transitive. This still constitutes active gap filling, in the respect that the ultimate gap position may turn out to be somewhere later than the object position [e.g., after a late-arriving preposition gap, as in (2) and (3)]. Let us call this a *conservative active gap filling* mechanism, since the bottom–up subcategorization information from the verb still plays a critical role in the parser's decision on whether to postulate an object gap or not. This view of active gap filling is rather standard for explaining filler-gap dependency completion in verb medial languages like English. For example, McElree and Griffith (1998) and McElree et al. (2003) have argued that the dependency completion process is triggered when the parser accesses information from the verb and initiates the retrieval process for the filler that is stored in working memory (see also Pickering and Barry, 1991; Lewis and Vasishth, 2005).

On the other hand, pre-verbal object gap creation in verb-final languages may reflect a language-general property of the processing architecture, although evidence for such mechanisms may be simply more difficult to obtain in verb-medial languages. In the English filler-gap case, for example, in any parser that adopts some form of left-corner strategy (Kimball, 1975; Abney and Johnson, 1991; Resnik, 1992; Shieber and Johnson, 1993; Stabler, 1994; Crocker, 1996; Lewis and Vasishth, 2005; Gibson, unpublished doctoral dissertation), the presence of the subject NP allows the parser to predict the presence of a VP. Given that a VP can contain an object NP position, the parser could project a VP with an object NP slot and assign the filler to this object position before confirming whether the upcoming verb is a transitive verb or not. Let us call this a *hyper-active gap filling* mechanism, because this involves a more risky predictive structure building process than is standardly assumed for active object gap creation in English. Filler retrieval and structural integration is still integral to the hyper-active gap filling mechanism, but the crucial difference is in what information triggers retrieval and integration, and consequently, at what point in the sentence this process is executed.

It is important to note that either of these two active gap filling mechanisms is compatible with the existing data on active object gap creation reviewed above. A filled gap effect only indicates that the gap had been created before the actual object NP is processed, and this result is compatible with both accounts, given that both hyper-active gap filling and conservative active gap filling mechanisms assume that object NP gap creation happens before or on the verb. A plausibility mismatch effect indicates that when the verb is potentially transitive, then the semantic fit between the filler and the verb is immediately assessed. This is also predicted by both accounts. The assessment of the semantic relation between the filler and the verb requires the parser to access the content of the verb, by which point the object gap

position should have been created on either account. Thus, neither paradigm allows us to tease apart the two hypotheses on what kind of information is sufficient for triggering object gap creation.

In the current study we aim to tease apart the predictions of two hypothesized mechanisms for active object gap creation processes. If English speakers construct the gap site before encountering the verb, just like speakers of verb-final languages, then disruption should be observed in filler-gap configurations when the verb turns out to be intransitive, relative to transitive verbs (e.g., *The party that the student arrived/planned. . .*). According to the conservative active gap filling mechanism outlined above, the parser waits for a transitive verb before postulating the corresponding gap structure. Here, no disruption is expected at an intransitive verb, since the parser has not postulated a gap that would require a transitive verb.

Two previous studies are relevant to the two hypotheses about active object gap creation in English. Previous work by Pickering and Traxler (2003) examined the effect of subcategorization frequency in optionally transitive verbs (e.g., *Those are the lines/props that the author spoke [about]. . .*). It was found that readers did not take subcategorization frequency into account in deciding where to posit a gap, as there was a strong preference to posit a gap in the verb object position (NP complement) even with verbs that more frequently take a PP complement. The absence of subcategorization frequency effect in active object gap creation could be taken to indicate that verb information is not relevant for object gap creation processes. However, all of the verbs in Pickering and Traxler's study could grammatically accommodate an NP complement, and the parser may therefore have relied on the transitivity information of the verb to create an object gap. Therefore, this finding does not distinguish the predictions of the two proposed mechanisms for active object gap creation.

To our knowledge, the only previous test of these two active object gap creation hypotheses is in Experiment 3 of Staub (2007). The test sentences in this experiment (5a–d) manipulated the transitivity of the verb (*called* vs. *arrived*) and sentence structure (relative clause with a gap vs. simple declarative with no gap). The filler was manipulated to be an implausible object of the transitive verb (*gadget-called*). Under the hyper-active gap filling hypothesis, the parser in effect predicts the presence of a transitive verb, and therefore the reading processes in the gap conditions should be disrupted in either intransitive or transitive condition, but for different reasons: when the verb turns out to be intransitive, and processing should also be disrupted when the verb is transitive because of the plausibility mismatch effect. On the other hand, the conservative active gap filling mechanism postulates a gap only after checking whether the verb is capable of hosting an object NP, and therefore reading disruption is predicted only in the transitive gap condition due to the plausibility mismatch effect.

(5) a. The gadget that the manager *called* occasionally about. . .
b. The manager *called* occasionally about the gadget . . .
c. The party that the student *arrived* promptly for . . .
d. The student *arrived* promptly for the party . . .

Staub (2007) found longer first-fixation durations in the transitive gap condition (5a) than in the transitive no-gap condition (5b), but no such difference was observed between the intransitive gap and no-gap conditions (5c) and (5d). This pattern of data supports the prediction of the conservative active gap filling hypothesis, suggesting that the parser does not create an object gap until it checks the transitivity information of the verb. One concern about this design, however, is whether the no-gap condition was truly a neutral baseline against which a transitivity mismatch could be measured, as the gap and no-gap conditions differed substantially in both the linear and structural position of the verb. As Staub (2007) points out, one piece of data suggesting that the control may not have been completely neutral is the fact that reading times on the intransitives were numerically (but non-significantly) shorter in the gap condition than in the no-gap condition. It is important to note here that the gap conditions (5a) and (5c) contain an extra NP (i.e., the head of the relative clause) prior to the critical verb region in comparison to the no-gap conditions (5b) and (5d). This may have led to a difference in the amount of contextual information available prior to the verb. Increased contextual information can facilitate processing for subsequent lexical items (Stanovich and West, 1983; Van Petten and Kutas, 1990; Kutas and Federmeier, 2000), and for this reason, lexical access for the intransitive verb in the gap condition may have become faster and masked the potential reading time slowdown associated with the structural manipulation. In an attempt to provide a better test of the predictions of the hyper-active and conservative active gap filling accounts, the current study used relative clause islands as a control condition, which allowed the target sentences to more closely match in informational content and word position.

Experiment 1

Experiment 1 was a self-paced reading study that was designed to test the predictions of the hyper-active and conservative active gap filling hypotheses, while addressing methodological concerns about previous work. We employed the transitivity mismatch paradigm used in Staub (2007) in order to test whether a verb transitivity manipulation affects reading time at the verb. Critically, in the baseline conditions the critical verb was embedded inside a relative clause structure, a syntactic 'island' domain that prohibits filler-gap dependency formation (Ross, unpublished doctoral dissertation; for a review, see Szabolcsi and den Dikken, 2003). A sample set of stimuli is shown in **Table 1**.

A number of previous studies have shown that the parser respects island constraints in real-time syntactic processing, such that it avoids actively constructing filler-gap dependencies that span syntactic island boundaries (Stowe, 1986; Kluender and Kutas, 1993; McKinnon and Osterhout, 1996; Traxler and Pickering, 1996; McElree and Griffith, 1998; Wagers and Phillips, 2009; Omaki and Schulz, 2011; Yoshida, unpublished doctoral dissertation). The relative clause island condition thus provided a baseline measure of reading times for the critical transitive and intransitive verbs, independent of processes of filler-gap

TABLE 1 | Sample materials and conditions for Experiment 1.

	Analysis regions										
	1	2	3	4	5	6	7	8	9	10	11
Transitive, non-island	The	city	that	the	author		wrote	regularly	about	was	named for an explorer
Transitive, island	The	city	that	the	author	who	wrote	regularly	saw	was	named for an explorer
Intransitive, non-island	The	city	that	the	author		chatted	regularly	about	was	named for an explorer
Intransitive, island	The	city	that	the	author	who	chatted	regularly	saw	was	named for an explorer
Example question	Was the city named for an explorer?										

dependency completion. The use of island configurations allowed us to address the methodological concerns with previous work.

First, this design allowed the baseline condition to present a filler NP prior to the critical region, such that the same amount of contextual information from the lexical items was present in advance of the critical verb region across the four conditions. Second, the word position for the critical regions (Regions 7 and 8 in **Table 1**) was closely matched across conditions (word positions 6 and 7 in the non-island conditions, word positions 7 and 8 in the island conditions), and it was also placed away from the early portion of the sentence.

Furthermore, following Staub's design, we selected transitive verbs that are implausible hosts for the filler. Under this design, the hyper-active gap filling hypothesis predicted a reading time slowdown in both the non-island transitive and the non-island intransitive conditions relative to their island counterparts, but for a different reason in the two cases. In the transitive condition, the slowdown would reflect a plausibility mismatch effect triggered by the poor semantic fit between the filler and the verb. In the intransitive condition, the slowdown would result from a transitivity mismatch effect due to the mismatch between the expected subcategorization property of the verb (i.e., transitive) and the actual subcategorization property of the verb. On the other hand, the conservative active gap filling hypothesis predicted an interaction. A reading time contrast should be observed between the non-island transitive condition and the island transitive condition due to the plausibility mismatch effect, but no corresponding contrast should be observed between the two intransitive conditions, given that the parser should not actively create an object gap in either condition. Note that the lexical difference in the critical verb region across conditions was not problematic, since the critical contrast was between non-island and island conditions within each verb type.

Method

Participants

We recruited 32 native speakers of American English from the University of Maryland community. They received a course credit or were paid $10 for their participation and were naïve to the purpose of the experiment.

Materials

We used 28 sets of four sentences like those shown in **Table 1**. All of the stimuli from experiments reported in this paper are made available in Supplementary Materials. The transitive non-island and island conditions were taken from the implausible semantic fit conditions in Omaki and Schulz (2011), who used a modified version of the plausibility manipulation materials from Traxler and Pickering (1996). Omaki and Schulz replicated Traxler and Pickering's plausibility mismatch effect with native and non-native speakers alike, confirming that the semantic fit between the filler and the verb affects the reading time for the verb when the verb is in a gap filling (i.e., non-island) environment, but not when the verb is inside a relative clause island. Critically, it was also found that the implausible verb-filler combination in a non-island environment (e.g., *city-wrote*) led to a significant slow down at the verb compared to its island counterpart with the same implausible verb-filler combination. Thus, even though the current experiment did not include a plausible counterpart of the implausible transitive verb condition, we could be confident that a reading time contrast between the transitive non-island and island conditions results from the semantic misfit between the filler and the verb. In other words, the finding in Omaki and Schulz's study supports the notion that island conditions in general can be used as baseline conditions for a reading disruption associated with active object gap creation. The intransitive conditions were modeled after the transitive conditions by replacing the optionally transitive verb with unergative or unaccusative intransitive verbs (Levin and Rappaport Hovav, 1995).

The non-island and island conditions differed in the number of relative clauses. The non-island condition had only one relative clause (*the city that the author wrote/chatted regularly about*), such that the object position of the verb *wrote/chatted* was the first potential gap position after the embedded subject was encountered. In the island conditions, the critical verb was embedded inside another relative clause *the author who wrote/chatted regularly*, such that linearly this was still the first verb but grammatically the filler should not be accessible to the verb due to the relative clause island constraint. Thus, the first verb served as the critical region for testing the plausibility and transitivity mismatch effects. All the transitive verbs were optionally transitive, such that the sentences in the island conditions were all ultimately grammatical. The subcategorization frequency of the optionally transitive verbs was not controlled, since Pickering and Traxler (2003) have demonstrated that plausibility mismatch effects are attested for optionally transitive verbs regardless of subcategorization frequency. In all four conditions the same adverb immediately followed the verb, making

it possible to observe potential spill-over effects. The 28 sentence sets were counter-balanced across four lists so that each participant saw only one version of the target items and consequently read seven tokens of each condition. In addition, 72 fillers of similar length and complexity were constructed and added to each list.

Procedure

The self-paced reading task was implemented on the Linger software developed by Doug Rohde (http://tedlab.mit.edu/~dr/Linger/). We used a word-by-word, non-cumulative moving window presentation (Just et al., 1982). In this design, each sentence initially appears as a series of dashes, and these dashes are replaced by a word from left to right every time the participant presses the space bar. In order to ensure that the participants were paying attention while reading the sentences, all sentences were followed by yes-no comprehension questions, and feedback was provided if the questions were answered incorrectly. Comprehension questions never addressed the critical filler-gap portion of the sentence. At the beginning of the experiment, participants were instructed to read at a natural pace and to answer the questions as accurately as possible. Seven practice items preceded the self-paced reading experiment, and the order of presentation was randomized for each participant. The experiment took ~30 min. The experiment protocol for this study was approved by the Institutional Review Board at the University of Maryland.

Data Analysis

The data from two items were excluded from analyses due to coding errors. Only trials in which the comprehension question was answered accurately were included in the analysis, which affected 5.7% of the trials. We also analyzed the data without excluding the trials based on comprehension accuracy, but the overall pattern of results did not change.

Self-paced reading times for the target sentences were examined for each successive region, although the words after the auxiliary *was* were combined into a single region because these lay beyond the critical regions and were unlikely to show effects relevant for the critical manipulation. The critical regions where a potential plausibility or transitivity mismatch effect was expected consist of Region 7 (i.e., the verb *wrote/chatted*) and the following Region 8 (i.e., the adverb *regularly*), in which spill-over effects could be observed. Regions 1 through 6 were predicted to show no difference across conditions, since they were lexically matched. Regions 9 through 11 could reveal reading time differences after the filler-gap dependency is completed (Region 9 hosts the true gap site), and with a possible additional difference in the island conditions due to the structural complexity associated with the extra relative clause in these conditions.

Reading time data that exceeded three standard deviations from the group mean at each region and in each condition were excluded, affecting 1.7% of the data. The remaining reading time data were analyzed using linear mixed effects models (Baayen et al., 2008). These analyses were conducted in the R environment (R Development Core Team, 2011), using the lme4 package for R (Bates et al., 2014). The fixed effects of island structure type (non-island vs. island) and verb transitivity (transitive vs. intransitive) were coded using sum contrasts, with one level of the factor coded as −0.5, and the other as 0.5. This sum contrast coding makes the mixed effect model estimates roughly comparable to the actual average reading time contrasts. The model included random intercepts for participants and items. For random slopes, we used the following procedure to determine the optimal random effect structure (for discussions: Jaeger, 2011; Barr et al., 2013). First, we constructed a fully crossed model that included the fixed effects and an interaction term as random slopes for both participants and items. This fully specified model failed to converge, plausibly due to the complexity of the model and missing data points in some of the trials (Barr et al., 2013). Next, we simplified the random effect structure by only keeping the verb transitivity factor as a random slope for participants and items. In our experimental design, the island structure is invariant across all items, and it is also known to be robust across individuals, regardless of working memory capacity (see Sprouse et al., 2012). On the other hand, the verbs differed across items, and it is possible that the subcategorization bias differs across participants. This mixed effects model converged for all regions. We computed p values for linear mixed effects models using the lmerTest R package (Kuznetsova et al., 2014).

Results

Comprehension accuracy

The mean comprehension question accuracy for experimental items across participants and items was 93.0%. For the non-island conditions, the transitive items were answered with an accuracy of 93.7% (SE = 1.9), and the intransitive items with an accuracy of 94.6% (SE = 1.4). For the island conditions, the transitive items were answered with an accuracy of 91.5% (SE = 1.7), and the intransitive items with an accuracy of 92.0% (SE = 2.2). The mean accuracy did not differ reliably across conditions, although the fact that the mean accuracy for island conditions was numerically lower may reflect the complexity difference between non-island and island conditions.

Reading time data

The region-by-region mean reading time for the transitive conditions is presented in **Figure 1**, and the mean region-by-region reading time for the intransitive conditions is presented in **Figure 2**.

In the non-critical Regions 1–6, there were no significant differences in Regions 1, 2, 4–6 (ps > 0.06). In Region 3 there was a main effect of verb type (Estimate = −17.3, SE = 7.6, $t = -2.27$, $p < 0.05$), due to slower reading times in the transitive conditions than in the intransitive conditions (381 vs. 358 ms). Since this region was lexically matched across conditions, we conclude that this is a spurious effect. But given that the effect was small and occurred well ahead of the critical regions, this unexpected effect was unlikely to have impacted the observations in the critical regions.

At the critical verb in Region 7 there were no significant differences (ps > 0.1). The following spill-over region

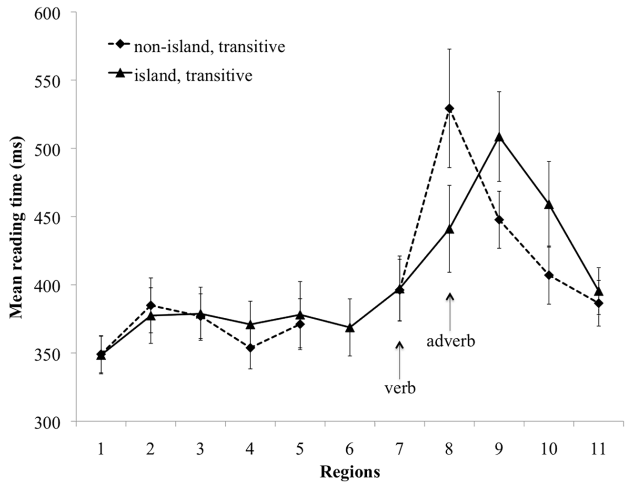


FIGURE 1 | Mean reading time (ms) for the transitive non-island and island conditions. Error bars indicate standard error of the mean.

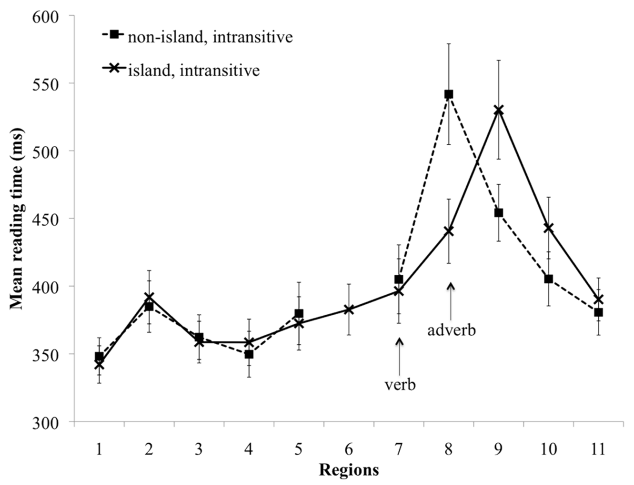


FIGURE 2 | Mean reading time (ms) for the intransitive non-island and island conditions. Error bars indicate standard error of the mean.

(Region 8) revealed no main effect of verb type, but there was a main effect of structure type (Estimate = −92.0, SE = 16.4, $t = -5.61$, $p < 0.001$), reflecting the fact that the non-island conditions produced significantly slower reading times than the island conditions (529 vs. 435 ms). There was no significant interaction of verb type and structure type ($p > 0.1$).

Region 9 consisted of a second verb in the island conditions and a preposition in the non-island conditions. We observed a main effect of structure type in Region 9 (Estimate = 63.7, SE = 15.9, $t = 4.01$, $p < 0.001$), as well as in Region 10 (Estimate = 46.1, SE = 11.5, $t = 4.0$, $p < 0.001$), in these cases due to slower reading times in the island conditions (Region 9: 519 vs. 451 ms, Region 10: 451 vs. 406 ms). Region 11 revealed no significant differences (ps > 0.09).

Discussion

In Experiment 1, we tested the predictions of two hypotheses about active object gap creation. The hyper-active gap filling hypothesis predicted the presence of reading disruption at intransitive verbs, because encountering an intransitive verb in a filler-gap context would be incompatible with the object gap structure constructed earlier. On the other hand, the conservative active gap filling hypothesis predicted no such reading disruption, because the parser should first consult the transitivity information of the verb to decide whether to posit an object gap or not. As a baseline for estimating the degree of disruption at the verb, we used relative clause island constructions, which block the association of the filler with the critical verb. The results were consistent with the predictions of the hyper-active account: in the region following the verb, we observed slower reading times for intransitive verbs in non-island conditions than in corresponding island conditions.

Previous work has shown a filler-gap plausibility mismatch effect at the verb such that mismatched transitive verbs in a non-island environment elicit longer reading times than their plausible non-island or plausible/implausible island counterparts (Traxler and Pickering, 1996; Omaki and Schulz, 2011), and here we replicated this finding. This effect can be interpreted as the result of active association of the filler with the transitive verb, which in these stimuli resulted in a verb–object plausibility mismatch. On the other hand, the slowdown observed in the intransitive non-island condition relative to the intransitive island condition can be interpreted as a *transitivity* mismatch. This suggests that the parser does not wait for bottom–up evidence from the verb that the verb can syntactically license a gap, but rather attempts to construct the dependency before this information is available. This slowdown cannot reflect the cost of maintaining the filler in working memory, because a filler is also being maintained at this position in the baseline island condition.

It is also important to note that the shorter reading times in the critical regions of the island conditions are theoretically informative. These findings suggest that the reading time increase in the non-island conditions is specifically due to an expectation violation following premature gap creation. A plausible alternative explanation of the reading disruption in the non-island conditions is that it reflects a more general cost associated with delaying gap creation decisions. Under this alternative account, we should expect to observe reading disruption in the island conditions as well, because gap creation must wait until the verb that follows the relative clause island region (e.g., *saw* in Region 9). However, this prediction is not supported by the data, as the reading time in the adverb region (Region 8) of the island conditions was reliably shorter than in non-island conditions.

In Regions 9 and 10, the island conditions were read more slowly for both levels of verb type. Region 9 corresponds to the word that licensed the true gap site across all conditions, and hence this slowdown could reflect a difference in the so-called integration cost (Gibson, 1998, 2000) between non-island and island conditions. Previous work on filler-gap dependency processing has demonstrated that increased complexity and length differences result in increased processing difficulties at the gap

site, as measured by reading time (Gibson and Warren, 2004; Wagers and Phillips, 2014) and reduced accuracy in speeded acceptability judgment tasks (McElree et al., 2003). However, the reading time difference in Region 9 may simply be due to lexical differences (prepositions in the non-island conditions vs. verbs in the island conditions), so the reading time contrast between the island and non-island conditions may not reflect an integration cost difference.

Note that it is unlikely that the reading time contrast between non-island and island conditions in Region 8 is related to the overall complexity of the constructions used in our stimuli, given that on all accounts that we are aware of, island domains have been argued to be syntactically more complex and more taxing for working memory resources (Deane, 1991; Kluender and Kutas, 1993; Kluender, 1998, 2004; Hofmeister and Sag, 2010). The fact that the putatively less complex non-island conditions were read more slowly allows us to attribute the slowdown to processes that uniquely occur in the non-island conditions, namely filler-verb association.

In summary, the presence of both a plausibility mismatch effect and a transitivity mismatch effect lends support to the hyper-active gap filling hypothesis, and argues against a conservative active gap filling hypothesis under which transitivity information is consulted before attempting to create an object gap. This finding directly contrasts with that of Staub (2007), who did not find evidence for a transitivity mismatch effect.

However, this conclusion is not warranted until two methodological concerns are addressed. First, the design in Experiment 1 was modeled after Staub (2007), who used a plausibility mismatch design for transitive verb conditions, and transitivity mismatch design for intransitive verb conditions. Our findings differed from Staub's as we found mismatch effects for both transitive and intransitive non-island conditions, but it is possible that some nuisance factor common to both non-island conditions led to a slow-down across the board. Stronger evidence for the hyper-active gap filling hypothesis can be obtained if we replicate the transitivity mismatch slowdown in the intransitive non-island condition, while at the same time observing no reading disruption in the transitive non-island condition. Experiment 2 accomplished this by making the filler and the verb semantically fit in the transitive conditions. The absence of reading disruption in the transitive conditions would suggest that the disruption in the non-island, intransitive condition is due to the intransitivity of the verb.

Second, it is important to note that our evidence for reading disruption for transitive and intransitive verbs (i.e., the slowdown in non-island conditions compared to island conditions) was not observed until the spill-over adverb region. Spill-over effects are widely observed in self-paced reading experiments, and it is thus common to attribute spill-over effects to processes triggered in a preceding region. However, in our experiment there is an alternative explanation for the effect in the adverb region that would not require hyper-active gap filling. For the intransitive condition, the slowdown in the adverb region could indicate that the parser had expected the presence of a preposition, which would allow structural integration of the filler. Under this alternative account, the slowdown is not due to a transitivity mismatch on the verb, but rather to a word category expectation mismatch in the adverb region that was triggered by the verb itself. This account is consistent with the conservative active gap filling hypothesis, since the parser's expectation regarding filler-gap dependency completion is based on the information from the verb. Incidentally, the reading disruption observed in the transitive conditions of Staub (2007) was at the verb region. One possible reason for this discrepancy is the difference in the dependent measure: Staub (2007) used an eye-tracking during reading method while we used self-paced reading in Experiment 1. An eye-tracking during reading method generally provides better temporal precision than the self-paced reading method (Rayner, 1998; Rayner and Pollatsek, 2006). Thus, an eye-tracking replication of Experiment 1 may yield a transitivity mismatch effect on the verb region, and provide stronger evidence for the hyper-active gap filling hypothesis. This is addressed in Experiment 2.

Experiment 2

Experiment 2 addressed two methodological concerns raised in Experiment 1 by removing sources of slowdown in the transitive conditions, and also by using the eye-tracking during reading method.

Method

Participants

We recruited 33 native speakers of American English from the Johns Hopkins University community, but data from one participant were removed due to calibration errors. Participants received course credit or $10 for their participation. They were all naïve to the purpose of the experiment.

Materials

We used 28 sets of four sentences as shown in **Table 2**. This experiment used the same transitivity mismatch logic as Experiment 1 and manipulated the verb transitivity type (intransitive vs. transitive). However, in this experiment the semantic fit between the filler and the transitive verb was always plausible, such that no reading disruption was expected at the transitive verb in the non-island condition. As in Experiment 1 we manipulated structure type (non-island vs. island), using conditions with relative clause island structures as baseline conditions. Relative clause islands provide an effective baseline, since they include the same filler NP and other lexical material as the non-island condition, while preventing dependency completion at the critical verb. As in Experiment 1, the transitive verbs were optionally transitive and the true gap position occurred outside the island domain, allowing the sentence to continue grammatically.

The 28 sentence sets were counter-balanced across four lists so that each participant saw only one version of the target items and consequently read seven tokens of each condition. In addition, 76 fillers of similar length and complexity were constructed and added to each list.

Procedure

An Eyelink 1000 eye-tracker (SR Research: Mississauga, ON, Canada) was used to record eye movements. The participant's

TABLE 2 | Sample materials and conditions for Experiment 2.

	Analysis regions				
	Sentence initial	**Pre-verb**	**Verb**	**Post-verb**	**Sentence final**
Transitive, non-island	The book that	the author	wrote	regularly	about was named for an explorer
Transitive, island	The book that	the author who	wrote	regularly	saw was named for an explorer
Intransitive, non-island	The book that	the author	chatted	regularly	about was named for an explorer
Intransitive, island	The book that	the author who	chatted	regularly	saw was named for an explorer
Example question	Was the book named for an explorer?				

head was stabilized by a chin rest and a forehead rest. The position of the right eye only was monitored at a sampling rate of 1000 Hz. The eye-tracker display allowed a maximum of 120 characters per line, in 10 pt Monaco font. Some filler sentences were displayed on two lines, but all target sentences were displayed on one line. Stimuli were displayed on a 21.5-inch Samsung SyncMaster monitor, and participants were seated 65 cm from the computer screen. Before the experiment started, participants were seated in front of the eye-tracker and received instructions for the experiment. A calibration routine was performed at the beginning of the experiment, and the experimenter monitored the calibration accuracy throughout the session, recalibrating when necessary. The experiment started with written instruction on the display and seven practice trials. At the beginning of each trial, a black circle was displayed on the left side of the monitor, which corresponded to the location of the beginning of the sentence. The text was displayed after the participant successfully fixated on the circle. After reading each sentence, the participant pressed a button to remove the sentence display. Each sentence was followed by a yes-no comprehension question, and the participant answered the comprehension question by pressing a left or right button. Comprehension questions never addressed the critical filler-gap portion of the sentence. The entire experiment lasted ∼35 min. The experiment protocol for this study was approved by the Homewood Institutional Review Board at the Johns Hopkins University.

Data Analysis

Comprehension accuracy for the target trials was 90.7%, and trials in which participants answered the comprehension question incorrectly were removed from the eye movement analyses, as data from these trials may reflect inattentive reading. For the remaining data, an automatic procedure pooled short contiguous fixations. The procedure incorporated fixations of less than 80 ms into larger fixations when they occurred within one character of each other and deleted any remaining fixations of less than 80 ms, because little information can be extracted during such short fixations (Rayner and Pollatsek, 1989). Unusually long fixations greater than 800 ms were also removed, because they usually reflect tracker losses or other anomalous events. This procedure resulted in the exclusion of 4.86% of all fixations.

For the purpose of analysis of the eye movement data, the sentences were divided into five analysis regions, as shown in **Table 2**. We report eye movement data in the following three regions: (a) the pre-verb region (*the author* in non-island conditions, *the author who* in island conditions), in order to ensure that there were no unexpected reading behavior differences that might compromise the interpretation of the data from the critical region, (b) the verb region, which is the critical region where potential transitivity mismatch effects might be observed, and (c) the post-verb region, which corresponds to the post-verbal adverb and could be used to probe for potential spill-over effects. The data in the remaining regions are not reported, because reading times at these regions are not critical for distinguishing the competing hypotheses. Moreover, after the post-verb region, the lexical items were not held constant across conditions and therefore any observed differences would be difficult to interpret. The island conditions contained one extra word, i.e., the relative pronoun (e.g., *who*), which could have affected reading times in the pre-verb region as well as regression measures for subsequent regions.

Following the data analysis procedures used in Staub (2007), four reading time measures were computed for the three regions of interests: *first fixation duration, first pass time, regression path time,* and *percent regressions* (Rayner, 1998; Rayner and Pollatsek, 2006; Staub and Rayner, 2007). First fixation duration is the duration of the very first fixation in a region, regardless of whether there is a single word or multiple words in that region. This measure is often used as an index of lexical difficulty (e.g., Reichle et al., 2003) but is also informative about the earliest syntactic processes that immediately follow lexical access (e.g., Frazier and Rayner, 1982; Sturt, 2003).

The *first-pass reading time* is calculated by summing the fixations in a region between the time when the eye-gaze first enters the region from the left and the time when the eye-gaze exits the region either to the left or the right. First-pass reading times also index early lexical and syntactic processes associated with a region, but given that they consist of multiple fixations on the same region, they may also reflect slightly later processes than the first fixation measure.

Regression path times are the sum of fixations from the time when the eye-gaze first enters a region from the left to the time when the eye-gaze exits the region to the right. Regression path time is identical to first-pass reading time if the eye-gaze first exits the region to the right, but if the eye-gaze exits the region to the left, then regression path times are longer than the first-pass time as they include all fixations in previous regions as well as re-fixations on the region before exiting the region to the right. Thus, regression path times are likely to reflect slightly later processes, such as integration of the critical region with the preceding

context. The *percent regressions* indicate the probability that a reader made a regressive eye movement to preceding regions after fixating a given region. This measure includes only regressions made during the reader's first pass through the region, and does not include regression made after re-fixating the region.

Reading time data (i.e., first fixation, first pass, and regression path durations) were analyzed using linear mixed effects models (Baayen et al., 2008), and percent regressions were analyzed by mixed effects logistic regression, as the dependent measure was categorical (see Jaeger, 2008). The mixed effects models included random intercepts for participants and items. We used the same procedure as Experiment 1 to simplify the random slope structure until the models converged in all regions and eye movement measures. This procedure led us to adopt verb transitivity as a random slope for participants and items for all fixation measures and regions, except for percent regression measures in the post-verb region. Here, we removed the verb transitivity random slope for participants, as the transitivity bias variance across different verbs (if any) is more likely to influence the data than variance in participants' experience with the verbs.

When the critical region demonstrated a significant interaction of verb and structure type, a planned comparison was conducted with separate mixed effects models to test for systematic differences between the island and non-island conditions within each verb type. These models included participants and items as random intercepts.

Results

Table 3 presents the participant means on each measure for each region as well as the standard errors of the participant means, and **Table 4** presents a summary of the statistical analyses.

TABLE 3 | Experiment 2 participant mean reading times in milliseconds (standard error) and percent regressions.

Measure	Pre-verb region	Verb region	Post-verb region
First fixation			
Transitive, non-island	212 (8)	249 (12)	242 (9)
Transitive, island	217 (7)	240 (7)	243 (9)
Intransitive, non-island	207 (8)	256 (10)	246 (10)
Intransitive, island	208 (5)	231 (8)	237 (7)
First-pass time			
Transitive, non-island	287 (14)	277 (13)	283 (13)
Transitive, island	386 (20)	275 (10)	287 (12)
Intransitive, non-island	299 (15)	303 (13)	296 (14)
Intransitive, island	396 (19)	266 (11)	284 (10)
Regression path time			
Transitive, non-island	463 (28)	373 (24)	402 (30)
Transitive, island	636 (43)	406 (31)	447 (35)
Intransitive, non-island	472 (38)	397 (23)	492 (35)
Intransitive, island	619 (41)	425 (38)	469 (26)
Percent regressions			
Transitive, non-island	33.1 (5.0)	17.1 (3.5)	17.9 (3.4)
Transitive, island	33.2 (4.0)	23.0 (4.4)	24.4 (3.4)
Intransitive, non-island	27.1 (4.8)	16.2 (2.8)	27.5 (3.7)
Intransitive, island	32.7 (4.7)	24.4 (3.7)	22.4 (3.1)

In the pre-verb region, the first pass time and regression path measures showed a main effect of structure ($p < 0.001$), with longer reading times in the island conditions than in the non-island conditions. This effect was expected because the pre-verb region in the island conditions contained the extra word *who*, which made it more likely to attract multiple fixations in that region. No other significant effects were observed in this region.

In the verb region, evidence for the hyper-active gap filling hypothesis was found in first fixation durations as well as in first pass measures. Both measures showed a main effect of structure with longer reading times for non-island conditions ($ps < 0.05$). First fixation durations showed a marginal interaction of structure and verb transitivity ($p = 0.06$), and first pass times showed a significant interaction ($p < 0.05$). Planned pairwise comparisons on first fixation durations and first pass times revealed that reading times in the non-island, intransitive condition were significantly longer than in the island, intransitive condition ($ps < 0.01$), but no significant difference was observed between the transitive conditions. No significant effect was observed for the regression path durations. There was a main effect of structure in percent regressions ($p < 0.05$), with a higher percentage of regression in the island conditions, which likely reflected the greater structural complexity in the island conditions.

In the post-verb region, there was a marginally significant interaction of verb and structure type ($p = 0.066$), but no significant effect was observed in other eye-movement measures.

Discussion

Experiment 2 used an eye-tracking during reading method to investigate whether the parser uses verb transitivity information in deciding whether to postulate a gap at the verb object position. First fixation durations and first pass times for intransitive verbs were significantly longer in a structure that allows a gap (non-island condition) than when the same verb appeared in an island configuration. This effect was not observed when the critical verb was transitive. The fact that there was a reading disruption for intransitive verbs but not for transitive verbs is consistent with the prediction of the hyper-active gap filling hypothesis. If the parser creates an object gap and integrates the filler into the object position before having access to verb transitivity information, reading disruption in the non-island intransitive condition should result from the mismatch between the predicted transitivity and actual transitivity of the verb.

It is also important to note that in this experiment the critical mismatch effects were observed in the verb region, unlike in Experiment 1 where the mismatch effects were observed only in the spill-over adverb region. This constitutes stronger evidence for hyper-active gap filling, because the mismatch effect must have resulted from properties of the verb itself. The question of why the critical effects were observed in the verb region in Experiment 2 (unlike in Experiment 1, where the effect was found in the spill-over region) likely reflects task-based differences whose effects are seen well beyond the current studies. Inhibition of the button pressing action in self-paced reading tasks is likely more difficult than inhibition of saccades in an eye-tracking task.

TABLE 4 | Summary of model estimates, standard errors, and *t*-values (for linear mixed effects models) and z-values (for logit mixed effects models) for four eye movement measures in Experiment 2.

Measure	Pre-verb region		Verb region		Post-verb region	
	Estimate	*t* (*Z*)	Estimate	*t* (*Z*)	Estimate	*t* (*Z*)
First fixation						
(Intercept)	210 (6)	37.361	242 (6)	38.800	241 (7)	33.526
Verb	−6 (6)	−1.71	2 (8)	0.279	3 (6)	−0.418
Structure	3 (5)	0.559	15 (6)	−2.425*	5 (6)	−0.780
Verb * Structure	−1 (10)	−0.141	24 (13)	−1.887†	6 (12)	−0.529
First-pass time						
(Intercept)	344 (14)	24.386	279 (8)	35.233	288 (11)	25.819
Verb	12 (17)	0.747	8 (12)	0.696	2 (9)	0.186
Structure	97 (13)	7.402**	−18 (9)	−2.076*	−2 (9)	−0.227
Verb * Structure	−0.8 (26)	−0.031	−45 (17)	−2.562*	−12 (18)	−0.655
Regression path time						
(Intercept)	551 (33)	16.495	398 (23)	17.29	458 (26)	17.793
Verb	2 (28)	0.086	12 (30)	0.414	53 (37)	1.420
Structure	162 (27)	5.930**	27 (25)	1.092	17 (30)	0.579
Verb * Structure	−41 (55)	−0.750	−12 (49)	−0.238	−53 (60)	−0.880
Percent regressions						
(Intercept)	−0.96 (0.21)	−4.350	−1.63 (0.20)	−8.192	−1.36 (0.18)	−7.672
Verb	−0.18 (0.17)	−0.919	0.05 (0.25)	0.202	0.20 (0.19)	1.060
Structure	0.18 (0.17)	1.057	0.44 (0.20)	2.182*	0.06 (0.18)	0.344
Verb * Structure	0.29 (0.34)	0.854	0.28 (0.40)	0.693	−0.66 (0.36)	−1.835†

Verb = verb transitivity (transitive vs. intransitive); Structure = island type (non-island vs. island).
†$p < 0.10$; *$p < 0.05$; **$p < 0.001$.

We note that one other methodological difference between our experiments and Staub (2007) regards the types of intransitive verbs used. Our intransitive materials consisted of two types of intransitive verbs: we mainly used unergative verbs which only take a semantic agent as an argument, but we also used unaccusative intransitive verbs that only take a theme/experiencer as an argument (Perlmutter, 1978; Levin and Rappaport Hovav, 1995). On the other hand, Staub's intransitive condition used only unaccusative intransitive verbs. Both types of intransitive verbs are generally incompatible with an overt direct object NP, but in some restricted contexts unergative intransitive verbs are capable of hosting an NP object (e.g., "laugh a big laugh"; see Keyser and Roeper, 1984). It is possible that this special property of unergative verbs may have led the parser to treat it in the same way as transitive verbs in our experiments, whereas unaccusative intransitive verbs admit no such exceptions.

It is important to note that this difference in materials design does not challenge our interpretation of the data. First, our stimuli did not meet the lexical or structural condition for allowing unergative verbs to behave as transitive verbs. Second, if our participants treated the unergative verbs as transitive verbs, then there should have been no reason to observe a slow-down in the intransitive, non-island condition, contrary to the findings in Experiments 1 and 2. However, in order to ascertain that our findings are not restricted to unergative intransitive verbs, we conducted Experiment 3 in which we used only the unaccusative intransitive verbs that were used in Staub (2007).

Experiment 3

The goal of Experiment 3 was to replicate the findings from Experiments 1 and 2 with a different set of intransitive verbs. We constructed new sets of stimuli that used only the unaccusative intransitive verbs used in Staub (2007). Given that unaccusative intransitive verbs are syntactically incapable of hosting an overt direct object NP, this class of intransitive verbs provides a stronger test of the transitivity mismatch effect.

Method

Participants

We recruited 44 native speakers of American English from the University of Maryland community. All had normal or corrected-to-normal vision, and were naïve to the purpose of the experiment. They received course credit or were paid $10 for their participation, which lasted around 40 min.

Materials

We created 24 sets of four sentences. The experimental design in this study is identical to that of Experiment 2 (see **Table 2**), except that the sentences were modified such that the critical verbs in all items were unaccusative intransitive verbs used in Staub (2007). These verbs included *remain, depart, prevail, emerge, arise, die, persist, disappear, erupt, appear, vanish, arrive*. According to Staub (2007), these verbs are considered to disallow transitive frames. Although it may be possible to find some rare counter-examples, we note that this should only

work against the hyper-active gap filling hypothesis, because the possibility of transitive frame would eliminate reasons to observe a reading time slow-down. Thus, finding a robust mismatch effect on the intransitive verb region should eliminate any concerns about the potential transitivity of the intransitive verbs.

The 24 sentence sets were counter-balanced across four lists, such that each participant saw only one version of each of the target sentences. We used 12 intransitive verbs from Staub (2007), such that 2 of the 24 items used the same verb with a different context. Participants saw each intransitive verb twice across the course of the experiment, once in an island context and once in a non-island context. The target sentences were combined with 108 fillers of similar length and complexity.

Procedure

An SR Research (Mississauga, ON, Canada) Eyelink 1000 eye-tracker at the University of Maryland was used to record eye movements. The basic configuration of this eye-tracker as well as the instruction for participants was the same as for Experiment 2, except that the stimuli were displayed on a 17-inch monitor, which allowed a maximum of 100 characters per line. The entire experiment lasted ∼40 min. The experiment protocol for this study was approved by the Institutional Review Board at the University of Maryland.

Data Analysis

The data analysis procedure was the same as that of Experiment 2. The mixed effects models included random intercepts for participants and items. We used the same procedure as Experiment 2 to simplify the random slope structure until the models converged in all regions and eye movement measures. This procedure led us to adopt verb transitivity as a random slope for participants only.

Results

Mean comprehension accuracy for the experimental items was 91.9% across the four conditions, and did not differ across the four conditions. **Table 5** presents the participant means on each measure for each region as well as the standard errors of the participant means, and **Table 6** presents a summary of the statistical analyses.

TABLE 5 | Experiment 3 participant mean reading times in milliseconds (standard error) and percent regressions.

Measure	Pre-verb region	Verb region	Post-verb region
First fixation			
Transitive, non-island	229 (8)	277 (8)	268 (11)
Transitive, island	237 (8)	266 (8)	258 (9)
Intransitive, non-island	226 (7)	299 (10)	271 (9)
Intransitive, island	222 (6)	270 (9)	260 (9)
First-pass time			
Transitive, non-island	367 (22)	319 (12)	330 (21)
Transitive, island	468 (29)	316 (14)	321 (16)
Intransitive, non-island	349 (19)	379 (13)	340 (15)
Intransitive, island	461 (21)	345 (20)	308 (14)
Regression path time			
Transitive, non-island	529 (29)	386 (20)	553 (79)
Transitive, island	706 (47)	520 (44)	529 (45)
Intransitive, non-island	538 (43)	528 (38)	545 (40)
Intransitive, island	762 (48)	527 (54)	497 (43)
Percent regressions			
Transitive, non-island	31.0 (3.7)	11.7 (2.7)	26.4 (3.8)
Transitive, island	26.3 (4.1)	28.4 (3.6)	25.9 (2.9)
Intransitive, non-island	26.7 (3.9)	14.4 (2.3)	24.0 (3.5)
Intransitive, island	32.1 (3.6)	24.0 (3.3)	21.2 (3.1)

Overall, the statistical analysis revealed a similar pattern to the results of Experiment 2. In the pre-verb region, first pass and regression path times showed a main effect of structure type (ps < 0.001), with longer reading times in the island conditions than in the non-island conditions. As explained above, this effect was expected since the pre-verb region in the island conditions contained the extra word *who*. Percent regressions showed a main effect of verb type ($p < 0.05$), with more frequent regressions in the intransitive conditions. Although this was unexpected, the regression frequency in the pre-verb region is unlikely to have affected reading times in subsequent regions.

In the verb region, first fixation durations revealed a main effect of structure type ($p < 0.05$), with longer reading times in the intransitive conditions, but the interaction was not significant. In first pass times, however, a significant interaction of verb and structure type effect was observed ($p < 0.05$). A pairwise comparison revealed that reading times in the non-island, intransitive condition were longer than in the island, intransitive condition ($p < 0.001$), but no significant difference was observed between the transitive conditions.

Because the regression path duration measure reflects differences in the probability of regressing from this region, we discuss the percent regressions results at the verb region first. There was a main effect of structure in percent regressions ($p < 0.05$). The greater percent regression in the island conditions most likely reflects the structural complexity of the island conditions. Next, regression path durations also revealed a main effect of structure ($p < 0.05$), as well as a significant interaction of verb and structure ($p < 0.05$). Pairwise comparisons revealed that the direction of this effect was the opposite of the expected pattern: a significant difference between the transitive conditions ($p < 0.01$), but no difference between the intransitive conditions.

This interaction was unexpected, but it receives a straightforward explanation once we consider the fact that regression path times reflects two different underlying measures: the first pass time and time spent regressing to earlier regions (for discussion see Staub and Clifton, 2006). As described above, transitivity mismatch was associated with longer first pass times and increased regressions in the intransitive non-island condition. However, the presence of an island appeared to have an independent cost as evidenced by the fact that the two island conditions had high percentages of regressions (24.0 and 28.4%), and this is reflected in the large regression path time in these conditions. In other words, the interaction in regression path may have resulted from the combination of complexity slowdowns in the two island

TABLE 6 | Summary of model estimates, standard errors, and *t*-values (for linear mixed effects models) and z-values (for logit mixed effects models) for four eye movement measures in Experiment 3.

Measure	Pre-verb region		Verb region		Post-verb region	
	Estimate	*t* (*Z*)	Estimate	*t* (*Z*)	Estimate	*t* (*Z*)
First fixation(Intercept)	229 (5)	45.584	279 (6)	47.583	264 (8)	35.193
Verb	6 (6)	1.005	8 (8)	0.974	2 (8)	0.295
Structure	3 (6)	0.456	−20 (8)	−2.504*	−10 (7)	−1.436
Verb * Structure	19 (12)	1.561	−21 (16)	−1.327	−8 (14)	−0.593
First-pass time(Intercept)	411 (21)	19.502	341 (10)	35.383	324 (15)	21.715
Verb	−5 (17)	−0.269	17 (12)	1.400	12 (12)	1.015
Structure	103 (17)	6.046***	−23 (12)	−1.896†	−18 (12)	−1.551
Verb * Structure	31 (34)	0.906	−78 (24)	−3.240*	−1 (23)	0.052
Regression path time(Intercept)	663 (44)	14.345	491 (26)	18.914	527 (42)	12.408
Verb	−29 (37)	−0.782	71 (38)	1.869†	28 (50)	0.573
Structure	188 (29)	6.413***	64 (32)	1.978*	−20 (37)	−0.545
Verb * Structure	−68 (59)	−1.167	−129 (65)	−2.002*	−4 (73)	−0.058
Percent regressions(Intercept)	−1.02 (0.15)	−6.935	−1.56 (0.15)	−10.328	−1.25 (0.15)	−8.300
Verb	−0.39 (0.19)	−2.074*	0.28 (0.22)	1.301	0.19 (0.20)	0.998
Structure	0.06 (0.16)	0.374	0.93 (0.20)	4.668***	−0.09 (0.17)	−0.517
Verb * Structure	−0.17 (0.33)	−0.533	−0.04 (0.40)	−0.088	0.33 (0.35)	0.959

Verb = verb transitivity (transitive vs. intransitive); Structure = island type (non-island vs. island).
*†p < 0.10; *p < 0.05; ***p < 0.001.*

conditions and transitivity mismatch slowdown in the intransitive non-island condition, such that all three were slower than the transitive non-island condition.

In the post-verb region, no significant effect was observed in any of the eye-movement measures.

Discussion

Experiment 3 was designed to replicate the results of Experiment 2 with the same intransitive verbs used by Staub (2007). We again observed that first pass times for intransitive verbs in a structure that would allow a gap (non-island condition) were significantly longer than when the same verb appeared within an island configuration. This contrast was not observed when the critical verb was transitive with a plausible direct object. This contrast is consistent with the hyper-active gap filling hypothesis, which states that the parser creates an object gap and integrates the filler into the object position before having access to verb transitivity information. This hypothesis predicts that reading disruption in the non-island intransitive condition should result from the mismatch between the predicted transitivity and actual transitivity of the verb.

We also found that regression path times at the verb region were much shorter for the transitive non-island condition than the other three conditions, a pattern that was also present but unreliable in Experiment 2. As discussed in the results section, this was due to a combination of the higher percentage of regressions in the island conditions and the longer first pass time in the intransitive non-island condition. Although speculative, one possible interpretation of the larger percentage of regressions in the island conditions is that island conditions contain an extra word (i.e., the relative pronoun *who*) and incur greater complexity.

General Discussion

Experiments 1, 2, and 3 all demonstrated evidence for reading disruption at an intransitive verb when the verb was in a potential gap-filling environment. The reading disruption that can be attributed to a transitivity mismatch effect was observed in the same region as the plausibility mismatch effect (Experiment 1), and this reading disruption for an intransitive verb was observed as early as the first fixation on the intransitive verb (Experiments 2 and 3). These results lend support to the hyper-active gap filling hypothesis, which claims that in English filler-gap dependency processing, object gap creation can be initiated based on pre-verbal information and can thereby lead the parser to expect a transitive verb. This is indeed what has been proposed for filler-gap dependency processing mechanism in head-final languages (Aoshima et al., 2004), but the current work suggests that the same mechanism extends to the processing of filler-gap dependency in verb-medial languages like English as well.

The view that object gap creation is triggered by pre-verbal information contrasts with a standard view in English filler-gap dependency processing that object gap creation is driven by properties of the verb (e.g., Pickering and Barry, 1991; McElree et al., 2003). In fact, the hyper-active gap filling mechanism suggests an alternative interpretation of existing evidence for active object gap creation. For example, the plausibility mismatch effect found in Traxler and Pickering (1996) has been taken to suggest that filler-retrieval occurs after accessing the transitivity information on the verb, and that subsequent structural integration of the filler leads to the implausible verb–object composition, which in turn results in reading time slowdown. However, under the hyper-active gap filling account, prior to the verb the reader analyzes

the filler as a direct object of the upcoming verb, and given the combination of the subject NP and the hypothesized object NP, the reader may already expect a certain class of transitive verbs that would be semantically compatible with the filler NP. In other words, plausibility mismatch effects could be reconsidered as a reflection of a violation of lexical expectations, which result from predictive structural analysis. Future studies are needed to examine to what extent this reinterpretation of plausibility mismatch effects is feasible.

The present study has focused on filler-gap dependency processing, but the current conclusion is consistent with a broader class of models of sentence processing that propose that the parser utilizes a variety of sources of linguistic and contextual information to predictively build structural representations (Kimball, 1975; Gibson, 1998; Hale, 2003; Kamide et al., 2003; Staub and Clifton, 2006; Levy, 2008). On the other hand, the present study does not reveal what kind of pre-verbal information is critical for triggering object gap creation in advance of the verb. One possible source that was already discussed in the Introduction is the grammatical knowledge of phrase structure rules, which suggest that the upcoming VP representation can contain an object NP slot. However, it is equally feasible that the parser could use non-grammatical information in predictively positing the object gap, such as differences in the relative conditional probabilities derived from the lexical and contextual information from the combination of the filler NP and the subject. For example, even when a clause appears to resemble a gap structure like a relative clause, with a certain combination an adjunct gap may seem much more plausible than an object gap analysis (e.g., *the day that*. . . can continue as involving an adjunct gap as in *the day that I was born,* or an object gap as in *the day that I have been looking forward to*). Further studies are needed to investigate what kind of information contributes to such predictive object gap creation processes (Chow et al., 2013).

We acknowledge that the data reported in this paper are compatible with an alternative explanation that assumes that verb information plays a critical enabling role in English filler-gap dependency formation. For example, it is possible that filler retrieval processes are automatically activated as soon as the parser accesses the category information of the verb without accessing the transitivity information of the verb. Such a procedure could be motivated by an incremental interpretation strategy that attempts to combine any N-N-V sequence into a proposition (for discussion, see e.g., Goodluck et al., 1991, 1995). Under this alternative account, the transitivity mismatch effect arises because the filler that was 'blindly' retrieved based on the verb category information mismatches the subcategorization property of the verb that is accessed later (see van Gompel and Liversedge, 2003, for a similar proposal for a gender mismatch effect in pronoun processing, and see Kazanina et al., 2007 for an alternative account based on predictive mechanisms).

Although our study does not completely rule out a non-predictive account, these data place important constraints on the form that such an account must take. Critically, a non-predictive account must assume that access to the contents of lexical information is ordered, such that category information is accessed earlier than the subcategorization property of the verb. However, as yet there is little evidence to support such ordered access to category vs. other contents of a verb (Farmer et al., 2006 is one rare case, but see Staub et al., 2009 for a counterargument), whereas there is an abundance of psycholinguistic and neurolinguistic research demonstrating extremely fast access to all aspects of lexical content (e.g., Federmeier et al., 2000; Dambacher et al., 2006; Hauk et al., 2006; Staub and Rayner, 2007; Tanenhaus, 2007; Almeida and Poeppel, 2013; Chow et al., 2014). Moreover, there has been a recent surge of empirical work demonstrating that structure building processes can proceed predictively based on various types of top–down linguistic and contextual information, as discussed above (e.g., Konieczny, 2000; Kamide et al., 2003; DeLong et al., 2005; Van Berkum et al., 2005; Lau et al., 2006; Staub and Clifton, 2006; Levy and Keller, 2013; Yoshida et al., 2013; Yoshida, unpublished doctoral dissertation), including access to transitivity information (Arai and Keller, 2013). The current work demonstrating extremely early object gap creation processes can be seen as another instance of such predictive structure building processes. While these other findings lead us to favor a predictive explanation, further work is needed to more firmly establish that the hyper-active gap filling hypothesis is a better account for the pattern of results observed across a variety of paradigms than this alternative category-driven approach.

The current finding may also seem to contradict findings by Boland et al. (1995) and Pickering and Traxler (2001). These authors tested the processing of filler-gap dependencies in sentences that contain verbs like *persuade* or *remind* that can have both an NP direct object slot and a clausal complement slot in their argument structure, and found no evidence for reading disruption when the filler was semantically incompatible with the direct object NP slot, but compatible with the complement slot. According to the hyper-active gap filling account, encountering a *persuade*-type verb should not result in a transitivity mismatch effect since *persuade* makes available an object position, but one may wonder whether it should result in a plausibility mismatch effect when the filler is a semantically incompatible object, since an object-gap structure is hypothesized to be predictively constructed before the verb.

We can see two ways of reconciling these findings with the results presented here. First, the plausibility mismatch slowdown observed for simple transitive verbs may largely reflect the cost of reanalyzing the predicted structure to one that is compatible with the new input, which may vary depending on the argument structure of the verb. Revision may be costly in the cases where the verb is intransitive or mono-transitive, where the argument structure does not provide sufficient information for the parser to anticipate an alternative structural position for the filler. In the *persuade/remind* cases, on the other hand, the revision may be less costly because the argument structure of the verb clearly indicates the presence of an upcoming clause in which the filler can be integrated. Second, the predicted filler-gap structure may be more abstract than we have indicated so far. Rather than specifically predicting an object gap when the filler and relative clause subject are encountered, the parser may simply predict an argument gap position somewhere inside the complement domain of an upcoming VP representation, such that a gap in either the NP slot

or in the clausal complement slot of *persuade*-class ditransitive verbs would be consistent with the prediction. The current results are compatible with either account. In sum, under either account, reading disruption at the verb can be mitigated when the verb makes more than one argument position available. This might explain why having an adjunct PP continuation (e.g., about) for mono-transitive verbs (e.g., wrote) still causes reading disruption at the verb while ditransitive verbs like *persuade/remind* do not lead to such reading disruption.

In the sentences used here, the intransitive structures are eventually resolved by the appearance of a preposition, which provides another structural position for the filler. Although this could be recognized as a possible reanalysis even at the verb position, this adjunct position is not specifically licensed by the input until the preposition is actually encountered (in contrast with the *persuade/remind* cases, in which the object position is available at the verb due to its argument structure information). One interesting question for further research is whether the difficulty of recovering from the simple transitive analysis is modulated by the frequency with which a particular intransitive verb co-occurs with a prepositional phrase that could host the filler. For example, many intransitive verbs can be combined with a prepositional complement to form a phrasal verb that takes the object of the preposition as an argument, e.g., *listen to the music*. If a particular intransitive verb occurs very frequently in a phrasal verb configuration, reanalysis to this structure in a filler-gap configuration might be less costly, even prior to the presentation of the preposition.

Finally, the conclusion that the same filler-gap dependency completion procedure is used across head-initial and head-final languages suggests that the parser's structure building procedures, at least for filler-gap dependency completion, may not be qualitatively different across languages. However, future studies extending beyond Japanese and English are needed to test the robustness of this generalization. Moreover, predictive dependency formation processes are observed in domains other than filler-gap dependency processing (e.g., resolution of backward anaphora; Kazanina et al., 2007; Aoshima et al., 2009; Yoshida et al., 2013), but it is not yet known whether these other predictive structure building processes are also relatively constant across languages. Overall, we believe this line of cross-linguistic investigation has the potential to shed further light on fundamental questions about the relationship between linguistic representations and real-time processes for constructing those representations.

Conclusion

The present study tested the hypothesis that predictive structure building processes underlie filler-gap dependency completion in English. In the presence of a filler-gap dependency, intransitive verbs consistently led to reading disruption, and this pattern was replicated in self-paced reading measures as well as in eye movement measures. These findings show that English speakers do not wait to check that the verb makes an object position available, and are consistent with the hypothesis that the parser postulates an object gap at least as soon as it encounters a filler phrase and a subject NP. We suggest that the parser uses pre-verbal information to predictively create rich syntactic representations regardless of word order differences across languages.

Acknowledgments

This work was supported in part by a UMD Wylie Dissertation Fellowship to AO, NSF BCS-1423117 to AO, NIH F32-HD063221-02 to EL, NSF BCS-0954651 to CP and AO, and NSF BCS-0848554 to CP.

Supplementary Material

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fpsyg.2015.00384/abstract

References

Abney, S. P. (1989). A computational model of human parsing. *J. Psycholinguist. Res.* 18, 129–144. doi: 10.1007/BF01069051

Abney, S. P., and Johnson, M. (1991). Memory requirements and local ambiguities of parsing strategies. *J. Psycholinguist. Res.* 20, 233–250. doi: 10.1007/BF01067217

Almeida, D., and Poeppel, D. (2013). Word-specific repetition effects revealed by MEG and the implications for lexical access. *Brain Lang.* 127, 497–509. doi: 10.1016/j.bandl.2013.09.013

Aoshima, S., Phillips, C., and Weinberg, A. (2004). Processing filler-gap dependencies in a head-final language. *J. Mem. Lang.* 51, 23–54. doi: 10.1016/j.jml.2004.03.001

Aoshima, S., Yoshida, M., and Phillips, C. (2009). Incremental processing of co-reference and binding in Japanese. *Syntax* 12, 93–134. doi: 10.1111/j.1467-9612.2009.00123.x

Arai, M., and Keller, F. (2013). The use of verb-specific information for prediction in sentence processing. *Lang. Cogn. Process.* 28, 525–560. doi: 10.1080/01690965.2012.658072

Baayen, H., Davidson, J., and Bates, M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *J. Mem. Lang.* 59, 390–412. doi: 10.1016/j.jml.2007.12.005

Bader, M., and Lasser, I. (1994). "German verb-final clauses and sentence processing: evidence for immediate attachment," in *Perspectives on Sentence Processing*, eds C. Clifton, L. Frazier, and K. Rayner (Hillsdale, NJ: Lawrence Erlbaum), 225–242.

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: keep it maximal. *J. Mem. Lang.* 68, 255–278. doi: 10.1016/j.jml.2012.11.001

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2014). *Lme4: Linear Mixed-Effects Models Using Eigen and S4*. R Package Version 1.1–7. Available at: http://CRAN.R-project.org/package=lme4

Blodgett, A., and Boland, J. E. (2004). Differences in the timing of implausibility detection for recipient and instrument prepositional phrases. *J. Psycholinguist. Res.* 33, 1–24. doi: 10.1023/B:JOPR.0000010512.39960.27

Boland, J. E., Tanenhaus, M. K., and Garnsey, S. M. (1990). Evidence for the immediate use of verb control information in sentence processing. *J. Mem. Lang.* 29, 413–432. doi: 10.1016/0749-596X(90)90064-7

Boland, J. E., Tanenhaus, M. K., Garnsey, S. M., and Carlson, G. N. (1995). Verb argument structure in parsing and interpretation: evidence from wh-questions. *J. Mem. Lang.* 34, 774–806. doi: 10.1006/jmla.1995.1034

Bornkessel, I., Schlesewsky, M., and Friederici, A. D. (2002). Grammar overrides frequency: evidence from the online processing of flexible word order. *Cognition* 85, B21–B30. doi: 10.1016/S0010-0277(02)00076-8

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Chow, W. Y., Lago, S., Barrios, S., Parker, D., Morini, G., and Lau, E. (2014). Additive effects of repetition and predictability during comprehension: evidence from event-related potentials. *PLoS ONE* 9:e99199. doi: 10.1371/journal.pone.0099199

Chow, W. Y., Wang, S., Lau, E., and Phillips, C. (2013). "Predictive computations underlie the N400's sensitivity to thematic role-reversals," in *Poster Presented at the 26th Annual CUNY Human Sentence Processing Conference*, Columbia, SC.

Clahsen, H., and Featherston, S. (1999). Antecedent priming at trace positions: evidence from German scrambling. *J. Psycholinguist. Res.* 28, 415–437. doi: 10.1023/A:1023293132656

Crain, S., and Fodor, J. D. (1985). "How can grammars help parsers?" in *Natural Language Parsing: Psychological, Computational and Theoretical Perspectives*, eds D. Dowty, L. Kartunnen, and A. Zwicky (Cambridge: Cambridge University Press).

Crocker, M. W. (1996). *Computational Psycholinguistics: An Interdisciplinary Approach to the Study of Language*. Dordrecht: Springer. doi: 10.1007/978-94-009-1600-5

Dambacher, M., Kliegl, R., Hofmann, M., and Jacobs, A. M. (2006). Frequency and predictability effects on event-related potentials during reading. *Brain Res.* 1084, 89–103. doi: 10.1016/j.brainres.2006.02.010

Deane, P. D. (1991). Limits to attention: a cognitive theory of island phenomena. *Cogn. Ling.* 2, 1–63. doi: 10.1515/cogl.1991.2.1.1

DeLong, K. A., Urbach, T. P., and Kutas, M. (2005). Probabilistic word preactivation during language comprehension inferred from electrical brain activity. *Nat. Neurosci.* 8, 1117–1121. doi: 10.1038/nn1504

de Vincenzi, M. (1991). *Syntactic Parsing Strategies in Italian*. Dordrecht: Kluwer. doi: 10.1007/978-94-011-3184-1

Farmer, T. A., Christiansen, M. H., and Monaghan, P. (2006). Phonological typicality influences on-line sentence comprehension. *Proc. Natl. Acad. Sci. U.S.A.* 103, 12203–12208. doi: 10.1073/pnas.0602173103

Featherston, S., Gross, M., Münte, T. F., and Clahsen, H. (2000). Brain potentials in the processing of complex sentences: an ERP study of control and raising constructions. *J. Psycholinguist. Res.* 29, 141–154. doi: 10.1023/A:1005188810604

Federmeier, K. D., Segal, J. B., Lombrozo, T., and Kutas, M. (2000). Brain responses to nouns, verbs and class-ambiguous words in context. *Brain* 123, 2552–2566. doi: 10.1093/brain/123.12.2552

Felser, C., Clahsen, H., and Münte, T. F. (2003). Storage and integration in the processing of filler-gap dependencies: an ERP study of topicalization and wh-movement in German. *Brain Lang.* 87, 345–354. doi: 10.1016/S0093-934X(03)00135-4

Fodor, J. D. (1978). Parsing strategies and constraints on transformations. *Linguist. Inquiry* 9, 427–473.

Ford, M., Bresnan, J., and Kaplan, R. (1982). "A competence-based theory of syntactic closure," in *The Mental Representation of Grammatical Relations*, ed. J. Bresnan (Cambridge, MA: MIT Press), 727–796.

Frazier, L. (1987). Syntactic processing: evidence from Dutch. *Nat. Lang. Linguist. Theor.* 5, 519–559. doi: 10.1007/BF00138988

Frazier, L., and Clifton, C. (1989). Successive cyclicity in the grammar and the parser. *Lang. Cogn. Process.* 4, 93–126. doi: 10.1080/01690968908406359

Frazier, L., and Flores d'Arcais, G. B. (1989). Filler driven parsing: a study of gap filling in Dutch. *J. Mem. Lang.* 28, 331–344. doi: 10.1016/0749-596X(89)90037-5

Frazier, L., and Rayner, K. (1982). Making and correcting errors during sentence comprehension: eye movements in the analysis of structurally ambiguous sentences. *Cogn. Psychol.* 14, 178–210. doi: 10.1016/0010-0285(82)90008-1

Garnsey, S. M., Pearlmutter, N. J., Myers, E., and Lotocky, M. A. (1997). The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *J. Mem. Lang.* 37, 58–93. doi: 10.1006/jmla.1997.2512

Garnsey, S. M., Tanenhaus, M. K., and Chapman, R. M. (1989). Evoked potentials and the study of sentence comprehension. *J. Psycholinguist. Res.* 18, 51–60. doi: 10.1007/BF01069046

Gibson, E. (1998). Linguistic complexity: locality of syntactic dependencies. *Cognition* 68, 1–76. doi: 10.1016/S0010-0277(98)00034-1

Gibson, E. (2000). "The dependency locality theory: a distance-based theory of linguistic complexity," in *Image, Language, Brain: Papers From the First Mind Articulation Project Symposium*, eds A. Marantz, Y. Miyashita, and W. O'Neil (Cambridge, MA: MIT Press), 95–126.

Gibson, E., and Warren, T. (2004). Reading-time evidence for intermediate linguistic structure in long-distance dependencies. *Syntax* 7, 55–78. doi: 10.1111/j.1368-0005.2004.00065.x

Goodluck, H., Finney, M., and Ling, M. (1995). Gap accessibility and children's sentence processing. *Cah. Linguist. Ottawa* 23(Suppl.), 1–32.

Goodluck, H., Finney, M., and Sedivy, J. (1991). "Sentence completeness and filler-gap dependency parsing," in *OTS Yearbook 1991*, eds P. Coopmans, B. Schouten and W. Zonneveld (Utrecht: University of Utrecht), 19–32.

Gordon, P. C., Hendrick, R., and Levine, W. H. (2002). Memory-Load interference in syntactic processing. *Psychol. Sci.* 13, 425–430. doi: 10.1111/1467-9280.00475

Gouvea, A. C., Phillips, C., Kazanina, N., and Poeppel, D. (2010). The linguistic processes underlying the P600. *Lang. Cogn. Process.* 25, 1–40. doi: 10.1080/01690960902965951

Grimshaw, J. B. (1990). *Argument Structure*. Cambridge, MA: MIT Press.

Haarmann, H. J., and Cameron, K. A. (2005). Active maintenance of sentence meaning in working memory: evidence from EEG coherences. *Int. J. Psychophysiol.* 57, 115–128. doi: 10.1016/j.ijpsycho.2005.03.017

Hale, J. (2003). The information conveyed by words in sentences. *J. Psycholinguist. Res.* 32, 101–123. doi: 10.1023/A:1022492123056

Hauk, O., Davis, M. H., Ford, M., Pulvermüller, F., and Marslen-Wilson, W. D. (2006). The time course of visual word recognition as revealed by linear regression analysis of ERP data. *Neuroimage* 30, 1383–1400. doi: 10.1016/j.neuroimage.2005.11.048

Hofmeister, P., and Sag, I. A. (2010). Cognitive constraints and island effects. *Language* 86, 366–415. doi: 10.1353/lan.0.0223

Inoue, A., and Fodor, J. D. (1995). "Information-paced parsing of Japanese," in *Japanese Sentence Processing*, eds R. Mazuka and N. Nagai (Hillsdale, NJ: Lawrence Erlbaum), 9–63.

Jaeger, T. F. (2008). Categorical data analysis: away from ANOVAs (transformation or not) and towards logit mixed models. *J. Mem. Lang.* 59, 434–446. doi: 10.1016/j.jml.2007.11.007

Jaeger, T. F. (2011). *Random Slopes in LME*. Available at: https://mailman.ucsd.edu/pipermail/ling-r-lang-l/2011-February/000225.html

Just, M. A., Carpenter, P., and Wooley, J. D. (1982). Paradigms and processes in reading comprehension. *J. Exp. Psychol. Gen.* 111, 228–238. doi: 10.1037/0096-3445.111.2.228

Kaan, E., Harris, A., Gibson, E., and Holcomb, P. (2000). The P600 as an index of syntactic integration difficulty. *Lang. Cogn. Process.* 15, 159–201. doi: 10.1080/016909600386084

Kamide, Y., Altmann, G., and Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: evidence from anticipatory eye movements. *J. Mem. Lang.* 49, 133–156. doi: 10.1016/S0749-596X(03)00023-8

Kamide, Y., and Mitchell, D. C. (1999). Incremental pre-head attachment in Japanese parsing. *Lang. Cogn. Process.* 14, 631–662. doi: 10.1080/016909699386211

Kazanina, N., Lau, E. F., Lieberman, M., Yoshida, M., and Phillips, C. (2007). The effect of syntactic constraints on the processing of backwards anaphora. *J. Mem. Lang.* 56, 384–409. doi: 10.1016/j.jml.2006.09.003

Keyser, S. J., and Roeper, T. (1984). On the middle and ergative constructions in English. *Linguist. Inquiry* 15, 381–416.

Kimball, J. (1975). Predictive analysis and over-the-top parsing. *Syntax Semant.* 4, 155–179.

King, J., and Just, M. A. (1991). Individual differences in syntactic processing: the role of working memory. *J. Mem. Lang.* 30, 580–602. doi: 10.1016/0749-596X(91)90027-H

Kluender, R. (1998). "On the distinction between strong and weak islands: a processing perspective," in *Syntax and Semantics: The Limits of Syntax*, eds P. Culicover and L. McNally (New York: Academic Press), 241–280.

Kluender, R. (2004). "Are subject islands subject to a processing account," in *Proceedings of the 23rd West Coast Conference on Formal Linguistics (WCCFL 23)*, eds V. Chand, A. Kelleher, A. Rodriguez, and B. Schmeiser (Somerville, MA: Cascadilla Press), 475–499.

Kluender, R., and Kutas, M. (1993). Subjacency as a processing phenomenon. *Lang. Cogn. Process.* 8, 573–633. doi: 10.1080/01690969308407588

Koh, S. (1997). The resolution of the dative NP ambiguity in korean. *J. Psycholinguist. Res.* 26, 265–273. doi: 10.1023/A:1025021918198

Konieczny, L. (2000). Locality and parsing complexity. *J. Psycholinguist. Res.* 29, 627–645. doi: 10.1023/A:1026528912821

Kutas, M., and Federmeier, K. D. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends Cogn. Sci.* 4, 463–470. doi: 10.1016/S1364-6613(00)01560-6

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (2014). *lmerTest: Tests for Random and Fixed Effects for Linear Mixed Effect Models*. R Package Version 2.0–11. Available at: http://CRAN.R-project.org/package = lmerTest

Lau, E., Stroud, C., Plesch, S., and Phillips, C. (2006). The role of structural prediction in rapid syntactic analysis. *Brain Lang.* 98, 74–88. doi: 10.1016/j.bandl.2006.02.003

Lee, M.-W. (2004). Another look at the role of empty categories in sentence processing (and grammar). *J. Psycholinguist. Res.* 33, 51–73. doi: 10.1023/B:JOPR.0000010514.50468.30

Levin, B., and Rappaport Hovav, M. (1995). *Unaccusativity: At the Syntax-Lexical Semantics Interface*. Cambridge, MA: MIT Press.

Levy, R. (2008). Expectation-Based syntactic comprehension. *Cognition* 106, 1126–1177. doi: 10.1016/j.cognition.2007.05.006

Levy, R. P., and Keller, F. (2013). Expectation and locality effects in German verb-final structures. *J. Mem. Lang.* 68, 199–222. doi: 10.1016/j.jml.2012.02.005

Lewis, R. L., and Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cogn. Sci.* 29, 375–419. doi: 10.1207/s15516709cog0000_25

MacDonald, M. C., Pearlmutter, N. J., and Seidenberg, M. S. (1994). Lexical nature of syntactic ambiguity resolution. *Psychol. Rev.* 101, 676–703. doi: 10.1037/0033-295X.101.4.676

Mauner, G., and Koenig, J. P. (2000). Linguistic vs. Conceptual sources of implicit agents in sentence comprehension. *J. Mem. Lang.* 43, 110–134. doi: 10.1006/jmla.1999.2703

Mazuka, R., and Lust, B. (1990). "On parameter setting and parsing: predictions for cross-linguistic differences in adult and child processing," in *Language Processing and Language Acquisition,* eds L. Frazier and J. de Villiers (Dordrecht: Kluwer), 163–206.

McElree, B., Foraker, S., and Dyer, L. (2003). Memory structures that subserve sentence comprehension. *J. Mem. Lang.* 48, 67–91. doi: 10.1016/S0749-596X(02)00515-6

McElree, B., and Griffith, T. (1998). Structural and lexical constraints on filling gaps during sentence comprehension: a time-course analysis. *Learn. Mem.* 24, 432–460. doi: 10.1037//0278-7393.24.2.432

McKinnon, R., and Osterhout, L. (1996). Constraints on movement phenomena in sentence processing: evidence from event-related brain potentials. *Lang. Cogn. Process.* 11, 495–524. doi: 10.1080/016909696387132

Nakano, Y., Felser, C., and Clahsen, H. (2002). Antecedent priming at trace positions in Japanese long-distance scrambling. *J. Psycholinguist. Res.* 31, 531–571. doi: 10.1023/A:1021260920232

Nicol, J. L. (1993). "Reconsidering reactivation," in *Cognitive Models of Speech Processing: The Second Sperlonga Meeting,* eds G. Altmann and R. Schillcock (Hillsdale, NJ: Lawrence Erlbaum Associates), 321–347.

Nicol, J., and Swinney, D. (1989). The role of structure in coreference assignment during sentence comprehension. *J. Psycholinguist. Res.* 18, 5–19. doi: 10.1007/BF01069043

Omaki, A., Davidson White, I., Goro, T., Lidz, J., and Phillips, C. (2014). No fear of commitment: children's incremental interpretation in English and Japanese wh-questions. *Lang. Learn. Dev.* 10, 206–233. doi: 10.1080/15475441.2013.844048

Omaki, A., and Schulz, B. (2011). Filler-gap dependencies and island constraints in second language sentence processing. *Stud. Second Lang. Acquis.* 33, 563–588. doi: 10.1017/S0272263111000313

Perlmutter, D. M. (1978). "Impersonal passives and the Unaccusative Hypothesis," in *Proceedings of the 4th Annual Meeting of the Berkeley Linguistics Society* (Berkeley, CA: University of California), 157–189.

Phillips, C. (2006). The real-time status of island phenomena. *Language* 82, 795–823. doi: 10.1353/lan.2006.0217

Phillips, C., Kazanina, N., and Abada, S. H. (2005). ERP effects of the processing of syntactic long-distance dependencies. *Cogn. Brain Res.* 22, 407–428. doi: 10.1016/j.cogbrainres.2004.09.012

Pickering, M., and Barry, G. (1991). Sentence processing without empty categories. *Lang. Cogn. Process.* 6, 229–259. doi: 10.1080/01690969108406944

Pickering, M. J., and Traxler, M. J. (2001). Strategies for processing unbounded dependencies: lexical information and verb-argument assignment. *J. Exp. Psychol. Learn. Mem. Cogn.* 27, 1401–1410. doi: 10.1037/0278-7393.27.6.1401

Pickering, M. J., and Traxler, M. J. (2003). Evidence against the use of subcategorisation frequency in the processing of unbounded dependencies. *Lang. Cogn. Process.* 18, 469–503. doi: 10.1080/01690960344000017

Pollard, C. J., and Sag, I. A. (1994). *Head-Driven Phrase Structure Grammar*. Chicago, IL: University of Chicago Press.

Pritchett, B. L. (1992). *Grammatical Competence and Parsing Performance*. Chicago, IL: University of Chicago Press.

R Development Core Team. (2011). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available at: http://www.R-project.org/

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychol. Bull.* 124, 372–422. doi: 10.1037/0033-2909.124.3.372

Rayner, K., and Pollatsek, A. (1989). *The Psychology of Reading*. Englewood Cliffs, NJ: Prentice Hall.

Rayner, K., and Pollatsek, A. (2006). "Eye-Movement control in reading," in *Handbook of Psycholinguistics*, 2nd Edn, eds M. J. Traxler and M. A. Gernsbacher (New York, NY: Academic Press), 613–657. doi: 10.1016/B978-012369374-7/50017-1

Reichle, E. D., Rayner, K., and Pollatsek, A. (2003). The E-Z reader model of eye-movement control in reading: comparisons to other models. *Behav. Brain Sci.* 26, 445–476. doi: 10.1017/S0140525X03000104

Resnik, P. (1992). "Left-Corner parsing and psychological plausibility," in *Proceedings of the 14th Conference on Computational Linguistics*, Stroudsburg, PA. doi: 10.3115/992066.992098

Shieber, S., and Johnson, M. (1993). Variations on incremental interpretation. *J. Psycholinguist. Res.* 22, 287–318. doi: 10.1007/BF01067835

Snedeker, J., and Trueswell, J. C. (2004). The developing constraints on parsing decisions: the role of lexical-biases and referential scenes in child and adult sentence processing. *Cogn. Psychol.* 49, 238–299. doi: 10.1016/j.cogpsych.2004.03.001

Spivey-Knowlton, M., and Sedivy, J. C. (1995). Resolving attachment ambiguities with multiple constraints. *Cognition* 55, 227–267. doi: 10.1016/0010-0277(94)00647-4

Sprouse, J., Wagers, M., and Phillips, C. (2012). A test of the relation between working-memory capacity and syntactic island effects. *Language* 88, 82–123. doi: 10.1353/lan.2012.0004

Stabler, E. P. (1994). "The finite connectivity of linguistic structure," in *Perspectives on Sentence Processing*, eds L. Frazier, C. Clifton, and K. Rayner (Hillsdale, NJ: Lawrence Erlbaum Associates), 303–336.

Stanovich, K. E., and West, R. F. (1983). On priming by a sentence context. *J. Exp. Psychol. Gen.* 112, 1–36. doi: 10.1037/0096-3445.112.1.1

Staub, A. (2007). The parser doesn't ignore intransitivity, after all. *J. Exp. Psychol. Learn. Mem. Cogn.* 33, 550–569. doi: 10.1037/0278-7393.33.3.550

Staub, A. (2010). Eye movements and processing difficulty in object relative clauses. *Cognition* 116, 71–86. doi: 10.1016/j.cognition.2010.04.002

Staub, A., and Clifton, C. (2006). Syntactic prediction in language comprehension: evidence from either.*Or. J. Exp. Psychol. Learn. Mem. Cogn.* 32, 425–436. doi: 10.1037/0278-7393.32.2.425

Staub, A., Grant, M., Clifton, C., and Rayner, K. (2009). Phonological typicality does not influence fixation durations in normal reading. *J. Exp. Psychol. Learn. Mem. Cogn.* 35, 806–814. doi: 10.1037/a0015123

Staub, A., and Rayner, K. (2007). "Eye movements and on-line comprehension processes," in *The Oxford Handbook of Psycholinguistics,* ed. G. Gaskell (Oxford: Oxford University Press), 327–342.

Stowe, L. (1986). Evidence for on-line gap location. *Lang. Cogn. Process.* 1, 227–245. doi: 10.1080/01690968608407062

Sturt, P. (2003). The time-course of the application of binding constraints in reference resolution. *J. Mem. Lang.* 48, 542–562. doi: 10.1016/S0749-596X(02)00536-3

Sussman, R. S., and Sedivy, J. C. (2003). The time-course of processing syntactic dependencies: evidence from eye movements. *Lang. Cogn. Process.* 18, 143–163. doi: 10.1080/01690960143000498

Szabolcsi, A., and den Dikken, M. (2003). "Islands," in *The Second Glot International State-of-the Article Book,* eds L. Cheng and R. Sybesma (Berlin: Mouton de Gruyter), 213–240.

Tanenhaus, M. K. (2007). "Spoken language comprehension: insights from eye movements," in *The Oxford Handbook of Psycholinguistics,* ed. G. Gaskell (Oxford: Oxford University Press), 309–326.

Tanenhaus, M. K., and Carlson, G. N. (1989). "Lexical structure and language comprehension," in *Lexical Representation and Process*, ed. W. D. Marslen-Wilson (Cambridge, MA: MIT Press), 529–561.

Traxler, M. J., and Pickering, M. J. (1996). Plausibility and the processing of unbounded dependencies: an eye-tracking study. *J. Mem. Lang.* 35, 454–475. doi: 10.1006/jmla.1996.0025

Traxler, M. J., Pickering, M. J., and McElree, B. (2002). Coercion in sentence processing: evidence from eye-movements and self-paced reading. *J. Mem. Lang.* 47, 530–547. doi: 10.1016/S0749-596X(02)00021-9

Van Berkum, J. J., Brown, C. M., Zwitserlood, P., Kooijman, V., and Hagoort, P. (2005). Anticipating upcoming words in discourse: evidence from erps and reading times. *J. Exp. Psychol. Learn. Mem. Cogn.* 31, 443–467. doi: 10.1037/0278-7393.31.3.443

van Gompel, R. P., and Liversedge, S. P. (2003). The influence of morphological information on cataphoric pronoun assignment. *J. Exp. Psychol. Learn. Mem. Cogn.* 29, 128–139. doi: 10.1037/0278-7393.29.1.128

Van Petten, C., and Kutas, M. (1990). Interactions between sentence context and word frequency in event-related brain potentials. *Mem. Cogn.* 18, 380–393. doi: 10.3758/BF03197127

Wagers, M. W., and Phillips, C. (2009). Multiple dependencies and the role of the grammar in real-time comprehension. *J. Linguist.* 45, 395–433. doi: 10.1017/S0022226709005726

Wagers, M. W., and Phillips, C. (2014). Going the distance: memory and control processes in active dependency construction. *Q. J. Exp. Psychol.* 67, 1274–1304. doi: 10.1080/17470218.2013.858363

Yoshida, M., Dickey, M. W., and Sturt, P. (2013). Predictive processing of syntactic structure: sluicing and ellipsis in real-time sentence processing. *Lang. Cogn. Process. 28,* 272–302. doi: 10.1080/01690965.2011.622905

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2015 Omaki, Lau, Davidson White, Dakan, Apple and Phillips. This is an open-access article distributed under the terms of the **Creative Commons Attribution License (CC BY).** *The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.*

frontiers in Psychology

ORIGINAL RESEARCH
published: 11 September 2015
doi: 10.3389/fpsyg.2015.01255

Thematic orders and the comprehension of subject-extracted relative clauses in Mandarin Chinese

Chien-Jer Charles Lin *

Department of East Asian Languages and Cultures, Indiana University, Bloomington, IN, USA

This study investigates the comprehension of three kinds of subject-extracted relative clauses (SRs) in Mandarin Chinese: standard SRs, relative clauses involving the disposal *ba* construction ("disposal SRs"), and relative clauses involving the long passive *bei* constructions ("passive SRs"). In a self-paced reading experiment, the regions before the relativizer (where the sentential fragments are temporarily ambiguous) showed reading patterns consistent with expectation-based incremental processing: standard SRs, with the highest constructional frequency and the least complex syntactic structure, were processed faster than the other two variants. However, in the regions after the relativizer and the head noun where the existence of a relative clause is unambiguously indicated, a top-down global effect of thematic ordering was observed: passive SRs, whose thematic role order conforms to the canonical thematic order of Chinese, were read faster than both the standard SRs and the disposal SRs. Taken together, these results suggest that two expectation-based processing factors are involved in the comprehension of Chinese relative clauses, including both the structural probabilities of pre-relativizer constituents and the overall surface thematic orders in the relative clauses.

Keywords: sentence comprehension, thematic orders, relative clauses, expectations, Mandarin Chinese

OPEN ACCESS

Edited by:
Matthew Wagers, University of California, Santa Cruz, USA

Reviewed by:
Ming Xiang, University of Chicago, USA
Lauren Clemens, McGill University, Canada

***Correspondence:**
Chien-Jer Charles Lin, Department of East Asian Languages and Cultures, Indiana University, 355 North Jordan Avenue, Bloomington, IN 47405-1105, USA
chiclin@indiana.edu

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: *10 April 2015*
Accepted: *05 August 2015*
Published: *11 September 2015*

Citation:
Lin C-JC (2015) Thematic orders and the comprehension of subject-extracted relative clauses in Mandarin Chinese. Front. Psychol. 6:1255. doi: 10.3389/fpsyg.2015.01255

Introduction

Relative clauses have been of great theoretical interest to sentence processing researchers, with decades of research comparing the processing of subject-extracted relative clauses (henceforth "SRs") to that of object-extracted relative clauses (henceforth "ORs"). A robust asymmetry has been repeatedly reported in languages where the relative clauses follow the nouns they modify (i.e., languages with head-initial relative clauses). In English, for example, relative clauses involving subject extractions like (1) have been found to be easier to comprehend than those involving object extractions like (2) (Ford, 1983; King and Just, 1991; King and Kutas, 1995; Gibson et al., 2005; Traxler et al., 2005). The head noun phrases in these constructions [indicated with boldface in (1, 2)] are conventionally referred to as the *fillers* in the sense that they fill the *gaps* located at the extracted positions in the subordinate clauses [indicated with underscores in (1, 2)].

(1) Subject-extracted relative clause:
{**The composer**$_i$ who $___i$ adored the musician} drank a glass of wine.
(2) Object-extracted relative clause:
{**The musician**$_i$ who the composer adored $___i$} drank a glass of wine.

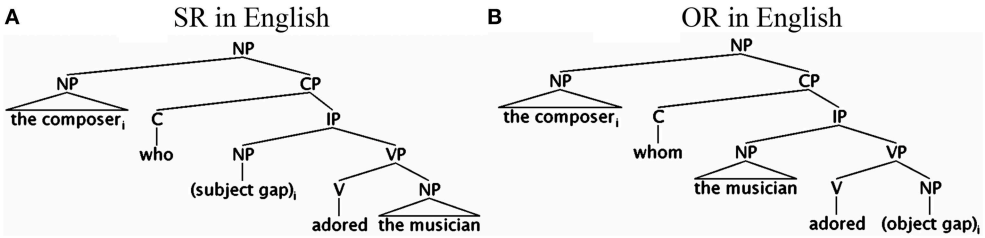


FIGURE 1 | Syntactic structure of relative clauses in English.

Two main groups of theories have been adopted to account for this processing asymmetry, here referred to as "integration-based theories" and "experience-based theories." The first group of theories focuses on the consumption of working memory in constructing filler-gap dependencies, suggesting that SRs in English are easier to comprehend (with shorter reading times and greater comprehension accuracies) because, relative to ORs, less working memory is required to process them. Within these integration-based theories, a number of proposals have been made as to precisely how the relevant processing costs are computed. A linearity account (e.g., Gibson, 1998) focuses on the number of referents intervening between the filler and the gap, attributing the easier comprehension of SRs to fewer new referents intervening between the filler and the gap. As a filler is assumed to remain active until a gap is reached in constructing filler-gap dependencies, the longer filler-gap distance in an OR consumes greater processing costs. A relevant variant of the linearity account focuses on the types of noun phrases intervening the filler and the gap, according to which similar types of noun phrases (NPs) create greater interference and therefore induce higher processing costs (Gordon et al., 2001). The activation and cue-based retrieval theory (Van Dyke and Lewis, 2003; Lewis and Vasishth, 2005) on the other hand takes into consideration the lexical items intervening a filler and a gap as contributors to activation and retrieval.

Within the integration-based theories, a structural account (e.g., O'Grady, 1997; Miyamoto and Nakamura, 2003; Hawkins, 2004; Lin, 2006) relies on the structural distance between the filler and the gap (e.g., computed by counting the number of intervening XPs) to compute processing costs. On this account, processing costs are determined by the number of intervening structural nodes inside a filler-gap dependency. Thus, an SR is easier to comprehend than an OR in English because a subject gap is structurally higher and closer to the relative clause operator (i.e., the complementizer *who/whom*) than an object gap (see **Figure 1**). Since fewer structural nodes intervene between the operator and the subject gap, less working memory is consumed in connecting the filler with the subject gap[1].

The second group of theories is experience-based, formalized either as constraints (e.g., MacDonald and Christiansen, 2002; Raeli and Christiansen, 2007; Gennari and MacDonald, 2008) or through a construct of "expectation" (Surprisal: Hale, 2001; Levy, 2008; Entropy Reduction: Hale, 2006). These theories account for the processing differences by resorting to probabilistic information associated with one's linguistic experiences, attributing the easier processing of SRs to the greater structural predictability associated with SRs than ORs. Since SRs have a higher frequency of occurrence than ORs in English (Roland et al., 2007), the parser is more likely to parse the head noun and relativizer in English as starting an SR than an OR. Thus, the increased predictability associated with SRs is claimed to be what induces the shorter reading times.

A related experience-based theory posits that the dominant (i.e., most frequent) thematic order in a language can be used as a perceptual strategy to facilitate sentence comprehension (Bever, 1970; Townsend and Bever, 2001; Lin, 2013). According to the thematic-order account, experience with thematic orders form *canonical thematic templates*, which may facilitate efficient thematic interpretations. Since the canonical word order in English is SVO and the canonical thematic order is agent-verb-patient, SRs, which present orders consistent with the dominant order, are predicted to be less costly to process[2]. The thematic order account predicts increases in reading time where word order mismatches take place.

This brief summary highlights the fact that the overall advantageous reading of SRs in English is consistent with multiple theories of sentence comprehension though specific predictions about where the processing differences should be observed may differ. Gibson and Wu (2013), for example, point out that an integration account predicts the increase in processing load where the filler-gap integration takes place (i.e., around the embedded verb region in head-initial relative clauses). An experience-based account, on the other hand, predicts the increase in processing load should occur where processing uncertainty increases (i.e., around the embedded subject but not on the embedded verb in an OR).

[1]The classification of these structure-based accounts as working memory accounts is my own. These original structure-based accounts did not necessarily specify the working memory component.

[2]The thematic template theory is of a similar flavor to the NVN strategy of Bever (1970). Whereas, Bever's NVN strategy focuses on the order of syntactic categories in a sentence, the thematic template account focuses on the linear orders of the semantic roles associated with noun phrases in relation to the verb position. What distinguishes the thematic order account from the word order account, therefore, is that the former does not link the semantic arguments to grammatical functions and therefore does not depend on the "structural" positions of the arguments.

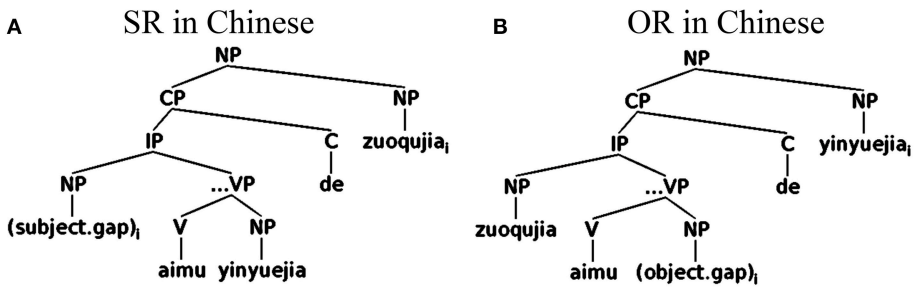


FIGURE 2 | Syntactic structure of relative clauses in Chinese.

An accumulating body of research over the past decade has painted a somewhat different picture of the processing of head-final relative clauses (Basque: Carreiras et al., 2010; Japanese: Miyamoto and Nakamura, 2003; Ueno and Garnsey, 2008; Korean: Kwon et al., 2010; Mandarin Chinese: Hsiao and Gibson, 2003; Lin and Bever, 2006; Lin and Garnsey, 2011; Packard et al., 2011; Qiao et al., 2012; Gibson and Wu, 2013; Jäger et al., 2015; Turkish: Kahraman et al., 2010). By definition, in a head-final relative clause construction, the relative clause appears before the head noun it modifies, meaning that the gap is encountered before the filler (rather than after it, as in English). Such structures are of crucial theoretical interest since they make it possible to reexamine the predictive power of the different competing sentence comprehension theories in a new context.

To illustrate the theoretical relevance of head-final relative clause processing, consider the Mandarin Chinese (henceforth "Chinese") sentences with relative clauses in (3) and (4)[3].

(3) Sentence with subject-extracted relative clause in Chinese:
_$_i$ aimu yinyuejia de zuoqujia$_i$ he-le yi bei jiu
_$_i$ adore musician REL composer$_i$ drink-ASP one glass wine
"The composer who adored the musician drank a glass of wine."

(4) Sentence with object-extracted relative clause in Chinese:
zuoqujia aimu _$_i$ de yinyuejia$_i$ he-le yi bei jiu
composer adore _$_i$ REL musician$_i$ drink-ASP one glass wine
"The musician whom the composer adored drank a glass of wine."

Chinese displays a head-initial structure in verb phrases: like English, Chinese is a Subject-Verb-Object (SVO) language, with verbs preceding their NP object complements. At the same time, however, Chinese displays a head-final structure in NP: modifiers of nouns exclusively appear before the head noun. Because of this combination, while subject gaps in Chinese are higher and structurally closer to the complementizer/relativizer than object gaps (as in English), subject gaps are linearly *farther* from the head noun (i.e., the filler) than object gaps, unlike English. These facts are illustrated in **Figure 2**, which diagrams the relative clauses from (3, 4)[4].

Regarding gap-filler integration, therefore, the linearity account predicts that the gap-filler relation in a Chinese OR should be less taxing to construct than that in an SR. The structure-based account predicts the opposite: since fewer structural nodes intervene between the head noun and a subject gap, the dependency between these two should be easier to construct compared to one involving an object gap. Both accounts would predict the locus of processing differences on the head noun where gap-filler integrations take place.

Regarding the effect of structural probabilities, given that SRs have higher frequencies than ORs in Chinese (Wu et al., 2011), greater surprisal values are associated with ORs and thus longer reading times in ORs are predicted. In terms of the effect of dominant thematic orders, since the canonical thematic order in Chinese is agent-verb-patient, ORs, which follow the dominant order, are predicted to be less costly to process (Lin, 2013, 2014). The experience-based effects make processing predictions for the whole sentences based on structural and word-order probabilities, not just for particular regions where integration costs incur.

Chinese has thus been taken as a valuable test case for validating the integration-based accounts and experience-based accounts depicted above (see also Jäger et al., 2015 for a review of the theoretical controversy). So far, research has provided a somewhat mixed picture. In the head-noun region, some studies have found that ORs took longer to read than SRs (Lin and Bever, 2006; Chen et al., 2012; Jäger et al., 2015) while others found the opposite (Gibson and Wu, 2013). One potential difficulty in acquiring consistent results is that studies differed regarding whether and how relative clauses are motivated. When relative clauses are not motivated (for example, when they appear in isolated sentences without referential contexts or structural cues preceding them), surprisal effects related to reanalyses may induce longer reading times in the disambiguating regions. Gibson and Wu (2013), for instance, pointed out that Chinese ORs may be more difficult to comprehend than SRs in neutral

[3]In these examples and throughout the paper, REL and ASP will be used as abbreviations for "relativizer" and "(perfective) aspect," respectively.

[4]We adopt a movement/raising analysis for the tree diagrams of subject-extracted and object-extracted relative clauses in Chinese (see Aoun and Li, 2003; Huang et al., 2009).

contexts because ORs are more likely to induce a garden path effect in the prenominal regions. Longer reading times for ORs are thus predicted in the head noun region, where disambiguation takes place.

On the other hand, when relative clauses are pragmatically motivated or structurally disambiguated, one needs to consider the potential effects of the different contextual cues. Chinese relative clauses have previously been pragmatically motivated by using discourse contexts (Gibson and Wu, 2013; Lin, 2014), and structurally disambiguated by using classifier-noun mismatching cues (Hsu et al., 2006) and classifier-adverbial sequences (Jäger et al., 2015). In studies that motivated relative clauses by using referential contexts (Lin, 2014; cf. Gibson and Wu, 2013), relative clause processing was shown to be sensitive to the order of thematic roles in the context: relative clauses whose thematic orders match those in the referential contexts showed shorter reading times in the regions after the head noun. In studies where relative clauses were structurally disambiguated, reading patterns have been found to be consistent with the conditional structural probabilities of SRs and ORs. Jäger et al. (2015), for instance, reported reading patterns consistent with surprisal predictions based on a corpus study and a sentence completion task. In Chinese relative clauses that follow disambiguating syntactic contexts like classifier-adverbial sequences, the conditional probability of an SR is higher than that of an OR in the embedded clause regions (i.e., IPs in **Figure 2**) but not on the head noun. Reading patterns confirmed that an SR advantage existed in the embedded clause regions but not on the head noun.

Methodologically, processing studies comparing Chinese SRs and ORs have reached a bottleneck. In most previous studies, SRs have been directly compared to ORs, meaning that SRs and ORs serve as each other's baseline conditions. Accordingly, any processing difference between the two has typically been associated with one single factor of theoretical interest. For instance, Gibson (1998) focuses on differences in linear distance between the gap and the filler, whereas theories of structural complexity (O'Grady, 1997; Hawkins, 2004) focus on differences in the number of structural layers/nodes intervening between the two. In fact, however, SRs and ORs are different from each other in multiple ways beyond these differences. In terms of constructional frequencies, SRs are more common than ORs (Lin, 2009; Wu et al., 2011). In terms of structural predictability, an SR is better expected than an OR (Jäger et al., 2015). In terms of nominal animacy preferences, the heads of SRs are preferably animate while those of ORs are preferably inanimate (Wu et al., 2012). Because SRs and ORs are simultaneously different from each other in so many ways, results from previous studies comparing the two are difficult to interpret.

The present study addresses this methodological issue by holding the extraction site constant (only SRs) and investigating the processing of three different *sub-types* of SRs: standard SRs, SRs with the disposal *ba* construction (henceforth "disposal SRs"), and SRs with the long passive *bei* construction (henceforth "passive SRs"). Both the disposal *ba* construction and the passive *bei* construction involve functional morphemes that have been analyzed as light verbs or grammaticalized verbs in Mandarin Chinese. An example for each construction is given in (5–7). Sentences with relative clauses appear after referential contexts, which are intended to pragmatically motivate relative clauses so that the prenominal relative clauses are parsed as relative clauses when they appear in sentences.

(5) Standard SR:
__$_i$ jiaoxing furen de zuoqujia$_i$ he yi bei jiu
__$_i$ wake.up lady REL composer$_i$ drink one glass wine
action *PATIENT* *AGENT*
"The composer that woke up the lady drank a glass of wine."

(6) Long passive (*bei*) SR:
__$_i$ bei furen jiaoxing de zuoqujia$_i$ he yi bei jiu
__$_i$ BEI lady wake.up REL composer$_i$ drink one glass wine
AGENT *action* *PATIENT*
"The composer that was woken up by the lady drank a glass of wine."

(7) Disposal (*ba*) SR:
__$_i$ ba furen jiaoxing de zuoqujia$_i$ he yi bei
__$_i$ BA lady wake.up REL composer$_i$ drink one glass
PATIENT *action* *AGENT*
jiu
wine
"The composer that woke up the lady drank a glass of wine."

Being SRs, all three structures involve the extraction and relativization of the subject NP, which, in Chinese, involves a movement type of dependency between the subject gap and the head NP (Aoun and Li, 2003). Where these three structures differ from one another is the internal structure of the pre-relativizer inflectional phrase (IP)—in particular, the structure of the verb phrase (VP) and the small verb phrase (vP) following the subject gap. Each of these three constructions will now be discussed in turn.

The syntactic structure of a standard SR is illustrated in **Figure 3**, representing the relative clause portion of (5). Standard SRs contain an SVO sequence with an empty subject NP inside the IP.

The syntactic structure of a passive SR is illustrated in **Figure 4**, representing the relative clause portion of (6) above. Under the main-verb analysis for Chinese long passives (Huang et al., 2009), this structure contains an empty subject and

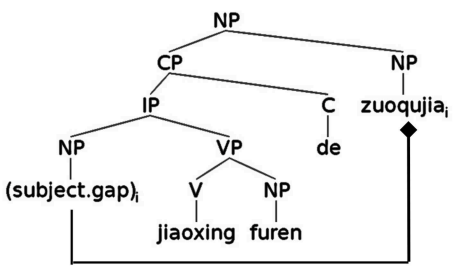


FIGURE 3 | Syntactic structure of a standard SR in Chinese.

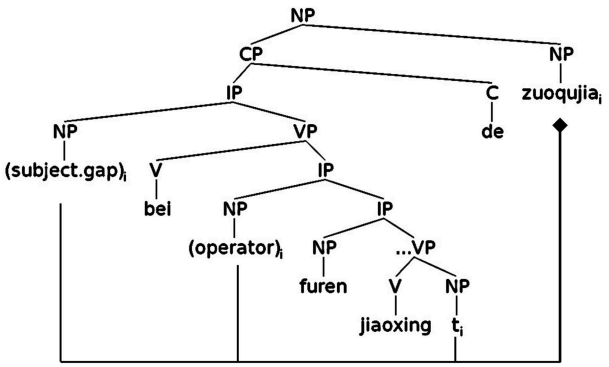


FIGURE 4 | Syntactic structure of a passive SR in Chinese.

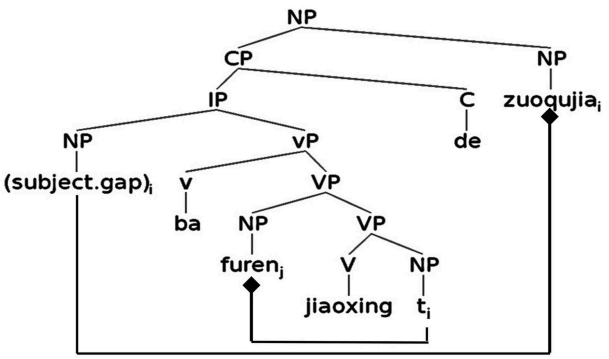


FIGURE 5 | Syntactic structure of a disposal SR in Chinese.

a VP headed by *bei*, followed by a secondary predicate IP[5]. Three dependencies are involved in this construction. First, as in all the SRs, a dependency exists between the subject gap and the head NP. Second, an additional dependency exists between the base generated object NP position in the lower VP and the NP operator at the periphery of the intermediate IP. Third, this NP operator holds the same identity as the subject gap. The three empty positions (the subject gap, the operator, and the trace) all bear the same identity as the head NP.

The structure of a disposal SR is illustrated in **Figure 5**, representing the relative clause portion of (7) above. Like passive SRs, disposal SRs also involve multiple dependencies. Under the light verb analysis of *ba* (e.g., Huang, 1997; Lin, 2001), the object NP of the lower VP is displaced to the specifier position. Two separate dependencies involving empty categories need to be constructed in the processing of a disposal SR: one between the subject gap and the head NP (outer connection in **Figure 5**), and the other between the moved object NP immediately following *ba* and the position of its trace (inner connection in **Figure 5**). Unlike passive SRs, the VP-internal dependency in a disposal SR is nested inside the dependency between the subject gap and the head NP[6].

The processing factors discussed above cast different effects on these three types of Chinese SRs. Let's first focus on the integration effects regarding the dependency between the subject gap and the head noun in each of the three structures, which are usually taken to be observable around the head noun region, where filler-gap integrations take place. In terms of linear locality (Gibson, 1998; Hsiao and Gibson, 2003), the same numbers of *new referents* intervene between the gap and the filler, thus predicting no processing differences. If linear distance is computed using the number of intervening *words*, then the passive SRs and the disposal SRs may both require greater processing load than the standard SRs because they involve an additional function word (*bei* and *ba*, respectively) between the gap and the filler. In terms of structural locality, since all three SRs involve extraction from subject position, the structural distance between the head noun and the gap are identical across all three structures (passing through two XP nodes—one CP and one IP), thus predicting no processing differences.

In addition to the gap-filler dependencies, the passive SRs and the disposal SRs involve additional displacement dependencies as depicted in **Figures 4**, **5**. For a passive SR, the sentence-initial passive marker *bei* indicates a missing subject NP that is to be connected with the object NP in the lower VP. Assuming that a relative clause parse has been adopted, the missing subject NP is taken to be connected both with the object NP and the head noun[7]. For a disposal SR, the sentence-initial light verb *ba* also indicates a missing subject NP. Assuming again the processing of a relative clause, this missing subject NP would be taken as a subject gap connected with the head noun. An additional dependency in a disposal SR involving the displaced object NP after *ba* would add to the processing cost already incurred by the SR. The integration-based accounts, taking into consideration these additional dependencies, would then predict that both a passive SR and a disposal SR should be harder than a standard SR because (a) the former SRs involve additional dependencies, and (b) the dependencies in the former SRs are longer and more complex than that of a standard SR.

[5]Alternatively, *bei* has also been analyzed as a preposition taking the NP following it as its oblique object in the long passives and the subject NP as an NP displaced from the object NP in the lower VP (Li, 1990). In this analysis, instead of three dependencies, two dependencies—one between the subject and the lower object and one between the subject gap and the head noun—are involved. While the main verb analysis, which Huang et al. (2009) persuasively argued for, is adopted in the present study, similar predictions about how passive SRs are processed in comparison with standard SRs and disposal SRs can be made when the alternative analysis is adopted.

[6]Like *bei*, the categorical status of *ba* is controversial. In addition to the light verb analysis adopted in the present article, it has also been analyzed as a lexical verb (Hashimoto, 1971), a preposition (Chao, 1968; Li, 1990), and a function word that assigns case (Huang, 1982; Goodall, 1987). In these analyses, the object NP forms a local syntactic constituent with *ba*, through which it is connected with the verb. The dependency between the object NP and the verb is still nested inside the dependency between the subject gap and the head noun.

[7]The integration cost associated with passive SRs may also need to consider the base-generated lowermost trace position, which is linearly closest to the head NP. Even though this short linear dependency may exist between the passivized NP trace and the head noun, it does not preclude the processor from establishing a dependency between the trace NP and the subject gap, which involves a longer linear distance than the dependency in a standard SR.

Processing differences are expected to appear on and before the head noun.

Next, we consider the overall structural complexity and structural frequencies involved in the three types of SRs. The standard SR is the simplest of the three constructions, as it contains the fewest number of structural layers and only has a single dependency relation (between the subject gap and the head). Passive SRs and disposal SRs are both more complicated in terms of the intricate dependency relations inside the VP/vP[8]. This hierarchy of complexity is consistent with the constructional frequencies of the 3075 relative clauses extracted from the Sinica Treebank (Version 3.0; Chen et al., 2003) by Lin (2009), among which standard SRs accounted for 53%, passive SRs accounted for 2%, and no instances of disposal SRs were found[9]. Thus, based on both structural complexity and constructional frequency, a standard SR should be the easiest to process among the three.

On the other hand, the thematic order effect predicts different processing preferences. Since the surface thematic order of a passive SR matches the canonical thematic order in Chinese (i.e., AGENT-*action*-PATIENT), a passive SR should be the easiest to process among the three. Conversely, the thematic orders of standard SRs and disposal SRs are inconsistent with this dominant thematic order and should be more difficult to process than the passive relatives.

One relevant hypothesis about effect locus proposed by Lin (2014) is that the pre-relativizer and post-relativizer regions of a head-final relative clause may reveal different processing effects. This hypothesis is directly related to the existence of uncertainty in processing head-final relative clauses: while the pre-relativizer regions are structurally ambiguous, the post-relativizer regions are structurally unambiguous. The pre-relativizer regions of an OR, for example, with the word order *Noun-Verb*, are more likely to be read as matrix clauses than subordinate relative clauses. The corresponding pre-relativizer *Verb-Noun* sequence of an SR would be parsed as a matrix clause with a missing subject argument before the verb (see Lin and Bever, 2011; Jäger et al., 2015 for more elaborated discussion on the issue of garden path in Chinese relative clauses). With the post-relativizer regions, however, no similar ambiguity exists since comprehenders tend to parse the functional morpheme *de* after the embedded clause as a relativizer. A corpus study and a sentence completion task by Jäger et al. (2015) have confirmed that a relative clause parse is already unambiguously established when the relativizer is reached. Lin (2014), in particular, proposes the effect of thematic templates, being a pattern matching effect, may be more observable in the post-relativizer regions where structural uncertainty has decreased.

In addition to the overall predictions of the effects, we thus further distinguish the processing effects in the pre-relativizer regions and the post-relativizer regions. In the pre-relativizer regions, disposal SRs and passive SRs are both expected to take longer to process than standard SRs given greater structural complexity and lower structural frequencies[10]. Integration effects based on linear locality and structural locality would make similar predictions given that simpler dependent relations exist in the standard SRs than in the disposal and passive SRs. The effect of thematic template mapping, on the other hand, predicts shorter reading time for passive SRs because they display thematic orders consistent with the canonical order in Chinese though this effect may emerge later in a prenominal relative clause construction.

In the post-relativizer regions, where the existence of relative clauses are clearly indicated by the relativizers and the head nouns, an integration account based on linear locality would predict that standard SRs be easier than both disposal and passive SRs, especially around the head noun region. An integration account based on structural locality would predict no processing differences, or easier processing for standard SRs due to the complexity effect possibly spilled over from the prenominal regions. The effect of thematic template mapping is the only theory that predicts an overturned reading pattern for passive SRs, with passive SRs being the least costly to read. The effect of thematic template mapping is expected to span across multiple post-relativizer regions.

The goal of the present study, in summary, is to examine the effect of thematic orders on Chinese relative clause processing. While Lin (2014) reported that the processing of SRs and ORs in Chinese is sensitive to the thematic orders presented in the context, it directly compared the processing of SRs and ORs, which as discussed, involve an array of differences that may obscure the effects. The present study contrasted the processing of three sub-types of SRs, thus keeping constant the extraction site regarding its grammatical function in the embedded clause. Furthermore, Lin (2014) studied the effect of thematic orders by varying the orders in the referential context while keeping the thematic orders in the relative clauses constant. The present study examined this effect by varying the thematic orders in the relative clauses while keeping the thematic orders in the context

[8]It is not a simple matter to determine whether a passive SR or a disposal SR is more complex. In terms of the number of dependencies and structural layers, a passive SR is more complex. In terms of the number of different NP identities involved in the dependencies, a disposal SR is more complex. Moreover, while both kinds of SRs exhibit nested dependencies, these dependencies are all associated with the same NP for passive SRs. This factor may make the construction of such dependencies easier to process than the multiple distinct dependencies of a disposal SR. In this sense, then, disposal SRs may be the more complex of the two.

[9]The Sinica Treebank can be found at the following URL: http://turing.iis.sinica.edu.tw/treesearch/. Passive constructions (using *bei*) and disposal constructions (using *ba*) are also less common in Chinese *overall* compared to canonical VO orders. In the Sinica Corpus, *ba* accounted for 14.4% of the words in the corpus and *bei* 14.8% (CKIP online word frequency list http://elearning.ling.sinica.edu.tw/eng_teaching.html, retrieved on September 17, 2012). These overall frequency differences are mirrored in processing differences: Lin (2006, 2008) found that, compared to canonical SVO sentences, disposal sentences and passive sentences showed lower acceptability ratings in naturalness-judgment questionnaires as well as longer reading times in online self-paced reading tasks.

[10]Constructional frequency is but one way to make expectation-based predictions. Alternatively, it is also possible to conduct a sentence completion task to generate word-by-word structural expectations (as has been done in Jäger et al., 2015). The sentence completion task will be particularly useful for distinguishing the processing of passive SRs and disposal SRs. Since the main contrast of interest in the present study is between standard SRs and passive SRs, using corpus counts and constructional frequencies should be sufficient for making the expectation-based predictions. Nonetheless, we leave an actual sentence completion task as an open possibility for generating more fine-grained word-by-word expectation-based predictions.

constant. It is hoped that this new manipulation can test the effect of thematic order on relative clause processing from a new angle.

Materials and Methods

Participants

Forty-eight Taiwanese college students at National Taiwan Normal University, all native speakers of Mandarin Chinese, participated in the experiment. The participants were screened for brain damage. All had normal (or corrected to normal) vision, and were naïve to the purpose of the experiment. Participants gave informed consent to take part in the study. The study protocol was approved by Indiana University's Institutional Review Board.

Materials

Twenty sets of sentences were included as the experimental trial, 16 of which were modified based on Gibson and Wu's (2013) stimuli. The experimental materials were created in such a way that they read naturally in Mandarin disposal and passive constructions. To motivate the relative clauses, each set consisted of a referential context introducing transitive relations in which three referents are involved, as in (8). The sentences in the context where these thematic relations were introduced present the thematic order of AGENT-*action*-PATIENT. Following each context was a dialogue between two interlocutors, Xiaoming and Xiaomei, in which Xiaoming asks Xiaomei to identify one referent out of the two active referents, as in (9). Xiaomei's response starts with the target relative clause presented in a word-by-word moving window format. A sample of the experimental materials is given below:

(8) Context:
Yidong gongyuli zhule fangdong yiji liangge fangke
one apartment lived landlord and two tenants
"A landlord and two tenants lived in an apartment."

Yiwei zhuhu chaoxingle fangdong
one tenant woke up landlord
"One of the tenants woke up the landlord."

Fangdong ze chaoxingle lingyiwei zhuhu
landlord then woke up the other tenant
"The landlord woke up the other tenant."

Xiaoming: Wo tingshuo qizhong yiming zhuhu hen gao
I heard among them one tenant very tall
"I heard one of the tenants was very tall."

Nayiwei zhuhu hen gao?
which.one tenant very tall
"Which tenant was very tall?"

(9) Target sentence with a relative clause:
(i) Standard SR
Xiaomei: Chaoxing fangdong de zhuhu hen gao
woke up landlord REL tenant very tall
V N REL Head Noun HN+1 HN+2
"The tenant that woke up the landlord was very tall."

(ii) Passive SR
Xiaomei: Bei fangdong chaoxing de zhuhu hen gao
BEI landlord woke up REL tenant very tall
BEI N V REL Head Noun HN+1 HN+2
"The tenant that was woken up by the landlord was very tall."

(iii) Disposal SR
Xiaomei: Ba fangdong chaoxing de zhuhu hen gao
BA landlord woke up rel tenant very tall
BA N V REL Head Noun HN+1 HN+2
"The tenant that woke the landlord up was very tall."

Forty-eight additional sets of sentences following a similar format served as fillers. Sixteen of these fillers had relative clauses of various types in them; the remaining 32 fillers did not contain relative clauses. Altogether, 68 sets of contexts and sentences were pseudorandomly presented so that no two experimental trials appeared consecutively. Comprehension questions followed each trial to ensure that participants paid attention in reading the experimental materials. The words used in the relative clauses are provided in the Supplementary Materials.

Procedure

The experiment followed the standard moving-window self-paced reading design and was conducted using Linger 2.94 (developed by Doug Rohde)[11]. In each trial, participants took their own pace hitting the spacebar to proceed to the next sentence or region. The contexts were presented sentence by sentence, and the target sentences (i.e., Xiaomei's response to Xiaoming's query) were presented word by word. For disposal and passive SRs, *ba,* and *bei* were presented in the same region as the following noun. After the last word, participants were given a true/false comprehension question focusing on the overall content of the context or the target sentence. Feedback was given if the participant's response was incorrect. Participants were instructed to read the sentences at a natural pace in order to answer the comprehension questions correctly. The reading time for each region, the time taken to answer the comprehension questions, and the responses to the comprehension questions were recorded. The whole experiment took an average of 40 min to complete.

Results

Linear mixed-effects models treating both subjects and items as random effects were fit to both the comprehension accuracy data and the region-by-region reading time data using the lme4 package version 1.1-7 in R (version 3.2.0; Bates et al., 2015). Two contrasts were defined comparing the passive SRs with the standard SRs (passive SR coded as +1, standard SR coded as −1) and comparing the passive SRs with the disposal SRs (passive SR coded as +1, disposal SR coded as −1). The analyses were carried out on log-transformed values of the reading times and residuals were checked to ensure that the normality requirement is met.

[11]See http://tedlab.mit.edu/~dr/Linger/ (retrieved on December 9, 2012) for documentation of Linger 2.94.

The package lmerTest (version 2.0-25) in R is used to verify the levels of statistical significance. The *t*-value of 2 is taken to be the threshold for statistical significance at $\alpha = 0.05$. Question-accuracies were analyzed using generalized linear mixed models with a binomial link function. The dependent measures included comprehension accuracies (binary results), latencies in answering comprehension questions, and region-by-region reading times.

Comprehension Accuracy

The mean comprehension accuracy for all items was 85% and the mean accuracy for the experimental trials was 90%. The accuracies of each of the three experimental conditions were 93.05% (passive SRs), 91.83% (standard SRs), and 86.28% (disposal SRs). These results are summarized in **Figure 6**. Statistical results are given in **Table 1**.

In terms of overall comprehension accuracy, passive SRs were comprehended more accurately than both standard SRs and disposal SRs. These results are consistent with the predictions of thematic order effect. Namely, passive SRs, whose thematic order followed the canonical thematic order, were comprehended with greater accuracies than both standard SRs and disposal SRs. No difference was found on the time taken to respond to the comprehension questions.

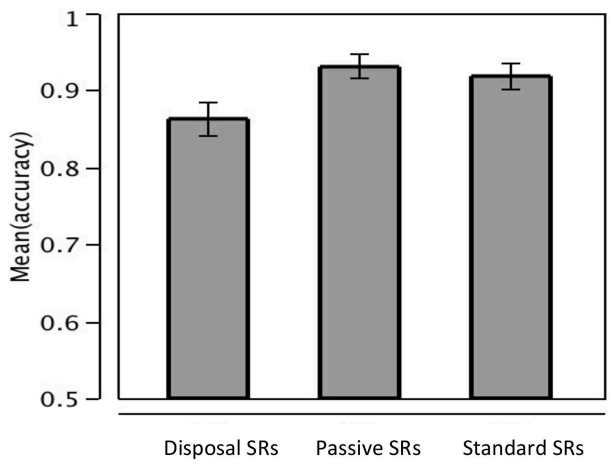


FIGURE 6 | Comprehension accuracies of disposal SRs, passive SRs, and standard SRs (error bars indicating one standard error).

Reading Times

Since the regions before and after the relativizers are hypothesized to be reflective of different processing effects, average reading times in the two pre-relativizer regions were compared to those in the post-relativizer regions (from the head noun to two regions after the head noun) across the three conditions. **Figure 7** illustrates the results of this analysis. Statistical results are given in **Table 2**.

In the pre-relativizer regions, passive SRs were read longer than standard SRs. In the post-relativizer regions, passive SRs were read faster than both the standard SRs and the disposal SRs. The reading time of each target region, including the two pre-relativizer regions, the relativizer, the head noun, and the two regions after the head noun, is further summarized in **Figure 8**. Statistical results of the by-region reading time analyses are given in **Table 3**.

Passive SRs were read longer than standard SRs in both regions inside the relative clause (i.e., the pre-relativizer regions), and faster than disposal SRs from the second region in the prenominal clause to the head noun region[12]. In the second region after the head noun, passive SRs were read faster than standard SRs.

To sum up, standard SRs were read with greatest ease in the earlier regions of the relative clauses. In contrast, in the regions following the relativizer, passive SRs were read more quickly than standard SRs and disposal SRs. The easier comprehension of standard SRs in the pre-relativizer regions is consistent both with integration effects (i.e., standard SRs having less complicated dependencies) and with expectation-based constructional frequency effects (i.e., standard SRs being more frequently experienced than passive SRs). The easier comprehension of passive SRs in the post-relativizer regions, on the other hand, is only consistent with the prediction of thematic template mapping.

General Discussion

The present study contrasted the reading patterns of three types of SRs in Chinese: standard SRs, passive SRs, and disposal SRs. Distinctive reading patterns were observed in the

[12] As a caveat to the advantage of standard SRs observed in the pre-relativizer regions, the disposal SRs and the passive SRs involve an additional function word (i.e., *ba* and *bei*) in the first region, which could induce longer reading times in these regions. The different words in these two pre-relativizer regions also make region-by-region comparisons less straightforward.

TABLE 1 | Summary of model estimates, standard errors, and the *t* or *z* values for comprehension accuracy and response latency.

	Comprehension accuracy				Comprehension latency		
Contrast	**Coef.**	***SE***	***z*-values**	***P*(>|*z*|)**	**Coef.**	**SE**	***t*-values**
Intercept	**2.61**	**0.00**	**2105.80**	**<0.001**	**7.67**	**0.06**	**121.44**
Passive_SR-Standard_SR	**0.06**	**0.00**	**51.70**	**<0.001**	0.02	0.03	0.95
Passive_SR-Disposal_SR	**0.71**	**0.00**	**573.80**	**<0.001**	0.01	0.03	0.55

Statistically significant ($\alpha = 0.05$) effects are highlighted in bold.

regions before and after the relativizer, suggesting the effects of different processing factors being operative. While the current experimental design intends to motivate relative clauses by using referential contexts, it is still unclear whether a relativized gap has indeed been postulated in the pre-relativizer regions given that a relative-clause parse is but one of several possible parses for the pre-relativizer regions. The structurally-ambiguous pre-relativizer regions showed reading patterns consistent with expectation-based theories of sentence comprehension (e.g., the uncertainty-reduction accounts of Hale, 2006 and Chen et al., 2012; see also Jäger et al., 2015), which rely on the probabilities of particular syntactic categories and constituents appearing at particular positions of a sentence. Standard SRs, being the most common prenominal structure of the three, are found to be easier to understand. Besides expectation-based effects, the reading patterns in the pre-relativizer regions are also compatible with integration-based effects, which, as discussed, predict easier processing on structures that involve simpler dependencies. In comparing the three types of SRs, a standard SR involves fewer dependencies and presents a simpler dependency structure.

When the relativizer region is reached, the existence of a relative clause is unambiguously indicated. Consistent with the prediction of the thematic order effect, a passive SR was read faster than the corresponding standard SR and disposal SR given that the thematic order in a passive SR is more frequently experienced than that in a standard SR and that in a disposal SR. All other theoretical factors, by contrast, favor the processing of a standard SR given its structural simplicity and greater constructional frequency. Moreover, this effect of thematic ordering was observed to span across several post-relativizer regions, being attested from the relativizer to the second region after the head noun individually as well as in the sum total. The thematic order effect therefore seems qualitatively different from the gap-filler integration effects, which are usually *localized* to the head noun region.

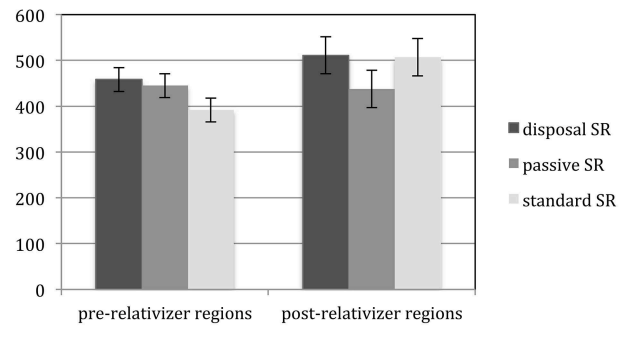


FIGURE 7 | Average reading times of disposal SRs, passive SRs, and standard SRs in the regions before and after the relativizer (error bars indicating one standard error).

In previous research on Chinese SR/OR processing, similar asymmetries have been found before and after the relativizer. Recall that, the thematic order of agent-verb-patient found in an OR, which is similar to that in a passive SR, may give a Chinese OR a processing edge over its SR counterpart owing to the thematic order effect. In contrast to an SR, the pre-relativizer regions of a Chinese OR present a word order (i.e., noun-verb) that matches the canonical order in a Chinese sentence and may be read with greater ease than those of a Chinese SR, whose pre-relativizer verb-noun sequence is non-canonical. In previous studies where relative clauses were not structurally disambiguated, greater processing costs were indeed associated with the pre-relativizer regions of an SR—an effect consistent with the prediction of structural probabilities as well as thematic orders (Hsiao and Gibson, 2003; Chen et al., 2008; Qiao et al., 2012). When the relative clauses were structurally disambiguated, however, SRs were processed with greater ease than ORs owing to SRs' greater structural predictability after disambiguating contexts (Jäger et al., 2015)—an effect that is consistent with the prediction of structural probabilities only.

In the post-relativizer regions, an OR disadvantage has been reported for relative clauses modifying the object of an SVO sequence (Lin and Bever, 2006). This effect has been attributed to the reanalysis of a garden-path parse in such structures given that no contextual cues indicated a relative clause parse on the left edge (Lin and Bever, 2011). Most relevant to the current findings, however, in studies that used referential contexts to motivate Chinese relative clauses, an OR advantage consistent with the thematic order effect reported in the current study was obtained (Gibson and Wu, 2013; Lin, 2014).

The thematic order effect on processing Chinese relative clauses is also supported by two offline studies on aphasic patients' processing of Chinese relative clauses: Law and Leung (2000) and Su et al. (2007). Using picture-matching tasks, both studies found better performance on ORs compared to SRs, which was attributed to the fact that Chinese ORs (but not Chinese SRs) match the canonical thematic order. These results are also compatible with the SR advantage of English-speaking aphasic patients (Caplan and Futter, 1986; Grodzinsky, 1986; Hagiwara and Caplan, 1990). An implication of the thematic

TABLE 2 | Summary of model estimates, standard errors, and the *t* values for reading times in the pre-relativizer and post-relativizer regions.

	Pre-relativizer regions			Post-relativizer regions		
Contrast	**Coef.**	***SE***	***t*-values**	**Coef.**	***SE***	***t*-values**
Intercept	**5.93**	**0.09**	**68.67**	**5.96**	**0.11**	**54.92**
Passive_SR-Standard_SR	**0.11**	**0.02**	**5.08**	**−0.06**	**0.02**	**−2.73**
Passive_SR-Disposal_SR	−0.03	0.02	−1.58	**−0.05**	**0.02**	**−2.37**

Statistically significant ($\alpha = 0.05$) effects are highlighted in bold.

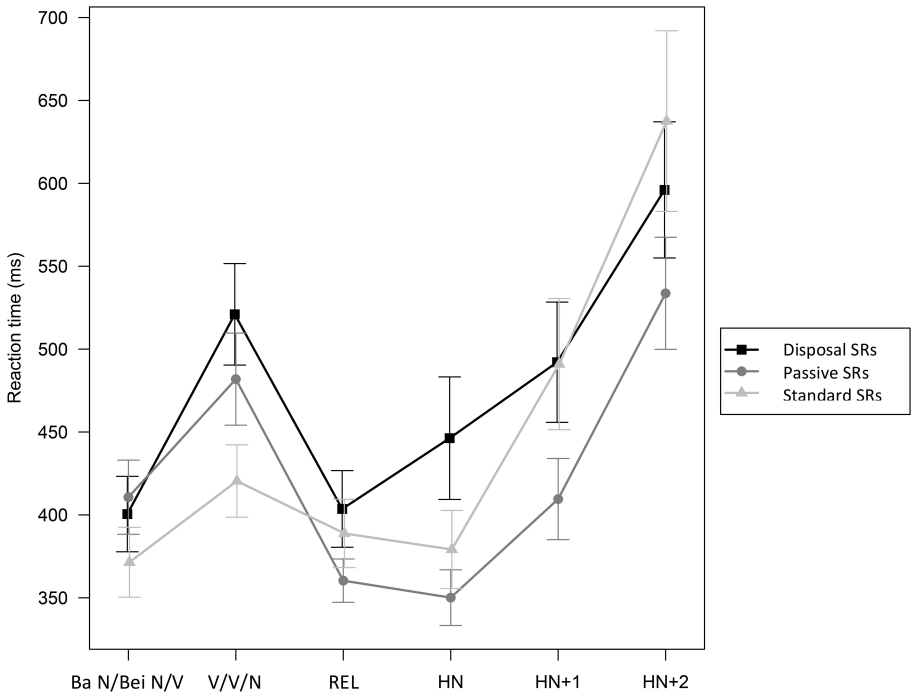


FIGURE 8 | Reading time of each critical region in the disposal SRs, passive SRs, and standard SRs (error bars indicating one standard error). See (9) for region coding.

TABLE 3 | Summary of model estimates, standard errors, and the *t* values for reading times in each of the critical regions.

	Ba N/Bei N/V			V/V/N			REL		
Contrast	**Coef.**	***SE***	***t*-values**	**Coef.**	***SE***	***t*-values**	**Coef.**	***SE***	***t*-values**
Intercept	**5.86**	**0.05**	**127.54**	**6.00**	**0.06**	**106.23**	**5.85**	**0.04**	**152.92**
Passive_SR-Standard_SR	**0.09**	**0.03**	**3.47**	**0.13**	**0.03**	**3.89**	−0.02	0.03	−0.91
Passive_SR-Disposal_SR	0.00	0.03	0.07	**−0.07**	**0.03**	**−2.08**	**−0.05**	**0.03**	**−1.97**
	Head Noun			**Head Noun + 1**			**Head Noun + 2**		
Contrast	**Coef.**	***SE***	***t*-values**	**Coef.**	***SE***	***t*-values**	**Coef.**	***SE***	***t*-values**
Intercept	**5.83**	**0.05**	**122.23**	**5.93**	**0.07**	**90.77**	**6.11**	**0.08**	**72.13**
Passive_SR-Standard_SR	−0.04	0.03	−1.11	−0.06	0.04	−1.78	**−0.08**	**0.04**	**−1.96**
Passive_SR-Disposal_SR	**−0.11**	**0.03**	**−3.27**	−0.06	0.04	−1.55	0.01	0.04	0.20

Statistically significant ($\alpha = 0.05$) effects are highlighted in bold.

order effect is that the advantage previously reported for an OR advantage in Mandarin and an SR advantage in English should be re-considered since Mandarin ORs and English SRs, like the passive SRs in the current study, present a canonical thematic order. When comparing SRs and ORs, the advantage for processing Chinese ORs may be due to the ORs presenting canonical thematic orders, but not the SRs.

In the current study, the reading patterns of disposal SRs are contrasted with those of standard SRs and passive SRs. Given their lower constructional frequency and greater number of dependencies involving empty categories, disposal SRs were expected to be the most difficult to process. Indeed, the reading patterns in the present study showed that disposal SRs were the most difficult among the three SRs examined in both the pre- and post-relativizer regions. Given the additional dependencies and lower structural probability associated with passive SRs, it may be expected that they should be equally difficult to process. This result was only obtained for the pre-relativizer regions, where passive SRs were read longer than the standard SRs. In the post-relativizer regions, the reading times of passive SRs were shorter than those of standard SRs and disposal SRs. This can be taken as evidence that the canonical thematic order found in a passive SR induced shorter reading times in its post-relativizer regions. The fact that structural probability effects and thematic template

effects have been observed in different regions of a relative clause does not imply that these processes are only operative in different regions of a sentence. Taken together, the results from these different studies suggest that the surprisal-related effect and the thematic template effect are both active and can be independently observed in different regions of a Chinese sentence.

The effect of thematic ordering on sentence comprehension can be understood as a processing heuristic used for efficiently coming up with thematic interpretations for sentences. The sentence processor keeps track of the linear positions of the content words in a sentence in forming thematic interpretations. The dominant thematic order of a language may serve as an "interpretation template," to which the content words of a sentence are matched. The comprehension of sentences with more complex structures such as relative clauses can be facilitated by matching thematic orders against the dominant thematic templates. Since the dominant thematic template in Chinese is AGENT-*action*-PATIENT, constructions matching this thematic order (such as ORs and passive SRs) may be comprehended with greater ease. This thematic template effect may also be effective in the comprehension of SRs in English, whose surface thematic order matches the dominant thematic order in the language.

These effects of thematic order are in line with several existing theories of sentence processing. The idea of thematic templates has a similar flavor to Bever's (1970) NVN heuristics—later referred to as "pseudosyntax" in Townsend and Bever (2001). In addition, mapping with thematic templates is also consistent with the "good enough" or "shallow processing" heuristics advanced by Ferreira (2003)[13]. We suggest that in order to arrive at a "good enough" impression of thematic relations, nouns and verbs are matched with the preexisting thematic templates. When the argument order in a sentence follows the dominant thematic template, the thematic roles of the nouns and verbs are easy to identify. Conversely, when the argument order is atypical, it is more difficult to identify thematic relations.

[13]In experiments requesting participants to identify the thematic roles of subjects and objects, Ferreira (2003) found that participants made more errors in sentences with atypical thematic orders (e.g., English passive sentences) than sentences with typical thematic orders (e.g., English active sentences). Moreover, this effect was found to be independent of the frequency of the relevant syntactic structures.

Conclusion

In conclusion, the reading time data for three sub-types of Chinese SRs reported in the present study supported two processes that are involved in the comprehension of Chinese relative clauses. Before reaching the relativizer, where the structure of the sentence is temporarily ambiguous, expectation-based incremental processing theories such as those of Hale (2001, 2006) and Levy (2008) can account for the processing differences across the three kinds of SRs though the results are also compatible with the integration-based predictions. Starting from the relativizer and the head noun, where the existence of a relative clause is unambiguously indicated, a global effect of thematic ordering was observed.

The critical evidence for the effect of thematic ordering comes from the easier processing of passive SRs, whose thematic role order conforms to the canonical thematic order of Chinese. Despite their more complex structural dependencies and lower constructional frequency compared with standard SRs, passive SRs were nevertheless comprehended with the greatest accuracy and processed with the shortest reading times in the post-relativizer regions. The current study therefore suggests that the comprehension of relative clauses in Chinese is sensitive to both the structural probabilities of constituents as well as the thematic orders involved in the relative clauses. In our effort to understand relative clause comprehension, it is important to take both of these factors into account.

Acknowledgments

This research was partially funded by Indiana University's Office of the Vice Provost for Research through the Faculty Research Support Program. I thank Aaron Albin, Yu-Jung Lin, and Chung-Lin Yang for their research assistance.

Supplementary Material

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpsyg.2015.01255

References

Aoun, J., and Li, Y.-h. A. (2003). *Essays on the Representational and Derivational Nature of Grammar: The Diversity of Wh-Constructions*. Cambridge, MA: MIT Press.

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2015). *lme4: Linear Mixed-effects Models using Eigen and S4 [Computer Software Manual]*. Available online at: http://CRAN.R-project.org/package=lme4 (R package version 1.1–7).

Bever, T. G. (1970). "Cognitive basis for linguistic structures," in *Cognition and the Development of Language,* ed J. R. Hayes (New York, NY: Wiley), 279–362.

Caplan, D., and Futter, C. (1986). Assignment of thematic roles to nouns in sentence comprehension by an agrammatic patient. *Brain Lang.* 27, 117–134. doi: 10.1016/0093-934X(86)90008-8

Carreiras, M., Duñabeitia, J. A., Vergara, M., de la Cruz-Pavía, I., and Laka, I. (2010). Subject relative clauses are not universally easier to process: evidence from Basque. *Cognition* 115, 79–92. doi: 10.1016/j.cognition.2009.11.012

Chao, Y.-R. (1968). *A Grammar of Spoken Chinese*. Berkeley, CA: University of California Press.

Chen, B., Ning, A., Bi, H., and Dunlap, S. (2008). Chinese subject-relative clauses are more difficult to process than the object-relative clauses. *Acta Psychol.* 129, 61–65. doi: 10.1016/j.actpsy.2008.04.005

Chen, K.-J., Huang, C.-R., Chen, F.-Y., Luo, C.-C., Chang, M.-C., Chen, C.-J., et al. (2003). "Sinica treebank: design criteria, representational issues and implementation," in *Building and Using Parsed Corpora,* ed A. Abeille (Dordrecht: Kluwer), 231–248. doi: 10.1007/978-94-010-0201-1_13

Chen, Z., Jäger, L., and Hale, J. (2012). "Uncertainty reduction as a predictor of reading difficulty in Chinese relative clauses," in *Proceedings of the 13th*

International Symposium on Chinese Languages and Linguistics (IsCLL), eds Y.-O. Biq and L. Chen (Taipei: National Taiwan Normal University), 245–261.

Ferreira, F. (2003). The misinterpretation of noncanonical sentences. *Cogn. Psychol.* 47, 164–203. doi: 10.1016/S0010-0285(03)00005-7

Ford, M. (1983). A method for obtaining measures of local parsing complexity throughout sentences. *J. Verbal Learn. Verbal Behav.* 22, 203–218. doi: 10.1016/S0022-5371(83)90156-1

Gennari, S. P., and MacDonald, M. C. (2008). Semantic indeterminancy and relative clause comprehension. *J. Mem. Lang.* 58, 161–187. doi: 10.1016/j.jml.2007.07.004

Gibson, E. (1998). Linguistic complexity: locality of syntactic dependencies. *Cognition* 68, 1–76. doi: 10.1016/S0010-0277(98)00034-1

Gibson, E., Desmet, T., Grodner, D., Watson, D., and Ko, K. (2005). Reading relative clauses in English. *Cogn. Linguist.* 16, 313–353. doi: 10.1515/cogl.2005.16.2.313

Gibson, E., and Wu, H.-H. I. (2013). Processing Chinese relative clauses in context. *Lang. Cogn. Process.* 28, 125–155. doi: 10.1080/01690965.2010.536656

Goodall, G. (1987). On argument structure and L-marking with Mandarin Chinese ba. *NELS* 17, 232–242.

Gordon, P. C., Hendrick, R., and Johnson, M. (2001). Memory interference during language processing. *J. Exp. Psychol. Learn. Mem. Cogn.* 27, 1411–1423. doi: 10.1037/0278-7393.27.6.1411

Grodzinsky, Y. (1986). Language deficits and the theory of syntax. *Brain Lang.* 27, 135–159. doi: 10.1016/0093-934X(86)90009-X

Hagiwara, H., and Caplan, D. (1990). Syntactic comprehension in Japanese aphasics: effects of category and thematic role order. *Brain Lang.* 38, 159–170. doi: 10.1016/0093-934X(90)90107-R

Hale, J. (2001). "A probabilistic Earley parser as a psycholinguistic model," in *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL 2001),* Vol. 2 (Stroudsburg, PA: NAACL), 159–166.

Hale, J. (2006). Uncertainty about the rest of the sentence. *Cogn. Sci.* 30, 609–642. doi: 10.1207/s15516709cog0000_64

Hashimoto, A. Y. (1971). Mandarin syntactic structures. *Unicorn* 8, 1–149.

Hawkins, J. A. (2004). *Efficiency and Complexity in Grammars*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780199252695.001.0001

Hsiao, F., and Gibson, E. (2003). Processing relative clauses in Chinese. *Cognition* 90, 3–27. doi: 10.1016/S0010-0277(03)00124-0

Hsu, C.-C. N., Hurewitz, F., and Phillips, C. (2006). "Contextual and syntactic cues for processing head-final relative clauses in Chinese," in *Paper Presented at the 19th Annual CUNY Conference on Human Sentence Processing* (New York, NY).

Huang, (1982). *Logical Relations in Chinese and the Theory of Grammar*. Doctoral dissertation, MIT.

Huang, C.-T. J. (1997). On lexical structure and syntactic projection. *Chinese Lang. Linguist.* 3, 45–89.

Huang, C.-T. J., Li, Y.-h. A., and Li, Y. (2009). *The Syntax of Chinese*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9781139166935

Jäger, L., Chen, Z., Li, Q., Lin, C.-J. C., and Vasishth, S. (2015). The subject-relative advantage in Chinese: evidence for expectation-based processing. *J. Mem. Lang.* 79–80, 97–120. doi: 10.1016/j.jml.2014.10.005

Kahraman, B., Sato, A., Ono, H., and Sakai, H. (2010). "Relative clause processing before the head noun: evidence for strong forward prediction in Turkish," in *Paper Presented at the 6th Workshop on Altaic Formal Linguistics (WAFL 6)* (Cambridge, MA).

King, J. W., and Just, M. A. (1991). Individual differences in syntactic processing: the role of working memory. *J. Mem. Lang.* 30, 580–602. doi: 10.1016/0749-596X(91)90027-H

King, J. W., and Kutas, M. (1995). Who did what and when? Using word-and clause-level ERPs to monitor working memory usage in reading. *J. Cogn. Neurosci.* 7, 376–395. doi: 10.1162/jocn.1995.7.3.376

Kwon, N., Gordon, P. C., Lee, Y., Kluender, R., and Polinsky, M. (2010). Cognitive and linguistic factors affecting subject/object asymmetry: an eye-tracking study of prenominal relative clauses in Korean. *Language* 86, 546–582. doi: 10.1353/lan.2010.0006

Law, S.-P., and Leung, M.-T. (2000). Sentence processing deficits in two Cantonese aphasic patients. *Brain Lang.* 72, 310–342. doi: 10.1006/brln.2000.2298

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition* 106, 1126–1177. doi: 10.1016/j.cognition.2007.05.006

Lewis, R. L., and Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cogn. Sci.* 29, 375–419. doi: 10.1207/s15516709cog0000_25

Li, Y.-h. A. (1990). *Order and Constituency in Mandarin Chinese*. Dordrecht: Kluwer. doi: 10.1007/978-94-009-1898-6

Lin, C.-J. C. (2006). *Grammar and Parsing: A Typological Investigation of Relative-clause Processing*. Ph.D. University of Arizona, Tucson.

Lin, C.-J. C. (2008). The processing foundation of head-final realtive clauses. *Lang. Linguist.* 9, 813–838.

Lin, C.-J. C. (2009). "Chinese relative clauses in corpus: processing considerations," in *Paper Presented at the 2009 International Conference on Applied Linguistics and Language Teaching (ALLT)* (Taipei).

Lin, C.-J. C. (2013). "Thematic templates and the comprehension of relative clauses," in *Language Down the Garden Path: The Cognitive and Biological Basis of Linguistic Structures,* eds M. Sanz, I. Laka, and M. K. Tanenhaus (Oxford: Oxford University Press), 141–148.

Lin, C.-J. C. (2014). Effect of thematic order on the comprehension of Chinese relative clauses. *Lingua* 140, 180–206. doi: 10.1016/j.lingua.2013.12.003

Lin, C.-J. C., and Bever, T. G. (2006). "Subject preference in the processing of relative clauses in Chinese," in *Proceedings of the 25th West Coast Conference on Formal Linguistics (WCCFL 25),* eds D. Baumer, D. Montero, and M. Scanlon (Somerville, MA: Cascadilla Press), 254–260.

Lin, C.-J. C., and Bever, T. G. (2011). "Garden path and the comprehension of head-final relative clauses," in *Processing and Producing Head-final Structures,* eds H. Yamashita, Y. Hirose, and J. L. Packard (New York, NY: Springer), 277–298.

Lin, T.-H. J. (2001). *Light Verb Syntax and the Theory of Phrase Structure.* Ph.D. University of California, Irvine.

Lin, Y., and Garnsey, S. M. (2011). "Animacy and the resolution of temporary ambiguity in relative clause comprehension in Mandarin," in *Processing and Producing Head-final Structures,* eds H. Yamashita, Y. Hirose, and J. L. Packard (New York, NY: Springer), 241–275.

MacDonald, M. C., and Christiansen, M. (2002). Reassessing working memory: comment on Just and Carpenter (1992) and Water and Caplan (1999). *Psychol. Rev.* 109, 35–54. doi: 10.1037/0033-295X.109.1.35

Miyamoto, E. T., and Nakamura, M. (2003). "Subject/object asymmetries in the processing of relative clauses in Japanese," in *Proceedings of the 22nd West Coast Conference on Formal Linguistics (WCCFL 22),* eds G. Garding and M. Tsujimura (Somerville, MA: Cascadilla Press), 342–355.

O'Grady, W. (1997). *Syntactic development*. Chicago: University of Chicago Press.

Packard, J. L., Ye, Z., and Zhou, X. (2011). "Filler-gap processing in Mandarin relative clauses: evidence from event-related potentials," in *Processing and Producing Head-final Structures,* eds H. Yamashita, Y. Hirose, and J. L. Packard (New York, NY: Springer), 219–240.

Qiao, X., Shen, L., and Forster, K. I. (2012). Relative clause processing in Mandarin: evidence from the Maze Task. *Lang. Cogn. Process.* 27, 611–630. doi: 10.1080/01690965.2011.578394

Raeli, F., and Christiansen, M. H. (2007). Processing of relative clauses is made easier by frequency of occurrence. *J. Mem. Lang.* 57, 1–23. doi: 10.1016/j.jml.2006.08.014

Roland, D., Dick, F., and Elman, J. L. (2007). Frequency of basic English grammatical structures: a corpus analysis. *J. Mem. Lang.* 57, 348–379. doi: 10.1016/j.jml.2007.03.002

Su, Y.-C., Lee, S.-E., and Chung, Y.-M. (2007). Asyntactic thematic role assignment by Mandarin aphasics: a test of the trace-deletion hypothesis and the double dependency hypothesis. *Brain Lang.* 101, 1–18. doi: 10.1016/j.bandl.2006.12.001

Townsend, D. J., and Bever, T. G. (2001). *Sentence Comprehension: The Integration of Habits and Rules*. Cambridge, MA: MIT Press.

Traxler, M. J., Williams, R. S., Blozis, S. A., and Morris, R. K. (2005). Working memory, animacy, and verb class in the processing of relative clauses. *J. Mem. Lang.* 53, 204–224. doi: 10.1016/j.jml.2005.02.010

Ueno, M., and Garnsey, S. M. (2008). An ERP study of the processing of subject and object relative clauses in Japanese. *Lang. Cogn. Process.* 23, 646–688. doi: 10.1080/01690960701653501

Van Dyke, J. A., and Lewis, R. (2003). Distinguishing effects of structure and decay on attachment and repair: a cue-based parsing account of recovery from misanalyzed ambiguities. *J. Mem. Lang.* 49, 285–316. doi: 10.1016/S0749-596X(03)00081-0

Wu, F., Kaiser, E., and Anderson, E. (2011). "Subject preference, head animacy and lexical cues: a corpus study of relative clauses in Chinese," in *Processing and Producing Head-final Structures* eds H. Yamashita, Y. Hirose, and J. L. Packard (New York, NY: Springer), 173–194.

Wu, F., Kaiser, E., and Anderson, E. (2012). Animacy effects in Chinese relative clause processing. *Lang. Cogn. Process.* 27, 1489–1524. doi: 10.1080/01690965.2011.614423

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2015 Lin. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 16 November 2015
doi: 10.3389/fpsyg.2015.01739

Filling Predictable and Unpredictable Gaps, with and without Similarity-Based Interference: Evidence for LIFG Effects of Dependency Processing

Kimberly Leiken[1,2], Brian McElree[3] and Liina Pylkkänen[1,3,4]*

[1] Department of Linguistics, New York University, New York, NY, USA, [2] Division of Neurology, MEG Center, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA, [3] Department of Psychology, New York University, New York, NY, USA, [4] NYUAD Institute, New York University Abu Dhabi, Abu Dhabi, UAE

OPEN ACCESS

Edited by:
Colin Phillips, University of Maryland, USA

Reviewed by:
Robert Fiorentino, University of Kansas, USA
Andrea Santi, University College London, UK

***Correspondence:**
Kimberly Leiken
kimberly.leiken@cchmc.org

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: *22 June 2015*
Accepted: *29 October 2015*
Published: *16 November 2015*

Citation:
Leiken K, McElree B and Pylkkänen L (2015) Filling Predictable and Unpredictable Gaps, with and without Similarity-Based Interference: Evidence for LIFG Effects of Dependency Processing. Front. Psychol. 6:1739. doi: 10.3389/fpsyg.2015.01739

One of the most replicated findings in neurolinguistic literature on syntax is the increase of hemodynamic activity in the left inferior frontal gyrus (LIFG) in response to object relative (OR) clauses compared to subject relative clauses. However, behavioral studies have shown that ORs are primarily only costly when similarity-based interference is involved and recently, Leiken and Pylkkänen (2014) showed with magnetoencephalography (MEG) that an LIFG increase at an OR gap is also dependent on such interference. However, since ORs always involve a cue indicating an upcoming dependency formation, OR dependencies could be processed already prior to the gap-site and thus show no sheer dependency effects at the gap itself. To investigate the role of gap predictability in LIFG dependency effects, this MEG study compared ORs to verb phrase ellipsis (VPE), which was used as an example of a non-predictable dependency. Additionally, we explored LIFG sensitivity to filler-gap order by including right node raising structures, in which the order of filler and gap is reverse to that of ORs and VPE. Half of the stimuli invoked similarity-based interference and half did not. Our results demonstrate that LIFG effects of dependency can be elicited regardless of whether the dependency is predictable, the stimulus materials evoke similarity-based interference, or the filler precedes the gap. Thus, contrary to our own prior data, the current findings suggest a highly general role for the LIFG in dependency interpretation that is not limited to environments involving similarity-based interference. Additionally, the millisecond time-resolution of MEG allowed for a detailed characterization of the temporal profiles of LIFG dependency effects across our three constructions, revealing that the timing of these effects is somewhat construction-specific.

Keywords: neurolinguistics, left inferior frontal gyrus, magnetoencephalography, Filler-gap dependency, object relative clause, verb phrase ellipsis, right node raising, similarity-based interference

INTRODUCTION

A classic finding within the cognitive neuroscience of language processing is that the comprehension of object relative (OR) clauses, such as (1), have been found to engender more hemodynamic activity in the left inferior frontal gyrus (LIFG, aka "Broca's Area") than subject relative (SR) clauses, such as (2) (e.g., Just et al., 1996; Stromswold et al., 1996; Caplan et al., 2000, 2008; Keller et al., 2001; Constable et al., 2004). This observed difference in hemodynamic activity mirrors the behavioral findings that ORs are more costly to process than SRs by various measures (Holmes, 1973; Hakes et al., 1976; Wanner and Maratsos, 1978; Holmes and O'Regan, 1981; Ford, 1983; Waters et al., 1987; King and Just, 1991).

(1) the fireman [(who$_i$) the deputy called $\underline{t_i}$] saved the sailor
(2) the fireman [who$_i$ $\underline{t_i}$ called the deputy] saved the sailor

In tandem with hypotheses developed from an older body of aphasia studies, this effect has given rise to the popular conception that Broca's area is somehow linked to syntactic processing (Berndt and Caramazza, 1980; Damasio and Damasio, 1989; Zurif, 1995; Grodzinsky, 2000). Of specific proposals, the narrowest in terms of LIFG function hypothesizes that this region is specifically responsible for the processing of displacement or movement (Grodzinsky, 2000; Ben-Shachar et al., 2003; Grodzinsky and Santi, 2008). It is important to note that these theories have primarily been tested using relative clause structures, such as those described above, which contain movement operations (or "transformations") that result in a long-distance dependency between two elements. Therefore, it is unclear whether it is the movement process itself or the consequential relation between non-adjacent components that induces the increase in activation. A more general set of hypotheses, but still specific to linguistic processing, includes the "linearization" computation (Bornkessel et al., 2005; Grewe et al., 2005), and the process of "unification" (Hagoort, 2003, 2005). Linearization involves maintaining hierarchical orderings of the members of a linguistic dependency. If this process takes place in the LIFG, then a violation of linguistic hierarchy should yield increased LIFG activity. Therefore, as English-type languages show a preference for subjects to precede objects, the LIFG effect for ORs could be taken to reflect the reversal of the subject–object order. Unification, on the other hand, is the process of integrating lexical information from a single word into a larger syntactic frame that has been retrieved from memory. Therefore, if this computation takes place in the LIFG, then integration of individual lexical items into the OR syntactic frame retrieved from memory might generate increased LIFG activity.

The above proposals contrast with hypotheses linking the LIFG primarily to non-language-specific processes, such as working memory (Caplan et al., 2000, 2008; Fiebach et al., 2001, 2005; Kaan and Swaab, 2002; Rogalsky et al., 2008; Makuuchi et al., 2009) or cognitive control (Botvinick et al., 2001; Miller and Cohen, 2001; Novick et al., 2005; Braver et al., 2007). Under both of these types of accounts, the LIFG increase relates not to any language-specific structural operation, but rather to the fact that, in ORs (but not in SRs), two noun phrases (NPs) are encountered prior to the verb, taxing working memory and/or inducing conflict.

The goal of the present work was to contribute to our understanding of LIFG function in language processing by examining LIFG dependency effects with a different methodology and a broader range of dependencies than previously studied, as well as by manipulating variables that affect memory retrieval operations in resolving a dependency. We employed magnetoencephalography (MEG), which, contrary to the traditional hemodynamic methods, allowed for a detailed temporal characterization of LIFG activity. Our design systematically varied not only the presence of dependency structures, but also the extent to which dependent structures elicited retrieval interference. In addition to commonly studied object extractions, we also explored dependencies resulting from verb-phrase ellipsis and right-node raising. Contrasting these constructions with ORs allowed us to narrow down the hypothesis space regarding the source of LIFG dependency effects. In sum, the central aim of this work was to assess whether dependency effects in the LIFG are only elicited for memory-intensive structures involving similarity-based interference or also for "easy" dependencies without much interference. The latter finding would conform to accounts implicating the LIFG for dependency resulting from movement operations (or long-distance dependency itself) whereas the emergence of LIFG effects only in the presence of interference would suggest a more memory-driven role. Note though that the interpretation of "movement" always involves retrieval whether or not the movement configuration places any extra burden on working memory. Thus a uniform effect of "movement" on the LIFG could reflect a generic role in retrieval, as opposed to a (language-) specific one in movement.

Retrieval Interference in Behavior and the LIFG

Enhanced LIFG activity for ORs as compared with SRs aligns with behavioral effects of increased processing time for ORs over SRs. However, recent behavioral studies have suggested that the retrieval process in ORs may only be more costly than that of SRs under conditions that engender retrieval interference (Gordon et al., 2001, 2002; Fedorenko et al., 2006; Van Dyke and McElree, 2006, 2011; Lee et al., 2007; Hofmeister, 2011). Evidence suggests that sentence comprehension relies upon a cue-driven, direct-access operation (e.g., McElree, 2000; McElree et al., 2003; Martin and McElree, 2008, 2009, 2011), in which cues formed at the retrieval site make contact with representations in memory that have matching content, without the need for a search process. Direct access is performed quickly, but can be highly susceptible to interference (Foraker and McElree, 2011). Basic memory research, as well as research on the role of memory in comprehension, indicates that the primary locus of interference occurs during retrieval (Van Dyke and McElree, 2006). Retrieval interference can result from "cue-overload," a

condition where retrieval cues are not distinctive enough to reliably elicit a desired target because they were associated with multiple items in memory (Watkins and Watkins, 1975; Nairne, 2002; Öztekin and McElree, 2007). In these circumstances, a sought-after element in memory may fail to be recovered, another element matching the retrieval cues may be recovered in its place, or "blend errors" may occur where two or more representations are "synthesized at retrieval" (Nystrom and McClelland, 1992).

It is natural to expect retrieval interference to be a key determinant of whether comprehension is successful. One type of retrieval interference that may impede the processing of OR dependencies is similarity-based interference: i.e., when two adjacent or nearby determiner-noun sequences are parallel in their surface syntax (Lewis, 1996). In fact, it has been shown that if the two NPs prior to the verb contrast in their surface structure, then the behavioral OR over SR effect diminishes (Gordon et al., 2001). This suggests that the processing delay is unrelated to the syntactic configuration of ORs. While a full spectrum of features that engender interference remains to be determined, interference has been recurrently observed when memory representations (i) overlap in their semantic category membership (Gardiner et al., 1972; Wickens, 1973; Dillon and Bittner, 1975; Watkins and Watkins, 1975; Crowder, 1976); (ii) have similar phonological forms (Haber and Haber, 1982; McCutchen et al., 1991; Acheson and Macdonald, 2011) or (iii) encode similar syntactic structures (Lewis, 1996; Gordon et al., 2001, 2002; Fedorenko et al., 2006; Van Dyke and McElree, 2006; Lee et al., 2007; Hofmeister, 2011).

Similarity-based interference effects have also been reported in neuroimaging studies. Increased activation in the LIFG has been associated with competition resolution in non-language-specific tasks (Thompson-Schill et al., 1997; Brandon et al., 2003, 2004; Derrfuss et al., 2004; Postle et al., 2004; Feredoes et al., 2006), and patients with lesions in this region have shown deficits when performing non-syntactic interference tasks (Costello and Warrington, 1989; Robinson et al., 1998; Thothathiri et al., 2010). This suggests that Broca's area should not merely be linked with syntactic processing, but instead plays a key role in resolving more domain-general interference.

In a previous study, we employed the time course sensitive technique of MEG to link the behavioral finding that OR effects depend on structural parallelism to the LIFG literature. Specifically, we investigated whether the LIFG effect would also be reduced when structural similarity between pre-verbal NPs is removed (Leiken and Pylkkänen, 2014). Our findings indicated that this was indeed the case; LIFG effects of similarity-based interference—but not the pure presence of a dependency—were found at the gap site in ORs. Thus, it was shown that MEG could indeed be employed for the study of object extraction, revealing effects at the gap site around 600 ms after verb onset; a time window consistent with the time course of EEG findings for dependency formation (King and Kutas, 1995; Kaan et al., 2000; Gouvea et al., 2010). Moreover, these results were consistent with working memory and/or conflict resolution-based hypotheses of the role of the LIFG, as opposed to purely syntactic accounts.

Three Constructions: Object Relatives, Verb Phrase Ellipsis, and Right Node Raising

Object Relative Clauses

In the current work, we engaged in a more large-scale investigation of the relationship between dependency formation and similarity-based interference. While the findings from our previous study—an LIFG increase only for ORs containing competing determiner-NPs—are consistent with a similarity-based retrieval interference account of LIFG activity (Öztekin et al., 2008, 2009), they do not yet conclusively rule out movement theories that LIFG activity increases for materials that contain a dependency resulting from a movement operation (Grodzinsky, 1986, 2000; Ben-Shachar et al., 2003, 2004; Grodzinsky and Friederici, 2006; Santi and Grodzinsky, 2007; Grodzinsky and Santi, 2008). Because of the predictable nature of an OR dependency, the lack of a gap-site effect in the LIFG for ORs without determiner-noun competitors may be due to the fact that dependency processing could primarily take place before the verb, prior to the gap. In fact, prior ERP studies have not only revealed P600 effects following the verb, but also a sustained anterior negativity at the point of the filler item in ORs prior to the verb (Phillips et al., 2005). This result could suggest that the bulk of dependency processing may occur in an anticipatory time window, preceding the completion of the gap-filling computation. Thus, one goal of the present study was to investigate whether the predictability of OR dependencies yields early effects in the LIFG, consistent with the dependency effects found in studies of movement theories, or whether ORs truly only elicit LIFG increases as a result of similarity-based interference. To address this question, the present MEG study (i) analyzed LIFG activity in earlier time windows prior to the gap site and (ii) compared LIFG activity elicited by ORs to LIFG activity elicited by non-predictable dependencies. With regards to the analysis of pre-gap LIFG activity, we employed OR clauses inside of sentence structures such as (3) to allow for more natural stimuli than in Leiken and Pylkkänen (2014).

(3) The husband hogged *the blankets that Jane grabbed* afterward.

The relative pronoun *that* may act as a signal that the object *the blankets* will be employed later on in a gap-filling dependency. It is possible that there are initial steps involved in computing a dependency, which may be able to be completed as soon as the filler item is recognized, even prior to the detection of a gap. Thus, once the gap is encountered, a sufficient portion of the processing sequence has been completed such that significant LIFG increases at the gap site would not be found. As this may have been the case in our previous study, in addition to measuring LIFG activity following the target verb *grabbed* prior to the gap site, the present study also analyzed activity in the earlier time window, following *that*, where predictive processing of the upcoming gap may take place.

Notice that the item of retrieval is a determiner-noun *the blankets*, which is different in its surface structure from the nearby proper name *Jane*. In order to investigate whether LIFG

activity results from similarity-based interference, we included OR conditions which, as in the previous study, allow for potential similarity-based inference between these two phrases by replacing *Jane* with a second determiner-NP (e.g., *the wife*).

To investigate whether LIFG activity, specifically at the gap site, results from a gap-filling process, non-gap-filling dependency constructions, such as (4), were employed as controls.

(4) The husband hogged the blankets and Jane grabbed *them* afterward.

Here, as there is no movement, there is no gap-filling dependency. However, it is important to note that the pronoun *them*, which replaces the gap in the OR construction, also forms a dependency with *the blankets*. Therefore, we might expect that a similar retrieval process occurs between retrieval at a pronoun and retrieval at an OR gap, and thus a comparison between condition (3) and condition (4) would yield little difference in LIFG activity. However, as we are using MEG, we will have the time course sensitivity to target activity immediately following the word *grabbed* in both conditions. We expect retrieval to be taking place in ORs during this time window, but later at *them* in controls. Furthermore, the results from Santi and Grodzinsky (2007) suggest that gap-filling shows a cost in the LIFG that retrieval at a pronoun does not. Therefore, we might expect increases for ORs in the LIFG over controls.

Verb Phrase Ellipsis

For the comparison of LIFG activity at the gap site in predictable ORs to the gap site of a non-predictable dependency, we employed a gap-filling dependency that does not contain a relative pronoun-like cue to the upcoming dependency; namely, verb phrase ellipsis (VPE). VPE is a two-clause construction that contains an overt verb phrase in the first clause, which, in the second clause, is interpreted, but replaced by an auxiliary verb. For example in (5), the dependency is between the overt verb phrase *called a cab* in the first clause and the gap resulting from ellipsis of the VP in the second clause.

(5) The pedestrian [called a cab]$_i$, and the bellhop did $\underline{t}_i$ too (Martin and McElree, 2008).

Like the retrieval of the filler at the gap site in ORs, the first-clause VP is retrieved later in the sentence. However, unlike ORs, VPE has no grammatical marking, like a relative pronoun, signaling that the verb phrase in the antecedent has a further role downstream (Martin and McElree, 2008).Without an early indication of a dependency, we expect the bulk of dependency-related neural activity to obligatorily take place after the ellipsis cue has been encountered. The present study, therefore, employs VPE constructions like (6), which will be analyzed following the ellipsis site, at *too*, in comparison with the OR gap site.

(6) The husband *hogged the blankets* and Jane did too.

If LIFG activity in response to gap-filling is reflective of a predictive process, then we expect ORs to show LIFG increases prior to the gap site, whereas we expect VPE to show LIFG increases following the ellipsis site.

On the other hand, LIFG activity for gap-filling constructions has previously been attributed to similarity-based interference. Therefore, to test this hypothesis this study included VPE conditions which contain a competitor for the VP item of retrieval. While a large literature exists for similarity-based interference in ORs, there is little precedent for what might induce this type of interference in VPE. Therefore, in order to introduce a rival VP, the present study interpolated an inner relative clause within VPE constructions, as in (7):

(7) The husband *hogged the blankets* and the wife who sometimes *nagged him* did too.

In this construction, the inner relative clause, *the wife who sometimes nagged him*, involves a VP *nagged him*, which may compete with *hogged the blankets* during retrieval at *too*. For consistency, similarity-based interference conditions of ORs also included these inner relative clauses. It is important to note that this yields parallel OR materials where one of the parallel NPs will contain an inner relative clause, while the other does not. This could potentially lower the similarity between these phrases, thus biasing against possible similarity-based interference effects in these conditions over non-parallel conditions.

According to Martin and McElree (2008), the information inside of the ellipsis site is not necessarily a structurally identical copy of the antecedent VP, as previously suggested (Frazier and Clifton, 2001). Instead, there was evidence that ellipsis may be interpreted using direct-access content cues. In this case, working memory will use a "pointer" mechanism to access the information in the antecedent. Therefore, measurements at *too* should essentially indicate the cost in the LIFG of this "pointing" mechanism. For comparison with a non-ellipsis construction, the present study will include controls, such as (8):

(8) The husband hogged the blankets and Jane did *that* too.

Note that this condition involves the pronoun *that* prior to the retrieval site. This pronoun forms a dependency with the VP from the first clause *hogged the blankets*. This pointing back to the antecedent VP is very similar to that in ellipsis. However, in VPE conditions the pointing is taking place at *too*, whereas in the controls this retrieval has already been completed at *that*. Therefore, a comparison at *too* might show increases for the ellipsis pointer mechanism over the control condition.

Right Node Raising

Object relatives and VPE not only differ in terms of their predictability, but also in terms of the syntactic category of their item of retrieval. That is, whereas ORs involve retrieval of an object or individual, VPE involves retrieval of a verb phrase. Therefore, any differences found in LIFG activity between these two constructions may not necessarily be attributable to predictability differences, but may be reflective of the difference in item retrieval. To control for this potential confound, we included a third predictable gap-filling construction, which also involved a dependency between a gap and verbal element; right node raising (RNR). There are several competing accounts of RNR,

including those that liken it to ORs[1] and others that associate it with VPE.[2] At present, however, it is a rather understudied construction, particularly in terms of how it is processed. Thus, we made a number of assumptions regarding the processing of RNR constructions.

(9) The husband *hogged* and Jane grabbed the pillows.

In (9), the verb in the first clause, *hogged*, requires an object, and the conjunction *and* indicates that what will follow will either be a VP-conjunct for *hogged*, or a larger DP-VP clause parallel to the one already presented. Therefore, when the DP, *Jane*, is encountered, it may lead to the expectation for the upcoming VP-conjunct, *grabbed the pillows*, which is parallel to the verb-gap phrase in the first clause. As a result, the object of the VP-conjunct in the second clause, *the pillows*, may be shared by the VP in the first clause. Under this set of assumptions, the item retrieved at the filler item, *the pillows*, is the verb-gap phrase *hogged*. This type of retrieval links RNR with VPE, which both have verbal items of retrieval, in contrast with ORs. It's important to note, however, that RNR is like ORs in terms of another property: predictability. If, as described above, the conjunction *and* followed by a DP acts as a cue to the upcoming VP, this would suggest that it is possible to begin processing the upcoming gap-filling computation prior to encountering the filler. Therefore, RNR will be included in the present study as a control for potential confounds, as it equates to VPE in terms of item of retrieval. Because of this shared property, in addition to the lack of existing literature on RNR processing, introducing the potential for similarity-based interference will be done in the same manner as VPE. That is, an inner relative clause will be interpolated, including a VP competitor for the item of retrieval, as in (10):

(10) The husband *hogged* and the wife who sometimes *nagged him* grabbed the pillows.

RNR will be analyzed at the retrieval site, *the pillows*, to examine similarity to VPE. Additionally, LIFG activity in the predictive region, *the wife*, will be analyzed. If RNR constructions are similarly predictable to ORs, then we might expect the bulk of LIFG activity to take place prior the filler item. RNR is also unique in that the gap precedes the filler, a configuration that is novel to the neurolinguistic literature. This reverse ordering of dependent elements might indicate a potential difference in the neural response between gap-filler RNR and filler-gap constructions.

In sum, using the temporal resolution of MEG, our aim was to assess to what extent the LIFG effect of dependency formation is modulated by predictability and/or syntactic similarity, in order to adjudicate between the multiple competing accounts of LIFG involvement in long-distance dependencies. Specifically, if the LIFG does not participate in dependency formation operations *per se*, but rather domain-general operations involved in retrieval and/or competition resolution, then LIFG effects should be modulated by similarity-based interference; specifically, conditions with the potential for high similarity-based interference should show strong LIFG effects. Additionally, if the absence of dependency effects in ORs without high similarity in Leiken and Pylkkänen (2014) was due to pre-gap predictive processing, then we would expect LIFG effects following the relative pronoun cue *that* in ORs. An unpredictable filler-gap construction, like VPE, which does not contain a cue to the upcoming dependency, would not allow for such predictive processing as in ORs. Therefore, we would expect LIFG effects to be delayed in VPEs until the late indication that ellipsis has taken place. Finally, RNRs, which enable prediction—albeit of a filler, rather than a gap—should pattern with ORs in allowing for pre-dependency LIFG effects. On the other hand, any similarity between RNR and VPE constructions (in contrast with ORs) would likely reflect retrieval of a verbal element as opposed to retrieval of an object. Unlike hemodynamic techniques, MEG provides the millisecond-by-millisecond temporal accuracy to attribute effects to specific portions of a trial. Thus, we can with confidence assess whether these effects are predictive of the upcoming gap, or result from encountering the gap.

Finally, it should be noted that although our region of interest (ROI) will simply be referred to as the LIFG, it is nowadays well-known that the LIFG (or "Broca's Area") is in fact a grouping of sub-regions with heterogeneous functionality, consisting of at least three Brodmann's areas (44, 45, and 47) and potentially further subdividing into multiple smaller regions according to evidence from multiple receptor mapping (Amunts et al., 2010). While both areas 44 and 45 have been implicated in sentence processing involving syntactic interference (e.g., Stowe et al., 1999; Cooke et al., 2002; Fiebach et al., 2004; Makuuchi et al., 2009), some linguistic competition tasks have distinguished between the two subparts, affecting only the pars triangularis/BA 45 (Gough et al., 2005; Guo et al., 2010) or only the pars opercularis/BA 44 (Mead et al., 2002; Gough et al., 2005). Crucially, MEG is unlikely to be able to distinguish between areas 44 and 45, and thus these two areas have been collapsed into a single region in our analysis. Therefore, our results will not inform any possible functional subdivision among these regions.

MATERIALS AND METHODS

Participants

Twenty-two right-handed native English speakers participated in the study (13 female; average age: 24.95 years). All had normal or corrected-to-normal vision. The study was formally approved by the New York University Institutional Review Board and all participants gave written informed consent.

[1]The gap-filling dependency of RNR is considered similar to ORs, in that it may involve the extraction of an element, resulting in a link between its overt position and its interpretation in the original location. While both ORs and RNR might involve extraction to a higher location, the extracted item in an OR undergoes leftward movement, whereas the extracted element in RNR would be moving rightward (Ross, 1968; Bresnan, 1974; Postal, 1974, 1998; Sabbagh, 2007). Therefore, the gap left behind in ORs follows the extracted item, whereas the gap left behind in RNR *precedes* the extracted item, as in the below example:

> Some people love $[t]_i$, but other people hate $[t]_i$, [the role that government plays in this country]$_i$. (Postal, 1974).

[2]On the other hand, non-movement analyses of RNR suggest that the shared object is simply an overt argument of the second clause, but deleted due to identity in the first clause. This representation of RNR would be more analogous to ellipsis accounts, which, as explained above, delete the VP in the second clause due to redundancy.

TABLE 1 | Experimental design with the critical items of MEG analysis bolded.

OR	Non-parallel	DEP	The husband hogged the blankets that Jane **grabbed** afterward
		Control	The husband hogged the blankets and Jane **grabbed** them afterward
	Parallel	DEP	The husband hogged the blankets that the wife who sometimes nagged him **grabbed** afterward
		Control	The husband hogged the blankets and the wife who sometimes nagged him **grabbed** them afterward
VPE	Non- parallel	DEP	The husband hogged the blankets and Jane did **too**
		Control	The husband hogged the blankets and Jane did that **too**
	Parallel	DEP	The husband hogged the blankets and the wife who sometimes nagged him did **too**
		Control	The husband hogged the blankets and the wife who sometimes nagged him did that **too**
RNR	Non- parallel	DEP	The husband hogged and Jane grabbed **the pillows**
		Control	The husband hogged the sheets and Jane grabbed **the pillows**
	Parallel	DEP	The husband hogged and the wife who sometimes nagged him grabbed **the pillows**
		Control	The husband hogged the sheets and the wife who sometimes nagged him grabbed **the pillows**

Stimuli and Task

As shown in **Table 1**, three different construction types were investigated: ORs (e.g., *The husband hogged the blankets that Jane grabbed afterward*); VPE (e.g., *The husband hogged the blankets and Jane did too*); and RNR (e.g., *The husband hogged and Jane grabbed the pillows*). Each type has two forms; one involving nearby elements that are parallel in their surface syntax to induce similarity-based interference ("par"), and one that contains elements which differ in their surface syntax ("nonpar"). Parallel types contain an interfering element designed to compete with the element being retrieved at a "gap" site. That is, a parallel NP in ORs, a parallel verb in RNR, and a parallel verb phrase in VPE. Sixty proper names (e.g., *Jane*), one for each set of non-parallel conditions, were employed. These names were replaced by a determiner-NP in parallel conditions (e.g., *the wife*). A non-dependency counterpart of each type was also included. This yielded a 2 × 2 design with similarity and dependency as factors within each construction type.

The specific conditions included: (i) sentences containing VPE (ellipsis-par; ellipsis-nonpar), (ii) VPE-controls containing *that* instead of ellipsis to point to the antecedent (ellipsis-control-par, ellipsis-control-nonpar), (iii) sentences containing ORs (OR-par, OR-nonpar), (iv) OR-controls containing a conjunction instead of a complementizer (OR-control-par; OR-control-nonpar), (v) sentences containing RNR (RNR-par; RNR-nonpar), (vi) RNR-controls in which an NP was inserted in the gap site resulting in a basic conjunction (RNR-par; RNR-nonpar), and (vii) filler sentences without syntactic dependencies for variability (filler-par; filler-nonpar). Each condition consisted of 60 trials, so altogether, each participant viewed 840 trials. The targets of MEG analysis were the dependency formation sites themselves (gap site in ORs, ellipsis-site in VPE, and filler-site in RNR) and anticipatory regions.

Obligatorily transitive verbs were used in the first clause of all conditions to prevent interpretation of RNRs as a basic conjunction. "*Did*" (which has an auxiliary verb use) was not used in the second conjunct of ORs/RNR to prevent a VPE reading. The item of retrieval in all conditions (the "filler" in the filler-gap construction), was designed to always been an inanimate object. For example, *the blankets, the bikes,* etc. (A full list of materials containing animate objects can be found in the Appendix.) Some psycholinguistic work has suggested that inanimacy of an object of retrieval may reduce the object-over-SR clause effect (Traxler et al., 2005), as well as neuroimaging work (Chen et al., 2006). Therefore, because the filler item was inanimate across all our sentence types, any effect of this type should be equivalent across conditions. Further, it should be more difficult to find an effect of similarity-based interference for ORs in the event that inanimacy diminishes processing load of OR clauses.

To ensure that the complexity of our materials did not sacrifice plausibility, we collected plausibility judgments on our stimuli using the Amazon Mechanical Turk (AMT) interface prior to the MEG recording (see Appendix 1). The test stimuli described above were complemented with syntactically grammatical, but highly implausible stimuli for comparison, following designs which similarly compare grammatical, but complex, stimuli with implausible items, such as Pylkkänen et al. (2004). The implausible stimuli were constructed by switching the verb in the first clause in the test items with the verb in the inner relative clause on one-third of the sentences in each condition, resulting in expressions such as *the husband nagged the blankets and the wife who seldom hogged him did too.* We gathered demographic information from 150 participants. Participants were obligated to indicate whether or not they were a native speaker of English and were informed that they were only permitted to participate in the experiment one time. Any participants who violated either of these criteria were rejected from analysis as were any participants who did not fill in the demographic survey. Also, those who far exceeded or fell below the average amount of time taken to complete their list were rejected if extreme durations were accompanied by unreasonable data (e.g., the same response for every trial). Items were distributed among 10 randomized different lists, so each list was completed by 15 subjects. Turk users saw each item and selected a plausibility rating on a 0–7 Likert scale (0 = completely implausible). Participants' raw ratings were averaged over each condition.

A *t*-test comparing the stimuli to be included in the MEG recording with implausible fillers showed that experimental stimuli ($M = 5.1408$) were rated significantly higher ($p < 0.001$) than their implausible counterparts ($M = 1.9199$) on a 0–7 scale, suggesting that the sentences intended for the MEG study were considered generally plausible.

Unsurprisingly, our stimulus manipulation affected the plausibility ratings, with a 2 × 2 × 3 ANOVA on the critical

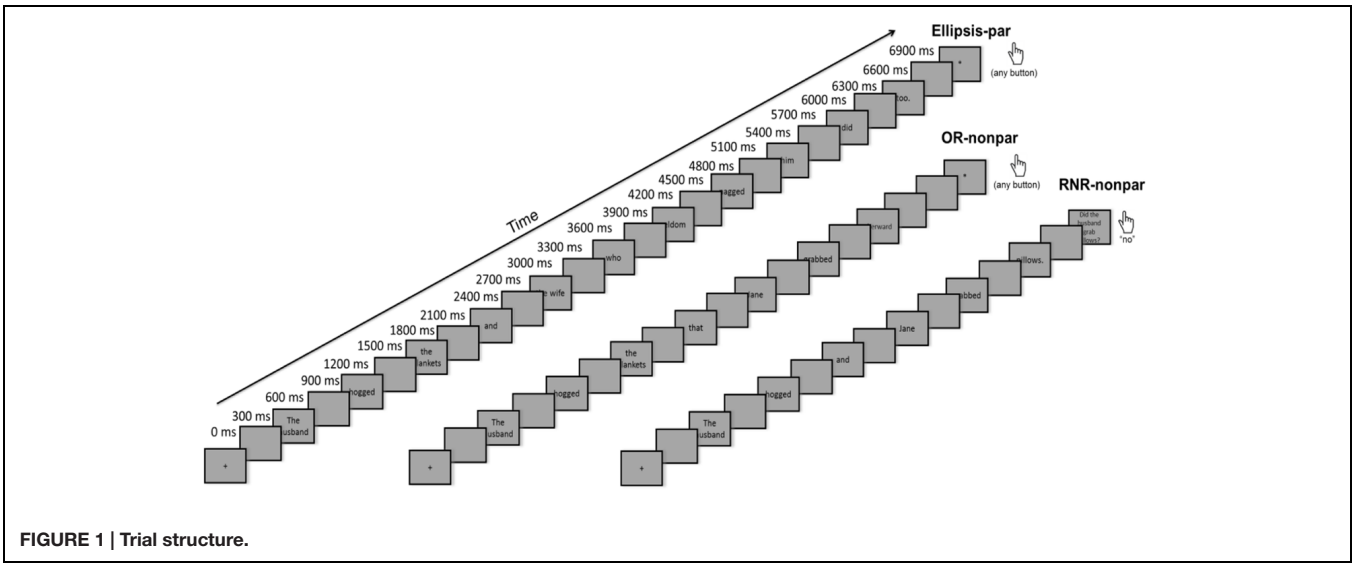


FIGURE 1 | Trial structure.

stimuli showing reliable main effects of all three factors. These effects were driven by lower ratings for parallel than non-parallel stimuli [Parallelism: $F(1,708) = 176.26$, $p < 0.001$; non-parallel $M = 5.51$, parallel $M = 4.81$], for dependency than control stimuli [Dependency $F(1,708) = 179.59$, $p < 0.001$, dependency $M = 4.80$, control $M = 5.51$], and by higher ratings for VPE than the other two constructions [Construction type: $F(2,708) = 41.858$, $p < 0.001$, VPE $M = 5.39$, OR $M = 5.25$, RNR $M = 4.83$].

The main effect of Parallelism was significant within each construction (all Fs > 25) though it was most robust within the OR-dependencies, as reflected by a reliable 2×2 interaction between Parallelism and Dependency within the ORs [$F(1,236) = 18.413$, $p < 0.001$, non-parallel control $M = 5.95$, parallel control $M = 5.29$, non-parallel dependency $M = 5.57$, parallel dependency $M = 4.21$], while no such interaction was observed within VPE [$F(1,236) = 1.998$, $p = 0.159$, non-parallel control $M = 5.45$, parallel control $M = 5.03$, non-parallel dependency $M = 5.87$, parallel dependency $M = 5.23$] or RNR [$F(1,236) = 0.072$, $p < 0.788$, non-parallel control $M = 5.97$, parallel control $M = 5.38$, non-parallel dependency $M = 4.25$, parallel dependency $M = 3.72$]. The three way interaction between Parallelism, Dependency and Construction was also significant [$F(2,708) = 4.43$, $p = 0.01$].

The main effect of Dependency was qualified by an interaction with Construction [$F(2,708) = 120.19$, $p < 0.001$], with reliably decreased ratings for dependency than control sentences for OR [$F(1,236) = 79.86$, $p < 0.001$; control $M = 5.62$, dependency $M = 4.89$] and RNR [$F(1,236) = 238.74$; $p < 0.001$, control $M = 5.67$, dependency $M = 3.98$], while the reverse held for VPE [$F(1,236) = 14.902$, $p < 0.001$; control $M = 5.24$, dependency $M = 5.55$].

In sum, parallelism uniformly decreased plausibility ratings, while the presence of a dependency decreased judgments for ORs and RNR but not for VPE. Thus any LIFG patterns tracking these effects could reflect plausibility instead of the independent variables of interest; we return to this in our report of the results.

During the MEG recordings, participants read all critical stimuli that were included in the MTurk study (with the exception of the implausible stimuli). Presentation was word-by-word (except in the case of determiner-NPs which were presented as a unit for time restriction purposes, e.g., *the wife*). After one-third of the linguistic stimuli, participants were presented with a comprehension question relating to the content of the previous text (e.g., *Did the husband grab the pillows?*) to which the answer was either "yes" or "no." For the purposes of this task, the participants were given practice outside the MEG machine and again inside the MEG machine prior to recording. Half of the questions had the answer "yes" and half had the answer "no." For a "yes," both the character and the action mentioned in the question needed to match those in the previous text.

Procedure

Before the MEG recordings, participants were instructed about the experimental task and their head shapes were digitized using a Polhemus (Colchester, VT, USA) FastSCAN COBRA 3D laser system. During the experiment, participants lay in a dimly lit, magnetically shielded room (Vacuumschmelze, Hanau, Germany). Using PsychToolbox, the experiment was presented on a 7x7-inch screen with a resolution of 1024×768 pixels placed approximately 9.5 inches above the subjects' eyes. Stimuli were presented word by word, 300 ms for each word, with a 300 ms blank screen between each word. To allow for longer processing time of complex stimuli, a blank screen was then presented for 700 ms prior to the question screen. Using a button press, the subject expressed whether the answer to the comprehension question was "yes" or "no" (**Figure 1**). Trial order was random. Subjects were in the machine for an hour, with five breaks (between each of the six blocks), and were then given an extended break outside of the MEG room, due to the length of the study. Subjects then returned to the machine for the next six blocks. The entire recording took about 2.5 h. MEG data were collected using a using a whole-head 157-channel axial gradiometer system (Kanazawa Institute of Technology, Nonoichi, Japan). For this

study, data were recorded at a sampling rate of 1000 Hz with a low-pass filter at 200 Hz using a DC recording and a notch filter at 60 Hz. Eye-blinks were recorded using an SR Research Eyelink 1000 Arm-Mounted Eyetracker sampling at 1000 Hz.

Data Analysis

Pre-processing of MEG Data

Raw data were noise-reduced (CALM; Adachi et al., 2001) and cleaned of artifacts (at a threshold of 4000fT). On average, no more than 25% of trials were lost during this procedure. Artifacts also included eye-blinks, which were removed by aligning the eye-tracking recording (described above) with the MEG recording. Data were high-pass filtered at 1 Hz. Data were then averaged by condition using a 200 ms pre-stimulus interval and a 1000 ms post-stimulus interval and baseline corrected using the 200 ms pre-stimulus interval. Data were low-pass filtered at 40 Hz after averaging, using the program BESA® 5.1 (MEGIS Software GmbH). Additionally, one subject was excluded as an outlier due to excessive blinking.

ROI Analysis of Minimum Norm Estimates

Magnetoencephalography data were analyzed as distributed sources using L2 minimum norm estimates calculated in BESA. The minimum norm images were depth weighted as well as spatiotemporally weighted, using a signal subspace correlation measure (Mosher and Leahy, 1998). LIFG activity at the site of dependency formation (OR: *grabbed*; VPE: *too*; RNR: *grabbed*) was examined via an ROI analysis. For the ROI analysis, sources were assigned to the anatomical LIFG region consisting of Brodmann's areas left 44 and 45, based on coordinates in Talairach space (Lancaster et al., 2000). Non-parametric, cluster-based permutation tests (Maris and Oostenveld, 2007) were performed in the same time windows as in Leiken and Pylkkänen (2014); an early "N400"-like time window (200–500 ms), associated with lexical access (Embick et al., 2001; Pylkkänen et al., 2002; Pylkkänen and Marantz, 2003), and basic combinatory effects (Bemis and Pylkkänen, 2011), and a late "P600" time window (500–800 ms) associated with OR versus SR P600 effects (Kaan et al., 2000). Additionally, due to the length and complexity of the current study's stimuli, a third, even later, analysis window was added (800–1000 ms). Permutation tests were employed to identify temporal clusters significantly affected by stimulus manipulation, corrected for multiple comparisons. Thresholds for initial cluster selection followed Leiken and Pylkkänen (2014), i.e., of waveform separations that lasted for 10 or more time points at $p < 0.3$, the one with the largest summed F or t statistic within each time-window was entered into 10,0000 permutations. The final corrected p-value for each cluster was calculated as the ratio of permutations yielding a test statistic greater than the actual observed test statistic ($\alpha = 0.05$). The tests were a 2 × 2 repeated measures ANOVA (Similarity × Dependency) over each time window ("N400," "P600," and late response) within each construction type at the site of dependency formation: at the onset of the verb preceding the gap in ORs, at the onset of the verb preceding the filler in RNR, and at the onset of the word following the ellipsis in VPE. The epochs were set to begin from the onset of each of these words through the following word. This extension allowed us to detect potential residual dependency effects which may have occurred early in the processing of the subsequent word. The ANOVA was then followed up with planned pairwise comparisons between parallel versus non-parallel subtypes, and dependency versus control subtypes. Effects at $p < 0.05$ will be discussed as significant and effects between this corrected level and $p < 0.10$ as marginal. Any p-values higher than this will be considered numerical trends. Our conclusions will, however, only rest on results reaching corrected significance at $p < 0.05$.

The above tests were followed by analyses at the pre-dependency time intervals (i.e., at the filler in ORs following the relative pronoun cue *that*, after the gap in RNR, and at a comparable lexical item in VPE). That is, analyses were performed in windows where the potential effects of dependency prediction may have taken place (i.e., prior to the gap in ORs, prior to the filler in RNR, and prior to the auxiliary verb in VPE). Unlike at the OR gap site, the lexical material in the predictive region was not matched in all four OR conditions (*that* in parallel and *and* non-parallel) Therefore, parallel and non-parallel ORs, along with their control counterparts could not be included in a single 2 × 2 ANOVA as above. Preverbal material was instead submitted to t-tests in order to examine potential anticipatory dependency processing in the same LIFG region. t-tests were performed after the presentation of the filler item: e.g., *the wife* in parallel conditions and *Jane* in non-parallels in the examples in **Table 1**. If effects of dependency prediction only occur in conjunction with similarity-based interference, then a difference would only be found between the two instances of *the wife* and not between the two instances of *Jane*. However, if anticipatory LIFG effects are independent of parallel syntactic structure, then both t-tests should show a difference. In RNR, the immediate post-gap lexical item *and* is the same in all four conditions, allowing for a 2 × 2 (similarity × dependency) ANOVA. This was then followed up by t-tests on the next word comparable to those performed on ORs; at *the wife* in parallel conditions, and *Jane* in non-parallel conditions. In VPE, no dependency distinction exists between the four conditions until after auxiliary verb. To confirm that no LIFG effect of dependency anticipation occurs in a time window prior to *did*, a 2 × 2 ANOVA at the conjunction *and* was performed. As in the ORs and RNR, t-tests were also performed within parallel conditions at *the wife* and within non-parallel conditions at *Jane*, to confirm the assumption that having no cue to an upcoming dependency prohibits dependency prediction. The t-tests employed the same settings as the above 2 × 2 ANOVAs; 10,000 permutations with the same cluster thresholds within the same three time intervals.

Full Brain Analyses

The ROI analyses were each supplemented by liberally thresholded uncorrected full brain contrasts. The goal of these analyses was to confirm that the effects found in the ROI analyses in fact reflected activity localized with the LIFG (as opposed to spillover from neighboring regions) and to reveal any other major cotemporaneous effects. We compared the minimum norm estimates of the activity elicited by the

experimental conditions sample-by-sample in the same pairwise comparisons described for the ROI analyses. Effects were visualized on the smooth BESA cortex when they remained reliable ($p < 0.05$, uncorrected) for at least three temporal samples and were observed in at least three spatially contiguous cortical sources.

RESULTS

Behavioral Data

After one-third of the sentences, participants were asked a comprehension question relating to the content of the stimulus sentence (e.g., Did the husband grab the pillows?) to which the answer was either "yes" or "no." Subjects performed fairly well on this complex task overall (M = 77.37%), and generally better (average accuracy ± SD) on the non-parallel (M = 82.75 ± 9.78%) than the parallel (M = 72.00 ± 8.12%) trials. In general, performance was slightly higher on control conditions (M = 77.60± 8.94%) compared with dependency conditions (M = 75.60± 9.28%). Performance was quite similar for the dependency version of each construction type: ORs (M = 76.89± 9.04%), VPE (M = 76.99± 9.20%), and RNR (M = 72.92± 9.61%).

MEG Data

Object Relative Clauses

As described above, only one of the two parallel NPs in parallel OR conditions contained an inner relative clause, potentially lessening similarity-based interference in these conditions. Nevertheless, our OR results showed a straightforward though late effect of parallelism after the gap-site, as well as a more complicated effect of dependency, as detailed below. No interactions between our two factors were observed. Test results are considered significant at $p < 0.05$, but for completeness in addressing our hypotheses we will report marginal results and numerical trends resulting from planned comparisons as well. Only significant findings will, however, contribute to our interpretation of results.

The early time window (200–500 ms) showed weak trends both toward a main effect of parallelism and for a main effect of dependency, but neither cluster survived the permutation correction for multiple comparisons (parallelism: $p = 0.1624$ at 397–500 ms; dependency: $p = 0.7253$ at 329–351 ms). The wave form separation during these non-significant main effects did, however, conform to the results of Leiken and Pylkkänen (2014), with only parallel dependencies eliciting increased amplitudes as compared to all other conditions.

No reliable effects were found in the 500–800 ms time window but the latest time-window, 800–1000 ms, showed both a reliable main effect of parallelism, with the cluster extending throughout this interval ($p = 0.0041$), as well as a reliable main effect of dependency, similarly covering the entire 800–1000 ms interval ($p = 0.0261$; **Figure 2**). These results reflected a pattern of parallel conditions eliciting increased LIFG amplitudes as compared to non-parallel conditions and dependency conditions eliciting increased amplitudes as compared to non-dependency controls. Planned pair-wise comparisons showed that within the dependency conditions, parallel ORs ($M = 6.16$) elicited significantly higher LIFG activity than non-parallel ORs ($M = 4.373$; 800–1000 ms, $p = 0.0087$) whereas within the control conditions, the increase for parallelism ($M = 4.504$) was only marginal (896–970 ms, $p = 0.0638$) versus non-parallel ORs ($M = 3.643$). The effect of dependency trended in the right direction for the parallel conditions (800–863 ms, $p = 0.1102$) but was significant for the non-parallel conditions (895–966 ms, $p = 0.0366$). Parallel ORs also elicited significantly higher LIFG activity than the non-parallel controls (800–1000 ms, $p = 0.001$). In sum, the pairwise comparisons showed increased activity for both dependency conditions over their controls and for both parallel conditions over their non-parallel versions.

However, before we can conclude that the LIFG ROI activity is modulated by the presence of a dependency, a complication arising from the lateness of this effect must be addressed. Namely, the effect occurred during the processing of the word immediately following the target verb; this word being *afterward* in the dependency conditions and *them* in the controls. Thus the LIFG increase could simply have reflected the increased activity for the longer and morphologically more complex *afterward* than *them*. However, since the word after *them* in the control conditions was *afterward* [the full contrast being *grabbed afterward* (OR) versus *grabbed them afterward* (control)], this lexically based explanation would predict that a comparison at *afterward* in the OR versus control conditions should not show the LIFG effect. In contrast, if the LIFG increase at *afterward* in the OR condition reflects dependency processing, it should replicate in a comparison of the two instances of *afterward*. To test this, the baseline was moved to 200 ms before the onset of *afterward* for all four conditions and 2 × 2 permutation ANOVAs were run in the 200–500 ms interval (covering the timing of the effect in the prior analysis) as well as in the 0–200 ms interval, covering any effects occurring at the very onset of this spill-over word. Though the ANOVA for 200–500 ms revealed a cluster replicating the pattern in the prior analysis (i.e., higher amplitudes for dependency than for control conditions), it did not survive correction for multiple comparison. However, in the earlier interval, 0–200 ms, a reliable main effect of dependency was observed (19–172 ms, $p = 0.0268$), with pairwise comparisons showing a significant increase for dependency ($M = 5.81$) over control ($M = 4.567$) within the non-parallel conditions (19–108 ms, $p = 0.0305$) and a similar trend for dependency ($M = 4.5.692$) over control ($M = 4.4.974$) within the parallel conditions (132–164 ms, $p = 0.1638$). These results converge on the finding that the presence of an OR dependency elicited a late LIFG increase occurring after 600 ms post-target verb onset. Interestingly, this effect was stronger for the non-parallel conditions, suggesting that it is not dependent on the presence of parallelism. This is in contrast to the findings of Leiken and Pylkkänen (2014), who for sentence fragments only found an LIFG increase for OR dependencies involving parallel NPs. Thus it is possible that the current full sentential stimuli may have been better test items for detecting a dependency effect in the absence of parallelism.

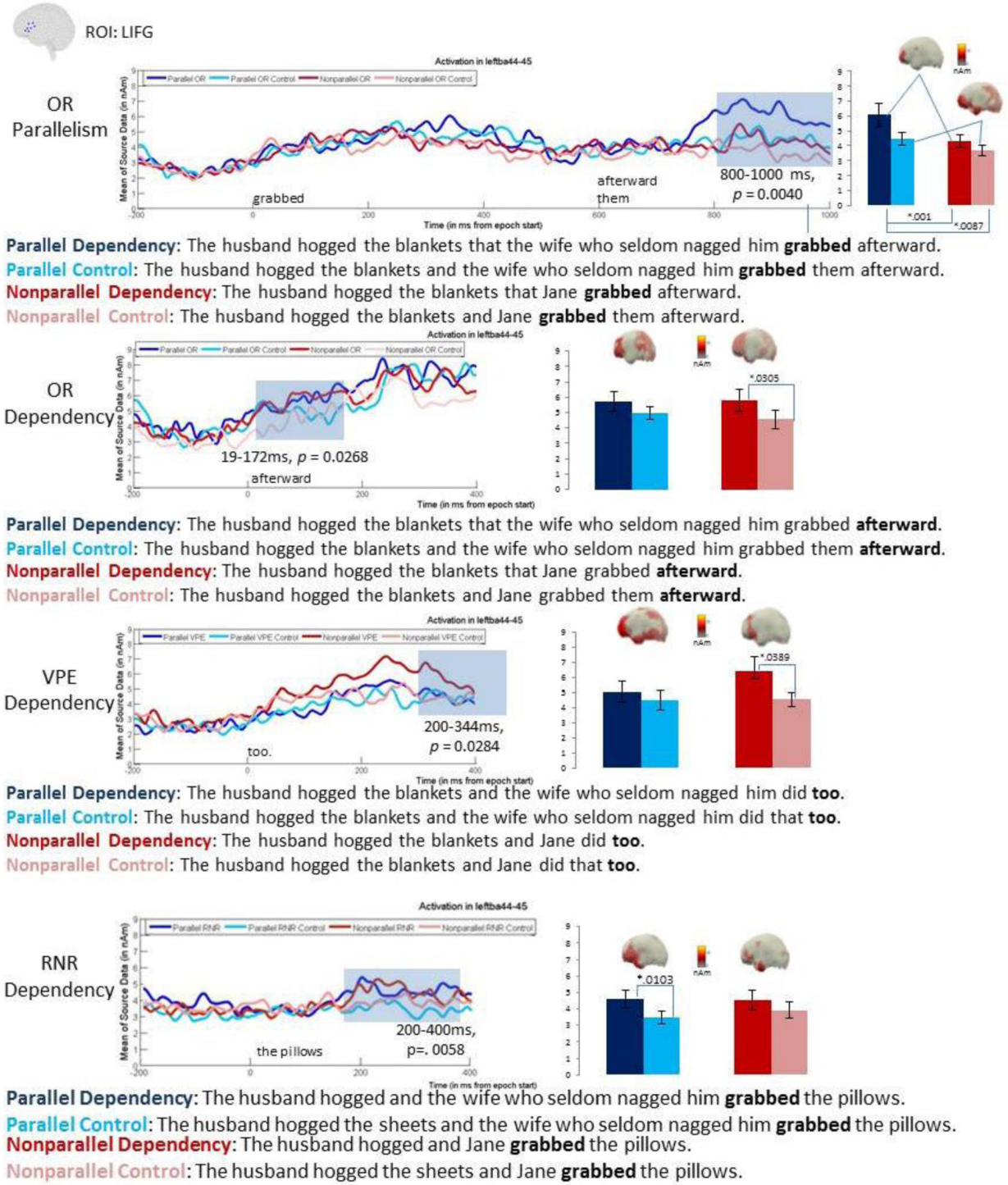


FIGURE 2 | Left inferior frontal gyrus (LIFG) ROI and Whole Brain results. The clusters of time points that were reliable in a cluster-based permutation *t*-test are shaded gray showing the LIFG increase for the dependency conditions as compared with control conditions. Red colors indicate non-parallel conditions, and blue colors indicate parallel conditions. Darker shades of each indicate dependency conditions, whereas the lighter shades stand for controls. Within 19–172 ms in the OR condition, there was a significant increase in the LIFG for OR dependencies versus control conditions ($p = 0.0268$). For the VPE condition, within 200–344 ms, there was a significant LIFG increase for VPE dependencies versus control conditions ($p = 0.0284$). The time window of 200–400 ms shows a significant LIFG increase for RNR ($p = 0.0058$). For each of the three constructions, an accompanying bar graph indicates the means for each condition (parallel dependency, non-parallel dependency, parallel control, non-parallel control) within the time window showing significant dependency increases in the LIFG. ROI findings were well-supported by full brain analyses, confirming LIFG increases within the time window showing significant clusters for each pairwise comparison.

Given that we did observe an effect of dependency after the gap-site, the prediction for such an effect in the predictive pre-gap region is weakened. In fact no such effects were observed when activity elicited by the filler items (*the wife* in parallels, and *Jane* in non-parallels) was compared in permutation *t*-tests. Thus our findings revealed no evidence for predictive gap-processing in the LIFG.

The whole brain graphs plot the same pair-wise comparisons as reported above on liberally thresholded whole brain minimum norms (time and space thresholds at 3 and *p*-value threshold at 0.05) at the time windows of the significant effects in the ROI analysis. The aim of this analysis was to ascertain that the ROI results in fact correspond to activity localized in the LIFG. These contrasts revealed activity overlapping with the BA 44–45 region during the time window of the parallelism main effect in ORs. Specifically, both parallel ORs and parallel control conditions showed an increase in this region compared with their non-parallel counterparts in the 800–1000 ms time window. The dependency effect early on after the presentation of *afterward* (19–172 ms after the onset) was also observable in the whole brain analyses for parallel ORs over parallel control conditions as well as for non-parallel ORs over non-parallel controls. In addition to left inferior frontal activity, posterior parieto-occipital activation was observed for the parallel control condition over non-parallel controls, as well as in both dependency contrasts.

Verb Phrase Ellipsis

For VPE, the cluster-based 2 × 2 ANOVA in the early time-window (200–500 ms) revealed a significant main effect of dependency at 200–344 ms ($p = 0.0284$). As with ORs, this effect was more strongly driven by a pair-wise effect in the non-parallel conditions (280–344 ms, $p = 0.0389$), with dependency ($M = 5.893$) over controls ($M = 4.330$). Parallel conditions showed a weaker trend ($p = 0.2123$ at 259–284 ms) of dependency ($M = 4.719$) over controls ($M = 4.178$). No reliable effect of parallelism was observed in this time-window nor any effects of either factor in the later time-windows. Finally, no effects were found in the pre-gap "predictive" time-windows (at *the wife* within parallels and at *Jane* within non-parallels), consistent with the fact that VPE dependencies are unpredictable.

Full brain analyses were also performed for the pairwise comparisons within the VPE conditions, specifically, contrasts between parallel VPE over parallel controls, and non-parallel VPE over non-parallel controls. These results confirmed to the ROI analyses in revealing LIFG effects within the time window of significant ROI findings. Namely, an effect was obtained at 200–344 ms for parallel VPEs over parallel controls, as well as for non-parallel VPEs over non-parallel controls. Again, the frontal effects were accompanied by more posterior activation in the parallel VPE condition over the parallel control condition, but not in the non-parallel contrast.

Right Node Raising

In the RNR analysis, the onset of the second verb (*grabbed* in **Table 1**) was treated as 0 ms, for consistency with the OR analyses. A reliable main effect of dependency was observed in the 800–1000 ms time window (or 200–400 ms following the RNR filler item, *the pillows*; $p = 0.0058$), with the cluster covering the entire analysis interval. In the pairwise comparison this effect was reliable within the parallels (267–400 ms, $p = 0.0103$), with dependency ($M = 4.600$) over controls ($M = 3.466$), and within the non-parallels, only trending in the same direction (299–332 ms, $p = 0.2251$) for dependency ($M = 4.528$) over controls ($M = 3.920$). No other effects were observed, including in the pre-gap "predictive" regions. Thus, like in VPE, parallelism did not appear to affect RNR processing in the LIFG. Timing wise, the RNR dependency effect occurred within 300 ms of encountering the site at which the dependency needs to be formed (which in RNR is the filler). This is similar to the dependency effect in VPE, suggesting that filler-gap order is not a strong modulator of LIFG dependency effects. This timing of course contrasts to the dependency effects observed for ORs, which were much later.

Whole brain pairwise comparisons were performed for the contrasts between parallel RNR and parallel controls, as well as between non-parallel RNR and non-parallel controls, with results conforming to the RNR ROI findings described above. That is, the time window of 200–400 ms after *the pillows*, which showed significant clusters of LIFG activity for parallel RNRs over parallel controls, also reveals significant effects in the full brain contrasts. Similarly, the increase of LIFG activity in non-parallel RNRs over non-parallel controls within 200–400 ms was evident from the full brain plots. For these contrasts, the effects were mostly anterior, though for both contrasts, the LIFG effect appears to be accompanied by an increase in activation in left anterior temporal cortex.

Results Summary

In sum, our results indicated an LIFG effect of Dependency for each construction type without interaction with Parallelism, suggesting that this effect is not dependent on similarity-based interference. In ORs, Parallelism had its own main effect, indicating that this factor can drive the LIFG even in the absence of a filler and gap. Importantly, neither effect tracked plausibility as rated in our MTurk norming study: parallelism lowered plausibility judgments across all constructions, but an MEG effect of Parallelism was only obtained for ORs; for Dependency, judgments were lower for dependency conditions in ORs and RNR but higher in VPE while in contrast, LIFG amplitudes increased for dependency conditions regardless of construction.

DISCUSSION

This study investigated LIFG activity during dependency processing using both a technique and constructions novel to the literature in order to shed light on the role of the LIFG in dependency formation. Our key question was whether dependency effects in the LIFG, whether the result of movement or not, require explicit taxing of working memory via similarity-based interference, or whether the sheer presence of a dependency is sufficient to drive this activity, as predicted by movement-based accounts of activity in this region. Our results show that similarity-based interference is a not a prerequisite for

LIFG effects: LIFG amplitudes showed a statistically significant increase when a dependency was present across our three constructions whether or not interference-inducing syntactic parallelism was built into the stimuli. Although sub-types of each construction contributed to the main dependency effect differently, we take the significant main effect within each construction type to show that for these materials, an activity increase was observed in the LIFG for any instance of retrieving the first member of a dependency chain. Thus, contrary to our own previous work, where we used sentence fragments as opposed to full sentences (Leiken and Pylkkänen, 2014), the current results support a role for the LIFG in dependency formation that generalizes across a variety of memory demands. The findings are compatible with the hypothesis that the LIFG computes syntactic movement, but also with the hypothesis that this region has a basic role in retrieval in a variety of non-movement contexts.

Given that Leiken and Pylkkänen (2014) found no purely dependency driven LIFG effects at an OR gap, an important goal for the current study was to investigate whether such effects could be observed in pre-gap LIFG activity, as potential reflections of gap anticipation. However, since here we did find an effect of dependency after OR gaps, the prediction for pre-gap dependency effects was somewhat weakened, and in fact, no anticipatory LIFG effects were observed for ORs or for either of the other construction types. Further, since significant LIFG increases for dependency were observed for ORs, specifically for the non-parallels, even though non-parallel control conditions also contained a type of retrieval at the pronoun *them*, these results conform to prior findings indicating that gap-filling may produce a greater LIFG cost than retrieval at a pronoun (Santi and Grodzinsky, 2007).

Left inferior frontal gyrus effects in OR dependencies were compared with VPE, which contains a non-predictable dependency and was thus anticipated to require the presentation of both filler and gap before the dependency could be processed. The VPE control condition, like that of ORs, also contained a type of non-gap-filling dependency between the pronoun *that* and the VP item of retrieval. Again, MEG time course sensitivity allows for fine-grained measurements at the post-gap word *too* in both conditions, where it was expected that retrieval takes place in VPE, but has already been completed in the control condition at *that*. Both of these expectations were supported, as VPE showed no anticipatory LIFG effects, but did show significant LIFG increases at the ellipsis-site.

While both OR and VPE constructions showed retrieval effects at the gap site in the LIFG, the timing of the effect was much earlier for VPE than that for ORs. While the OR results are compatible with the full time-course of gap-filling processes in previous SAT studies (McElree, 2001; McElree et al., 2003, 2006), their lateness with respect to VPE deserves some attention. Whereas the OR constructions involve retrieval of an object/individual, VPE involves retrieval of a verbal element. Thus, one possibility may be that the category of the retrieved item matters for retrieval time. Another possibility is that operations performed at the retrieval site differ for ORs and VPEs. Our RNR constructions bear on this issue, as they are predictable like ORs, but involve retrieval of a verbal element like VPE. A unique property of RNR constructions, however, is that they contain a gap-filling dependency where the gap precedes the filler, unlike in ORs or VPE. Despite this special property, RNR findings were closely linked with our VPE results. Specifically, RNR showed a significant increase for dependency in the LIFG. Therefore, we cannot attribute VPE-OR differences to predictability, as RNR is matched with ORs for this feature. Interestingly, we note that the similar processing profiles for VPEs and RNRs aligns with the theoretical proposal that RNRs are in fact a type of ellipsis (Wexler, 1980; Swingle, 1993; Kayne, 1994; Wilder, 1997; Hartmann, 2001, 2003; Abels, 2003; Ha, 2008). The slower timing effect for ORs could indicate that the word category of the item of retrieval affects retrieval speed, with access to verbal elements being faster than objects/individuals. Alternatively, and perhaps more plausibly, VPEs and RNRs could be processed more quickly than ORs because, as Martin and McElree (2008, 2009) argued, VPE can be resolved through a pointer mechanism, wherein retrieval consists of pointing to a structure in memory. On the other hand, processing ORs requires building the argument structure of the verb phrase after argument has been retrieved.

Taken together, the present set of findings can, in fact, be accounted for by hypotheses associating the LIFG with dependencies resulting from syntactic movement (Grodzinsky, 2000; Ben-Shachar et al., 2003; Grodzinsky and Santi, 2008), though the main effect of parallelism obtained for the ORs also provides evidence for the role of working memory independent of structure (Caplan et al., 2000, 2008; Fiebach et al., 2001, 2005; Kaan and Swaab, 2002; Rogalsky et al., 2008; Makuuchi et al., 2009) or cognitive control (Botvinick et al., 2001; Miller and Cohen, 2001; Novick et al., 2005; Braver et al., 2007). These results for parallelism are more convincing given that the item of retrieval in all conditions was an inanimate object. In other words, despite the fact that inanimacy has been associated with a reduction in OR processing load (Traxler et al., 2005), similarity-based interference effects were still found for parallel ORs versus non-parallels. Importantly, however, not only did these similarity-based interference effects appear rather late in the present study, at 800–1000 ms as opposed to at 300–400 ms in Leiken and Pylkkänen (2014), but they also only held for ORs, and not VPE or RNR. Regarding latency, it has been shown that the timing of P600 effects can be delayed when dependency length is increased (Phillips et al., 2005). Due to the full sentential stimuli of the present study, the distance between the gap and filler items was much greater than that in our previous study, where we employed minimal OR phrases. Therefore, it is possible that the later timing of similarity-based interference effects was due to the large amount of mediating material between filler and gap. The high complexity of the present study's materials may also be relevant for the fact that the parallelism effects were only found for ORs. That is, the ORs were contained inside of a complex sentential structure, as in (3), where the first constituent of the sentence, the matrix subject, (e.g., *the husband*)

was of the same determiner-noun structure as the item of retrieval (e.g., *the blankets*) and its competitor (e.g., *the wife*).

(1) *The husband* hogged *the blankets* that *the wife who sometimes nagged him* grabbed afterward.

This property may have induced *proactive* interference with the item of retrieval, where material prior to the initial encoding of the target item creates competition with it (Öztekin and McElree, 2007). The VPE and RNR stimuli did not have elements inducing possible proactive interference. This factor may also have contributed to the latency of the OR retrieval effect as compared with the other two conditions which showed similar effects in a much earlier time window.

In sum, our two main results are LIFG increases in response to similarity-based interference in ORs, and LIFG increases in response to the presence of the three different dependency types regardless of similarity-based interference. While both of these findings are attributed to "LIFG" activity, it is important to note that this region contains heterogeneous subparts. Therefore, it is possible that the interference effect is in one subdivision, whereas the effects for retrieval are in the other. The spatial resolution of MEG is, however, unlikely to be able to disambiguate the detailed localization of these effects within the LIFG and thus we must leave this question for future work.

CONCLUSION

This study took advantage of the detailed time-resolution of MEG and the stimulus properties of three different dependency constructions–ORs, VPE, and right-node-raising—to target several of the major competing accounts of the role of the LIFG in dependency processing. Our findings revealed that at the retrieval sites of each of these three dependencies, LIFG increases are observed, conforming to "movement" accounts. Additionally, in ORs only, effects of similarity-based interference were observed in the LIFG, consistent with working memory or cognitive control theories. Thus, our results add to the growing body of evidence that a complete understanding of "Broca's Area" must take into consideration both structure and memory related processes. Overall, our results are consistent with a hypothesis that the LIFG region subserves the recovery of an element from memory. The exact generality of this process across contexts remains a question for future work, but the current results enable a new level of temporal and computational precision in subsequent hypotheses about the type of retrieval that the LIFG contributes to.

ACKNOWLEDGMENTS

This research was supported by the National Science Foundation grant BCS-1221723 (LP) and grant G1001 from the NYUAD Institute, New York University Abu Dhabi (LP). We thank Jeffrey Walker, Miriam Lauter, and Rebecca Egbert for their assistance at various stages of this project.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpsyg.2015.01739

REFERENCES

Abels, K. (2003). "Right node raising: ellipsis or across the board movement," in *Proceedings of the North East Linguistic Society 34*, eds K. Moulton and M. Wolf (Amherst, MA: Graduate Linguistic Student Association).

Acheson, D. J., and Macdonald, M. C. (2011). The rhymes that the reader perused confused the meaning: phonological effects during on-line sentence comprehension. *J. Mem. Lang.* 65, 193–207. doi: 10.1016/j.jml.2011.04.006

Adachi, Y., Shimogawara, M., Higuchi, M., Haruta, Y., and Ochiai, M. (2001). Reduction of non-periodic environmental magnetic noise in MEG measurement by continuously adjusted least squares method. *IEEE Trans. Appl. Supercond.* 11, 669–672. doi: 10.1109/77.919433

Amunts, K., Lenzen, M., Friederici, A. D., Schleicher, A., Morosan, P., Palomero-Gallagher, N., et al. (2010). Broca's region: novel organizational principles and multiple receptor mapping. *PLoS Biol.* 8:e1000489. doi: 10.1371/journal.pbio.1000489

Bemis, D. K., and Pylkkänen, L. (2011). Simple composition: a magnetoencephalography investigation into the comprehension of minimal linguistic phrases. *J. Neurosci.* 31, 2801–2814. doi: 10.1523/JNEUROSCI.5003-10.2011

Ben-Shachar, M., Hendler, T., Kahn, I., Ben-Bashat, D., and Grodzinsky, Y. (2003). The neural reality of syntactic transformations: evidence from functional magnetic resonance imaging. *Psychol. Sci.* 14, 433–440. doi: 10.2307/40064164

Ben-Shachar, M., Palti, D., and Grodzinsky, Y. (2004). Neural correlates of syntactic movement: converging evidence from two fMRI experiments. *Neuroimage* 21, 1320–1336. doi: 10.1016/j.neuroimage.2003.11.027

Berndt, R. S., and Caramazza, A. (1980). A redefinition of the syndrome of broca's aphasia: implications for a neurological model of language. *Appl. Psychol.* 1, 225–278. doi: 10.1017/S0142716400000552

Bornkessel, I., Zysset, S., Friederici, A. D., von Cramon, D. Y., and Schlesewsky, M. (2005). Who did what to whom? The neural basis of argument hierarchies during language comprehension. *Neuroimage* 26, 221–233. doi: 10.1016/j.neuroimage.2005.01.032

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., and Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychol. Rev.* 108, 624–652. doi: 10.1037/0033-295X.108.3.624

Brandon, M., Hirshorn, E., Jha, A., Fabian, S., and Thompson-Schill, S. (2004). "Proactive interference resolution during nonverbal working memory: evidence for domain-general processing in LIFG," in *Poster Presented at the Cognitive Neuroscience Society Annual Meeting*, San Francisco, CA.

Brandon, M., Jha, A., Trueswell, J. C., Barde, L., and Thompson-Schill, S. (2003). "Proactive interference in verbal and non-verbal working memory," in *Poster Presented at the 44th Annual Meeting of Psychonomic Society*, Vancouver, BC.

Braver, T. S., Gray, J. R., and Burgess, G. C. (2007). "Explaining the many varieties of working memory variation: dual mechanisms of cognitive control," in *Variation in Working Memory*, eds A. Conway, C. Jarrold, M. Kane, A. Miyake, and J. Towse (New York, NY: Oxford University Press).

Bresnan, J. W. (1974). The position of certain clause-particles in phrase structure. *Linguist. Inq.* 5, 614–619. doi: 10.2307/4177846

Caplan, D., Alpert, N., Waters, G., and Olivieri, A. (2000). Activation of Broca's area by syntactic processing under conditions of concurrent articulation. *Hum. Brain Mapp.* 9, 65–71.

Caplan, D., Stanczak, L., and Waters, G. (2008). Syntactic and thematic constraint effects on blood oxygenation level dependent signal correlates of comprehension of relative clauses. *J. Cogn. Neurosci.* 20, 643–656. doi: 10.1162/jocn.2008.20044

Chen, E., West, W. C., Waters, G., and Caplan, D. (2006). Determinants of bold signal correlates of processing object-extracted relative clauses. *Cortex* 42, 591–604. doi: 10.1016/S0010-9452(08)70397-6

Constable, R. T., Pugh, K. R., Berroya, E., Mencl, W. E., Westerveld, M., Ni, W., et al. (2004). Sentence complexity and input modality effects in sentence comprehension: an fMRI study. *Neuroimage* 22, 11–21. doi: 10.1016/j.neuroimage.2004.01.001

Cooke, A., Zurif, E. B., DeVita, C., Alsop, D., Koenig, P., Detre, J., et al. (2002). Neural basis for sentence comprehension: grammatical and short-term memory components. *Hum. Brain Mapp.* 15, 80–94. doi: 10.1002/hbm.10006

Costello, A. L., and Warrington, E. K. (1989). Dynamic aphasia: the selective impairment of verbal planning. *Cortex* 25, 103–114. doi: 10.1016/S0010-9452(89)80010-3

Crowder, R. G. (1976). *Principles of Learning and Memory*. Mahwah, NJ: Lawrence Erlbaum Associates.

Damasio, H., and Damasio, A. R. (1989). *Lesion Analysis in Neuropsychology*. Oxford: Oxford University Press.

Derrfuss, J., Brass, M., and von Cramon, D. Y. (2004). Cognitive control in the posterior frontolateral cortex: evidence from common activations in task coordination, interference control, and working memory. *Neuroimage* 23, 604–612. doi: 10.1016/j.neuroimage.2004.06.007

Dillon, R. F., and Bittner, L. A. (1975). Analysis of retrieval cues and release from proactive inhibition. *J. Verbal Learn. Verbal Behav.* 14, 616–622. doi: 10.1016/S0022-5371(75)80048-X

Embick, D., Hackl, M., Schaeffer, J., Kelepir, M., and Marantz, A. (2001). A magnetoencephalographic component whose latency reflects lexical frequency. *Cogn. Brain Res.* 10, 345–348. doi: 10.1016/S0926-6410(00)00053-7

Fedorenko, E., Gibson, E., and Rohde, D. (2006). The nature of working memory capacity in sentence comprehension: evidence against domain-specific working memory resources. *J. Mem. Lang.* 54, 541–553. doi: 10.1016/j.jml.2005.12.006

Feredoes, E., Tononi, G., and Postle, B. R. (2006). Direct evidence for a prefrontal contribution to the control of proactive interference in verbal working memory. *Proc. Natl. Acad. Sci. U.S.A.* 103, 19530–19534. doi: 10.1073/pnas.0604509103

Fiebach, C. J., Schlesewsky, M., and Friederici, A. D. (2001). Syntactic working memory and the establishment of filler-gap dependencies: insights from ERPs and fMRI. *J. Psycholinguist. Res.* 30, 321–338. doi: 10.1023/A:1010447102554

Fiebach, C. J., Schlesewsky, M., Lohmann, G., von Cramon, D. Y., and Friederici, A. D. (2005). Revisiting the role of Broca's area in sentence processing: syntactic integration versus syntactic working memory. *Hum. Brain Mapp.* 24, 79–91. doi: 10.1002/hbm.20070

Fiebach, C. J., Vos, S. H., and Friederici, A. D. (2004). Neural correlates of syntactic ambiguity in sentence comprehension for low and high span readers. *J. Cogn. Neurosci.* 16, 1562–1575. doi: 10.1162/0898929042568479

Foraker, S., and McElree, B. (2011). Comprehension of linguistic dependencies: speed-accuracy tradeoff evidence for direct-access retrieval from memory. *Lang. Linguist. Compass* 5, 764–783. doi: 10.1111/j.1749-818X.2011.00313.x

Ford, M. (1983). A method for obtaining measures of local parsing complexity throughout sentences. *J. Verbal Learn. Verbal Behav.* 22, 203–218. doi: 10.1016/S0022-5371(83)90156-1

Frazier, L., and Clifton, C. (2001). Parsing coordinates and ellipsis: copy α. *Syntax* 4, 1–22. doi: 10.1111/1467-9612.00034

Gardiner, J. M., Craik, F. I. M., and Birtwistle, J. (1972). Retrieval cues and release from proactive inhibition. *J. Verbal Learn. Verbal Behav.* 11, 778–783. doi: 10.1016/S0022-5371(72)80012-4

Gordon, P. C., Hendrick, R., and Johnson, M. (2001). Memory interference during language processing. *J. Exp. Psychol. Learn. Mem. Cogn.* 27, 1411–1423. doi: 10.1037/0278-7393.27.6.1411

Gordon, P. C., Hendrick, R., and Levine, W. H. (2002). Memory-load interference in syntactic processing. *Psychol. Sci.* 13, 425–430. doi: 10.1111/1467-9280.00475

Gough, P. M., Nobre, A. C., and Devlin, J. T. (2005). Dissociating linguistic processes in the left inferior frontal cortex with transcranial magnetic stimulation. *J. Neurosci.* 25, 8010–8016. doi: 10.1523/JNEUROSCI.2307-05.2005

Gouvea, A. C., Phillips, C., Kazanina, N., and Poeppel, D. (2010). The linguistic processes underlying the P600. *Lang. Cogn. Process.* 25, 149–188. doi: 10.1080/01690960902965951

Grewe, T., Bornkessel, I., Zysset, S., Wiese, R., von Cramon, D. Y., and Schlesewsky, M. (2005). The emergence of the unmarked: a new perspective on the language-specific function of Broca's area. *Hum. Brain Mapp.* 26, 178–190. doi: 10.1002/hbm.20154

Grodzinsky, Y. (1986). Language deficits and the theory of syntax. *Brain Lang.* 27, 135–159. doi: 10.1016/0093-934X(86)90009-X

Grodzinsky, Y. (2000). The neurology of syntax: language use without Broca's area. Target article with 36 commentaries. *Behav. Brain Sci.* 23, 47–117. doi: 10.1017/S0140525X00002399

Grodzinsky, Y., and Friederici, A. D. (2006). Neuroimaging of syntax and syntactic processing. *Curr. Opin. Neurobiol.* 16, 240–246. doi: 10.1016/j.conb.2006.03.007

Grodzinsky, Y., and Santi, A. (2008). The battle for Broca's region. *Trends Cogn. Sci.* 12, 474–480. doi: 10.1016/j.tics.2008.09.001

Guo, Y., Martin, R., Van Dyke, J. A., and Hamilton, C. (2010). "Interference effects in sentence comprehension: an fMRI study," in *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, eds S. Ohlsson and R. Catrambone (Austin, TX: Cognitive Science Society), 1429–1434.

Ha, S. (2008). *Right Node Raising, and Across-the-Board Constructions*. Doctoral Dissertation, Boston University, Boston, MA.

Haber, L. R., and Haber, R. N. (1982). Does silent reading involve articulation? Evidence from tongue twisters. *Am. J. Psychol.* 95, 409–419. doi: 10.2307/1422133

Hagoort, P. (2003). How the brain solves the binding problem for language: a neurocomputational model of syntactic processing. *Neuroimage* 20(Suppl. 1), S18–S29. doi: 10.1016/j.neuroimage.2003.09.013

Hagoort, P. (2005). On Broca, brain, and binding: a new framework. *Trends Cogn. Sci.* 9, 416–423. doi: 10.1016/j.tics.2005.07.004

Hakes, D. T., Evans, J. S., and Brannon, L. L. (1976). Understanding sentences with relative clauses. *Mem. Cogn.* 4, 283–290. doi: 10.3758/BF03213177

Hartmann, K. (2001). *Right Node Raising and Gapping. Interface Conditions on Prosodic Deletion*. Philadelphia, PA: John Benjamins Publishing Company.

Hartmann, K. (2003). "Background matching in right node raising constructions," in *Interfaces: Deriving and Interpreting Omitted Structures*, eds K. Schwabe and S. Winkler (Philadelphia, PA: John Benjamins Publishing), 121–151.

Hofmeister, P. (2011). Representational complexity and memory retrieval in language comprehension. *Lang. Cogn. Process.* 26, 376–405. doi: 10.1080/01690965.2010.492642

Holmes, V. M. (1973). Order of main and subordinate clauses in sentence perception. *J. Verbal Learn. Verbal Behav.* 12, 285–293. doi: 10.1016/S0022-5371(73)80072-6

Holmes, V. M., and O'Regan, J. K. (1981). Eye fixation patterns during the reading of relative-clause sentences. *J. Verbal Learn. Verbal Behav.* 20, 417–430. doi: 10.1016/S0022-5371(81)90533-8

Just, M. A., Carpenter, P. A., Keller, T. A., Eddy, W. F., and Thulborn, K. R. (1996). Brain activation modulated by sentence comprehension. *Science* 274, 114–116. doi: 10.1126/science.274.5284.114

Kaan, E., Harris, A., Gibson, E., and Holcomb, P. (2000). The P600 as an index of syntactic integration difficulty. *Lang. Cogn. Process.* 15, 159–201. doi: 10.1080/016909600386084

Kaan, E., and Swaab, T. Y. (2002). The brain circuitry of syntactic comprehension. *Trends Cogn. Sci.* 6, 350–356. doi: 10.1016/S1364-6613(02)01947-2

Kayne, R. S. (1994). *The Antisymmetry of Syntax*. Cambridge, Mass: MIT Press.

Keller, T. A., Carpenter, P. A., and Just, M. A. (2001). The neural bases of sentence comprehension: a fMRI examination of syntactic and lexical processing. *Cereb. Cortex* 11, 223–237. doi: 10.1093/cercor/11.3.223

King, J., and Just, M. A. (1991). Individual differences in syntactic processing: the role of working memory. *J. Mem. Lang.* 30, 580–602. doi: 10.1016/0749-596X(91)90027-H

King, J. W., and Kutas, M. (1995). Who did what and when? Using word- and clause-level ERPs to monitor working memory usage in reading. *J. Cogn. Neurosci.* 7, 376–395. doi: 10.1162/jocn.1995.7.3.376

Lancaster, J. L., Woldorff, M. G., Parsons, L. M., Liotti, M., Freitas, C. S., Rainey, L., et al. (2000). Automated Talairach Atlas labels for functional brain mapping. *Hum. Brain Mapp.* 10, 120–131.

Lee, Y., Lee, H., and Gordon, P. C. (2007). Linguistic complexity and information structure in Korean: evidence from eye-tracking during reading. *Cognition* 104, 495–534. doi: 10.1016/j.cognition.2006.07.013

Leiken, K., and Pylkkänen, L. (2014). MEG evidence that the LIFG effect of object extraction requires similarity-based interference. *Lang. Cogn. Process.* 29, 381–389. doi: 10.1080/01690965.2013.863369

Lewis, R. L. (1996). Interference in short-term memory: the magical number two (or three) in sentence processing. *J. Psycholinguist. Res.* 25, 93–115. doi: 10.1007/BF01708421

Makuuchi, M., Bahlmann, J., Anwander, A., and Friederici, A. D. (2009). Segregating the core computational faculty of human language from working memory. *Proc. Natl. Acad. Sci. U.S.A.* 106, 8362–8367. doi: 10.1073/pnas.0810928106

Maris, E., and Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *J. Neurosci. Methods* 164, 177–190. doi: 10.1016/j.jneumeth.2007.03.024

Martin, A. E., and McElree, B. (2008). A content-addressable pointer mechanism underlies comprehension of verb-phrase ellipsis. *J. Mem. Lang.* 58, 879–906. doi: 10.1016/j.jml.2007.06.010

Martin, A. E., and McElree, B. (2009). Memory operations that support language comprehension: evidence from verb-phrase ellipsis. *J. Exp. Psychol. Learn. Mem. Cogn.* 35, 1231–1239. doi: 10.1037/a0016271

Martin, A. E., and McElree, B. (2011). Direct-access retrieval during sentence comprehension: evidence from sluicing. *J. Mem. Lang.* 64, 327–343. doi: 10.1016/j.jml.2010.12.006

McCutchen, D., Bell, L. C., France, I. M., and Perfetti, C. A. (1991). Phoneme-specific interference in reading: the tongue-twister effect revisited. *Read. Res. Q.* 26, 87–103. doi: 10.2307/747733

McElree, B. (2000). Sentence comprehension is mediated by content-addressable memory structures. *J. Psycholinguist. Res.* 29, 111–123. doi: 10.1023/A:1005184709695

McElree, B. (2001). Working memory and focal attention. *J. Exp. Psychol. Learn. Mem. Cogn.* 27, 817–835. doi: 10.1037/0278-7393.27.3.817

McElree, B., Foraker, S., and Dyer, L. (2003). Memory structures that subserve sentence comprehension. *J. Mem. Lang.* 48, 67–91. doi: 10.1016/S0749-596X(02)00515-6

McElree, B., Murphy, G. L., and Ochoa, T. (2006). Time-course of retrieving conceptual information: a speed-accuracy tradeoff study. *Psychon. Bull. Rev.* 13, 848–853. doi: 10.3758/BF03194008

Mead, L. A., Mayer, A. R., Bobholz, J. A., Woodley, S. J., Cunningham, J. M., Hammeke, T. A., et al. (2002). Neural basis of the Stroop interference task: response competition or selective attention? *J. Int. Neuropsychol. Soc. JINS* 8, 735–742. doi: 10.1017/S1355617702860015

Miller, E. K., and Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annu. Rev. Neurosci.* 24, 167–202. doi: 10.1146/annurev.neuro.24.1.167

Mosher, J. C., and Leahy, R. M. (1998). Recursive MUSIC: a framework for EEG and MEG source localization. *IEEE Trans. Biomed. Eng.* 45, 1342–1354. doi: 10.1109/10.725331

Nairne, J. S. (2002). Remembering over the short-term: the case against the standard model. *Annu. Rev. Psychol.* 53, 53–81. doi: 10.1146/annurev.psych.53.100901.135131

Novick, J. M., Trueswell, J. C., and Thompson-Schill, S. L. (2005). Cognitive control and parsing: reexamining the role of Broca's area in sentence comprehension. *Cogn. Affect. Behav. Neurosci.* 5, 263–281. doi: 10.3758/CABN.5.3.263

Nystrom, L. E., and McClelland, J. L. (1992). Trace synthesis in cued recall. *J. Mem. Lang.* 31, 591–614. doi: 10.1016/0749-596X(92)90030-2

Öztekin, I., Curtis, C. E., and McElree, B. (2009). The medial temporal lobe and the left inferior prefrontal cortex jointly support interference resolution in verbal working memory. *J. Cogn. Neurosci.* 21, 1967–1979. doi: 10.1162/jocn.2008.21146

Öztekin, I., and McElree, B. (2007). Proactive interference slows recognition by eliminating fast assessments of familiarity. *J. Mem. Lang.* 57, 126–149. doi: 10.1016/j.jml.2006.08.011

Öztekin, I., McElree, B., Staresina, B. P., and Davachi, L. (2008). Working memory retrieval: contributions of the left prefrontal cortex, the left posterior parietal cortex, and the hippocampus. *J. Cogn. Neurosci.* 21, 581–593. doi: 10.1162/jocn.2008.21016

Phillips, C., Kazanina, N., and Abada, S. H. (2005). ERP effects of the processing of syntactic long-distance dependencies. *Cogn. Brain Res.* 22, 407–428. doi: 10.1016/j.cogbrainres.2004.09.012

Postal, P. M. (1974). *On Raising: One Rule of English Grammar and its Theoretical Implications*. Cambridge, MA: MIT Press.

Postal, P. M. (1998). *Three Investigations of Extraction*. Cambridge, MA: MIT Press.

Postle, B. R., Brush, L. N., and Nick, A. M. (2004). Prefrontal cortex and the mediation of proactive interference in working memory. *Cogn. Affect. Behav. Neurosci.* 4, 600–608. doi: 10.3758/CABN.4.4.600

Pylkkänen, L., Llinás, R., and McElree, B. (2004). "Distinct effects of semantic plausibility and semantic composition in MEG," in *Proceedings of the 14th International Conference on Biomagnetism Biomag 2004*, Boston, MA.

Pylkkänen, L., and Marantz, A. (2003). Tracking the time course of word recognition with MEG. *Trends Cogn. Sci.* 7, 187–189. doi: 10.1016/S1364-6613(03)00092-5

Pylkkänen, L., Stringfellow, A., and Marantz, A. (2002). Neuromagnetic evidence for the timing of lexical activation: an MEG component sensitive to phonotactic probability but not to neighborhood density. *Brain Lang.* 81, 666–678. doi: 10.1006/brln.2001.2555

Robinson, G., Blair, J., and Cipolotti, L. (1998). Dynamic aphasia: an inability to select between competing verbal responses? *Brain* 121, 77–89. doi: 10.1093/brain/121.1.77

Rogalsky, C., Matchin, W., and Hickok, G. (2008). Broca's area, sentence comprehension, and working memory: an fMRI Study. *Front. Hum. Neurosci.* 2:14. doi: 10.3389/neuro.09.014.2008

Ross, J. R. (1968). *Constraints on Variables in Syntax*. Cambridge, MA: Massachusetts Institute of Technology.

Sabbagh, J. (2007). Ordering and linearizing rightward movement. *Nat. Lang. Linguist. Theory* 25, 349–401. doi: 10.1007/s11049-006-9011-8

Santi, A., and Grodzinsky, Y. (2007). Working memory and syntax interact in Broca's area. *Neuroimage* 37, 8–17. doi: 10.1016/j.neuroimage.2007.04.047

Stowe, L. A., Paans, A. M., Wijers, A. A., Zwarts, F., Mulder, G., and Vaalburg, W. (1999). Sentence comprehension and word repetition: a positron emission tomography investigation. *Psychophysiology* 36, 786–801. doi: 10.1111/1469-8986.3660786

Stromswold, K., Caplan, D., Alpert, N., and Rauch, S. (1996). Localization of syntactic comprehension by positron emission tomography. *Brain Lang.* 52, 452–473. doi: 10.1006/brln.1996.0024

Swingle, K. (1993). The role of prosody in right node raising. *Paper Presented at the Syntax at Santa Cruz*, Vol. 2, eds G. Pullum and E. Potsdam (Santa Cruz, CA: Syntax Research Center), 82–112.

Thompson-Schill, S. L., D'Esposito, M., Aguirre, G. K., and Farah, M. J. (1997). Role of left inferior prefrontal cortex in retrieval of semantic knowledge: a reevaluation. *Proc. Natl. Acad. Sci. U.S.A.* 94, 14792–14797. doi: 10.1073/pnas.94.26.14792

Thothathiri, M., Schwartz, M. F., and Thompson-Schill, S. L. (2010). Selection for position: the role of left ventrolateral prefrontal cortex in sequencing language. *Brain Lang.* 113, 28–38. doi: 10.1016/j.bandl.2010.01.002

Traxler, M. J., Williams, R. S., Blozis, S. A., and Morris, R. K. (2005). Working memory, animacy, and verb class in the processing of relative clauses. *J. Mem. Lang.* 53, 204–224. doi: 10.1016/j.jml.2005.02.010

Van Dyke, J. A., and McElree, B. (2006). Retrieval interference in sentence comprehension. *J. Mem. Lang.* 55, 157–166. doi: 10.1016/j.jml.2006.03.007

Van Dyke, J. A., and McElree, B. (2011). Cue-dependent interference in comprehension. *J. Mem. Lang.* 65, 247–263. doi: 10.1016/j.jml.2011.05.002

Wanner, H. E., and Maratsos, M. P. (1978). "An ATN approach to comprehension," in *Linguistic Theory and Psychological Reality*, eds M. A. Halle, J. Bresnan, and G. A. Miller (Cambridge, MA: MIT Press).

Waters, G., Caplan, D., and Hildebrandt, N. (1987). "Working memory and written sentence comprehension," in *Attention and Performance 12: The Psychology of Reading*, ed. M. Coltheart (Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.), 531–555.

Watkins, O. C., and Watkins, M. J. (1975). Buildup of proactive inhibition as a cue-overload effect. *J. Exp. Psychol. Hum. Learn. Mem.* 1, 442–452. doi: 10.1037/0278-7393.1.4.442

Wexler, K. (1980). *Formal Principles of Language Acquisition*. Cambridge, MA: MIT Press.

Wickens, D. D. (1973). Some characteristics of word encoding. *Mem. Cogn.* 1, 485–490. doi: 10.3758/BF03208913

Wilder, C. (1997). "Some properties of ellipsis in coordination," in *Studies on Universal Grammar and Typological Variation*, eds A. Alexiadou and T. A. Hall (Amsterdam: John Benjamins Publishing), 59–107.

Zurif, E. B. (1995). "Brain regions of relevance to syntactic processing," in *An Invitation to Cognitive Science: Language*, Vol. 1, 2nd Edn, eds L. Gleitman and M. Liberman (Cambridge, MA: MIT Press).

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2015 Leiken, McElree and Pylkkänen. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 10 April 2015
doi: 10.3389/fpsyg.2015.00349

Task-dependency and structure-dependency in number interference effects in sentence comprehension

Julie Franck[1]*, *Saveria Colonna*[2] and *Luigi Rizzi*[3,4]

[1] Laboratoire de Psycholinguistique, University of Geneva, Geneva, Switzerland, [2] Centre National de la Recherche Scientifique – University of Paris 8, Paris, France, [3] Department of Linguistics, University of Geneva, Geneva, Switzerland, [4] Interdepartmental Centre for Cognitive Studies of Language, University of Siena, Siena, Italy

We report three experiments on French that explore number mismatch effects in intervention configurations in the comprehension of object A'-dependencies, relative clauses and questions. The study capitalizes on the finding of object attraction in sentence production, in which speakers sometimes erroneously produce a verb that agrees in number with a plural object in object relative clauses. Evidence points to the role of three critical constructs from formal syntax: intervention, intermediate traces and c-command (Franck et al., 2010). Experiment 1, using a self-paced reading procedure on these grammatical structures with an agreement error on the verb, shows an enhancing effect of number mismatch in intervention configurations, with faster reading times with plural (mismatching) objects. Experiment 2, using an on-line grammaticality judgment task on the ungrammatical versions of these structures, shows an interference effect in the form of attraction, with slower response times with plural objects. Experiment 3 with a similar grammaticality judgment task shows stronger attraction from c-commanding than from preceding interveners. Overall, the data suggest that syntactic computations in performance refer to the same syntactic representations in production and comprehension, but that different tasks tap into different processes involved in parsing: whereas performance in self-paced reading reflects the intervention of the subject in the process of building an object A'-dependency, performance in grammaticality judgment reflects intervention of the object on the computation of the subject-verb agreement dependency. The latter shows the hallmarks of structure-dependent attraction effects in sentence production, in particular, a sensitivity to specific characteristics of hierarchical representations.

Keywords: number, agreement, attraction, intervention, intermediate traces, c-command, cue-based retrieval, comprehension

OPEN ACCESS

Edited by:
Claudia Felser, University of Potsdam, Germany

Reviewed by:
Jana Häussler, University of Potsdam, Germany
Laurel Brehm, Northwestern University, USA

***Correspondence:**
Julie Franck, Laboratoire de Psycholinguistique, University of Geneva, 40 Boulevard du Pont d'Arve, 1205 Geneva, Switzerland
julie.franck@unige.ch

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: *24 December 2014*
Accepted: *11 March 2015*
Published: *10 April 2015*

Citation:
Franck J, Colonna S and Rizzi L (2015) Task-dependency and structure-dependency in number interference effects in sentence comprehension. Front. Psychol. 6:349. doi: 10.3389/fpsyg.2015.00349

Introduction

The wide literature on agreement in sentence production has given rise to a large body of research on the phenomenon of interference called 'attraction.' In the standard and most explored case, the speaker incorrectly produces a verb that agrees with a plural noun situated in a modifying

prepositional phrase (PP) linearly intervening between the subject and the verb (e.g., **The time for fun and games are over*, from Bock and Miller, 1991; Bock and Cutting, 1992). As experimental evidence accumulated, it has become evident that various types of syntactic elements have the potential to trigger interference, including adjuncts (Franck et al., 2004) and immediately preverbal objects (Fayol et al., 1994; Hartsuiker et al., 2001; Hemforth and Konieczny, 2003; Konieczny et al., 2004; Franck et al., 2006, 2010; Häussler, unpublished), but also and more intriguingly elements that are not situated between the subject and the inflected verb in the linear word string. Such cases of interference have been reported in object relative clauses as illustrated in (1a) (Bock and Miller, 1991; Franck et al., 2006, 2010), questions (Vigliocco and Nicol, 1998), as well as cleft sentences (Franck et al., 2006). The present study addresses the question of whether similar interference effects are detectable in sentence comprehension. In particular, the work aims to address three questions. First, do the critical constructs from formal syntax, i.e., intervention, intermediate traces, and c-command, which capture attraction patterns in agreement production (Franck et al., 2006, 2010), also play a role in the computation of agreement in sentence comprehension? Second, are the processes involved in the computation of agreement features the same in production and comprehension? Third do different experimental techniques in comprehension tap on distinct aspects of agreement computation in performance?

(1) a. Jean parle aux **patientes** [$_{RC}$ que le médicament *guérissent.]
*Jean speaks to-the-PL patients-PL that the medicine-SG *cure-PL.*
*'John speaks to the patients whom the medicine *cure.'*
b. Jean dit aux **patientes** [$_{CC}$ que le médicament guérit.]
Jean tells to-the-PL patients-PL that the medicine-SG cures-SG.
'John tells the patients that the medicine cures.'

The paper is structured as follows. We first present the syntactic configurations underlying object interference. We then turn to the role of the experimental task in agreement computation. Subsequently, we report three experiments exploring the role of syntactic configurations in object interference in sentence comprehension. The first two experiments use different methodologies to test the role of movement and intermediate traces: they contrast attraction in the minimal structural pair consisting of object relative clauses, involving movement (as in 1a), and the superficially similar complement clauses without movement (as in 1b). The third experiment tests the role of c-command by manipulating attraction from moved complex objects. These objects involve both a c-commanding DP and a purely preceding DP within a PP modifier whose respective effects on attraction are systematically assessed (e.g., Quelles **patientes** du médecin dis-tu que le juriste défend? *Which-PL patients-PL of the doctor do you say that the-SG lawyer-SG defends-SG?* vs. Le chirurgien de quelles **patientes** dis-tu que le juriste défend? *The surgeon of which-PL patients-PL do you say that the-SG lawyer-SG defends-SG?*).

The Role of Syntactic Structure in Object Interference

In a detailed exploration of object attraction in sentence production, Franck et al. (2010) tested various hypotheses with respect to the structural conditions underlying agreement errors. The starting point of their work was the finding that despite its close superficial resemblance to the object relative (1a), the sentence complement clause (1b) fails to trigger attraction. Whereas in (1a) *patients* is the moved object of the target verb *cure* used transitively, in (1b) it is the unmoved object of the main verb *tells* while the target verb is used intransitively.

In four additional experiments, the authors explored the role of properties that distinguish relative clauses from complement clauses. Argumenthood was found to play no role in attraction since objects that are not part of the argument structure of the target verb, as in extraction from clausal complements, trigger similar attraction as thematic objects (e.g., Voici **les otages** que le journaliste *apprennent qu'on a blessés; *Here are the hostages-PL that the journalist-SG *learn-PL that someone injured*), while objects in their canonical post-verbal position did not generate attraction either. Participle agreement triggered by the object in French was also found to play no role in attraction, since attraction effects were found with elements that fail to trigger participle agreement like accusative clitics in the causative construction (e.g., Le directeur **les** *font acheter; *The director-SG them-PL make-PL *buy*). Moreover, the strength of object attraction in structures in which the object has moved to the front of the sentence and fails to intervene linearly between the subject and the verb (relatives and clefts) was shown to be of a similar strength to that of a linearly intervening object, as is the case of the clitic object pronoun (e.g., L'avocat **les** *défendent; *The-SG lawyer them-PL *defend*, Franck et al., 2006). All these cases involve an object (or its trace) intervening in a c-commanding position between the terms of the agreement relation, the subject and the agreeing verb (see Franck et al., 2006, 2010 for a graphical illustration of the hierarchical structure and c-command relations). Evidence suggests that attraction is significantly weaker if the attractor intervenes purely in terms of precedence, i.e., in a position situated lower down in the tree and that fails to c-command the agreement node, as is the case of the modifying PP (e.g., L'avocat des **patientes** *mentent; *The-SG lawyer of the patients-PL *lie-PL*) or the dative clitic (e.g., L'avocat **leur** *mentent; *The-SG lawyer to them-PL *lie, The lawyer lie to them*, Franck et al., 2006). In sum, the empirical evidence suggests that object movement is a necessary and sufficient condition for object attraction to arise in an SVO language like French (conditions may differ for SOV languages if the object originates in preverbal position), and that c-commanding attractors generate more interference than preceding ones.

But why does the object interfere with agreement even in cases like (1a), in which it is pronounced in a position from which it does not intervene, either linearly or hierarchically, between the subject and the verb? Franck et al. (2006, 2010) noted that interference with the object seems to occur at a position that is neither the final nor the initial position since these two positions failed to generate attraction. The authors proposed that the intermediate position of the object in the hierarchical structure, mediating

its initial position in the thematic structure and its final surface position, plays a crucial role. In a (much simplified) representation of example (1a) like, *Jean parle aux* ***patientes*** *[*$_{RC}$ *que le médicament* t_2 **guérissent* t_1*]*, the object *patientes* is initially generated in $\mathbf{t_1}$ and then moves in $\mathbf{t_2}$, a position which intervenes on the subject-verb agreement (AGREE) relation between the position hosting agreement morphology (ultimately attached to the verb) and the subject in its initial thematic position. Finally, the object moves higher to reach its final position. The intermediate position $\mathbf{t_2}$, unpronounced but with morphological reflexes in some cases such as participial agreement in French (Kayne, 1989) or wh-agreement in Austronesian languages (Chung, 1998), is postulated in formal syntactic analyses for locality reasons (e.g., to respect Phase Impenetrability in a system like Chomsky, 2001; see Gibson and Warren, 2004 for experimental evidence for the role of intermediate traces in the processing of long-distance dependencies). So, we argued that formal characteristics of abstract representations assumed in the "principles and parameters"/minimalist analysis of agreement form the representational basis over which agreement processes operate in performance. The intermediate trace of the object in the vP periphery may be thought of, in a phase-based architecture, as corresponding to a temporary memory buffer from which the object remains active and available for further processes (Chesi, 2005). The activation of the object in memory in this precise position intervening on agreement would be the locus of the interference effects observed in agreement production.

Task-Effects in Object Interference

Production and comprehension critically differ in that whereas in production the speaker has access to the conceptual structure of the sentence, this structure is incrementally built in sentence comprehension, under the strong guidance of predictive mechanisms (see e.g., Hale, 2001; Levy, 2008). Nevertheless, under the view that the effects reported in agreement production reflect properties of the hierarchical structure over which agreement is calculated, one expects the same effects to show up in sentence comprehension (e.g., see Kempen et al., 2012, for a model with shared representations and shared processes of syntactic encoding and decoding). Various studies have shown that plural attractors situated in the subject phrase interfere with verb agreement processing in sentence comprehension (e.g., Nicol et al., 1997; Pearlmutter et al., 1999; Pearlmutter, 2000; Kaan, 2002; Thornton and MacDonald, 2003; Häussler, unpublished). The typical finding is that participants spend more time reading or judging the acceptability of a sentence in the presence of a plural mismatching subject modifier as compared to when it is singular.

However, only a few studies have explored object attraction in comprehension. Clifton and colleagues conducted two grammaticality judgment studies exploring attraction in structures like (2) (Clifton et al., 1999). They found that although participants correctly reject ungrammatical sentences like (2a), they tend to accept ungrammatical sentences like (2b) in which a plural element (people) is situated higher than the subject and the verb. Clifton et al. argued that the relative acceptability of (2b) lies in the fact that the chain between the moved element and its trace is still active, and the agreement dependency is on its path. The effect was attributed to the passing of the plural feature on the agreement dependency and not to a late gap complexity effect affecting the general difficulty in processing (2b), since the same sentences with a singular attractor (*people* was replaced by *person*) were systematically rejected, attesting that the difficulty selectively arises in the presence of the plural feature (see also Kayne, 2000 for the hypothesis that the agreement pattern in (2b) is fully grammatical in some dialects). The authors concluded that interference arises because of the link of the projection path that is shared by two distinct syntactic dependencies (agreement and the NP-trace chain). These findings and the overall interpretation are consistent with the analysis (just summarized) in Franck et al. (2006, 2010), which would further specify that interference would be triggered in (2b) by the transit of the plural relative head *people* in the periphery of the vP headed by *think*, a position from which *people* hierarchically intervenes between *manager* and the agreeing head, thus giving rise to the plural form *think*.

(2) a. Lucine dislikes **the people** who think the manager *know the answers.
b. Lucine dislikes **the people** who the manager (*)think **t** know the answers.

Using a methodology based on a two-choice response time paradigm, Staub (2009, 2010) also found evidence for interference from moved objects when participants were asked to select between the two verb forms (singular vs. plural) presented simultaneously on the screen after being exposed to the subject phrase one word at a time in a Rapid Serial Visual Presentation (RSVP) procedure. Verb selection was found to be sensitive to attraction similarly to sentence production, with slower response times to select the correct verb form in the presence of a mismatching plural feature on the object of the sentence. Note, however, that it is unclear whether the task taps into comprehension or production, since both components are involved.

Wagers et al. (2009) investigated object interference in two experiments on object relative clauses (Experiments 2 and 3) involving a self-paced reading procedure. The results showed a significant effect of the object number in the region following the critical verb. However, the effect was restricted to ungrammatical sentences and showed up in the form of faster reading times in the presence of a mismatching plural attractor, suggesting that it reflects a grammatical illusion lying in the incorrect computation of agreement. More generally, across the five self-paced reading experiments on both object and subject modifier attraction, the authors consistently found no attraction in grammatical sentences. This finding appears prima facie to be in contradiction with the other reports of significant interference in grammatical sentences involving subject modifiers, but the careful testing of these structures by Wagers et al. (2009) suggests that the attraction effect observed at the verb in these studies was actually a spillover effect from the plural feature preceding the verb. In line with that interpretation, they found that the introduction of an adverb between the modifier and the verb dissolved the effect (e.g., The slogan on the posters unsurprisingly was designed to get attention). Recent evidence from on-line experimental techniques further supports the view that attraction is restricted to

ungrammatical sentences (eye-tracking: Dillon et al., 2013; ERP: Tanner et al., 2014). The possibility that attraction only arises in ungrammatical sentences in comprehension has important consequences for models of agreement computation. Wagers et al. (2009) suggest that attraction in sentence comprehension is driven by the properties of a cue-based retrieval process triggered when the parser encounters an agreement error: the system involves a predictive component by which the parser expects a particular number feature on the verb, and only if the bottom-up features of the verb mismatch the top–down predicted features is the cue-based-retrieval deployed to check whether the correct feature was missed during the first pass. On this view, number interference in comprehension arises from a fundamentally different cause from attraction errors in sentence production.

Nevertheless, as Wagers et al. (2009) point out, other studies reported reliable interference effects in grammatical sentences that cannot easily be explained by an extended effect of the plural attractor. In a series of five experiments using a Maze task (requiring for each upcoming word in the sentence to select amongst two words, and in the critical region, between a correctly agreeing verb and a word from a different grammatical category, e.g., *was* vs. *ink*) or a sentence classification task (requiring to determine whether the sentence is a legitimately ordered string of words), Nicol et al. (1997) found significant interference in grammatical sentences. Since in these tasks response times reflect either the selection of the correct grammatical category or the global assessment of the sentence word order, the interference found does not seem to be attributable to the spillover of the plural feature processing on the verb. In a speeded grammaticality judgment task on sentences containing an embedded clause with complex subjects modified by a genitive phrase, Häussler et al. (2003) reported interference from the plural attractor in grammatical sentences only; no interference was found in ungrammatical sentences. Using a similar procedure of grammaticality judgment, Häussler and Bader (2009) also found interference from a mismatching feature within a relative clause introduced by a possessive pronoun in both grammatical and ungrammatical sentences. Here again, the slow down observed in mismatch condition cannot be attributed to a spillover effect of processing the plural feature on the attractor linearly preceding the verb.

Summing up, while some studies of attraction in sentence comprehension point to similarities with sentence production, others suggest differences. Moreover, discrepancies are also found between studies of attraction in sentence comprehension, some reporting attraction in ungrammatical sentences only, others in both grammatical and ungrammatical sentences, and yet others finding attraction in grammatical sentences only. However, a direct comparison across these studies is difficult due to the fact that they involve different tasks and different syntactic structures. The role of the task in language performance gained increased interest in the recent years (e.g., Caplan et al., 2008; Caplan, 2010; see also Salverda et al., 2010 for an overview of task effects in the visual world paradigm), and it therefore seems crucial to collect finely controlled comparisons on agreement performance before conclusions be drawn with respect to the mechanism of attraction.

In order to test the potential influence of the task on attraction, Experiments 1 and 2 use the same materials tested in agreement production by Franck et al. (2010), but with two different experimental methods. Experiment 1 uses a similar self-paced reading procedure to that used by Wagers et al. (2009) but differs from it in that only grammatical sentences were introduced, to maximize the naturalness of the comprehension process and avoid any potential contamination from the presence of ungrammatical sentences. Experiment 2 uses a procedure of speeded grammaticality judgment on the ungrammatical versions of these sentences, with the aim of maximally promoting agreement computation in comprehension. Finally, since Experiment 2 only involved ungrammatical sentences, Experiment 3 manipulated grammaticality in order to assess its role in the same procedure of speeded grammaticality judgment used in Experiment 2.

Experiment 1: Object Interference in Self-Paced Reading

Experiment 1 explores the role of object movement, modulated by its number specification, in a maximally natural sentence comprehension task involving the self-paced reading of grammatical sentences. The same materials as in Experiment 1 from Franck et al. (2010) were used, involving a minimal contrast between a structure with object movement in an object relative clause (1a), and a structure without object movement in a complement clause (1b). The two structures are identical in surface order; the main difference lies in the selection of the main verb, which takes a single complement in (1a) (thus enforcing the analysis of the *que* clause as a relative modifying the object DP) and two complements in (1b) (thus giving rise to a sentential complement interpretation for the *que* clause). Interference is examined on the agreement of the verb in the subordinate clause (to cure). In the relative clause (1a), 'patient(s)' is the object of the target verb 'cures,' and is therefore assumed to transit via the intermediate position at the embedded vP periphery intervening on the subject-verb agreement relation (Franck et al., 2010). In the complement clause (1b), 'patient(s)' is the unmoved indirect object of the main verb while the embedded verb 'cures' is used intransitively. If the intervention of the intermediate trace of the moved object on agreement creates interference in sentence comprehension similarly to sentence production, a slow down is expected at the verb in the presence of a mismatching plural object as compared to a singular object in object relatives (1a), but not in sentence complements (1b). However, if attraction in sentence comprehension reflects a process of 'rechecking' triggered by an erroneous agreement, no interference is expected from the plural feature in either of the two structures.

Method

Participants

Seventy-two students from the University of Geneva, aged between 18 and 40, took part in the experiment. They

received course credit for their participation. The experiment was approved by the ethics committee of the Department of Psychology of the University of Geneva and informed consent was obtained from all participants.

Materials

The experimental materials consisted of the 24 sentences used in Franck et al. (2010) incorporated in a 2 × 2 design crossing structure (relative vs. complement) and the number of the object (singular vs. plural). All subject head nouns were singular. Subjects and objects were all animate. Since the initial sentences ended with the target verb, two windows were added after the verb in order to measure potential spillover effects. These windows contained an adverb followed by a locative phrase. Each sentence was decomposed into 8 windows corresponding to phrases (content word + grammatical word if present). All test sentences were grammatical with respect to subject-verb agreement. Each sentence was followed by a yes/no comprehension question that probed participants' interpretation of the thematic relations in the sentence. Examples of test items are presented in **Table 1** (the full list of items is available in the Supplementary Materials).

An additional set of 48 grammatical filler items were built. Half of them had the same structure as the experimental materials (12 Object relative clauses and 12 Complement clauses) but with plural subjects (half with singular objects). The other half involved a variety of syntactic structures (eight declaratives, eight relatives, four temporal modifiers, four PP modifiers) with a varying number of reading windows.

Procedure

Sentences were presented on a computer screen using the e-prime software in a self-paced moving window paradigm (Just et al., 1982). Each sentence was first presented with dashes replacing words. Participants were instructed to read sentences by pressing the space bar in order to have the segments appear. Once read, windows disappeared from the screen such that only one window was readable at a time. Participants were told that they would also have to answer yes/no comprehension questions about the content of these sentences. Instructions encouraged both rapid reading and correctness in answering the questions. Order of presentation of the sentences was randomized. The experiment started with four practice trials.

Data analyses

Analyses of reading times were run after excluding incorrect responses to comprehension questions (181 data points rejected representing 10.2% of the data). Remaining response times were then trimmed for outliers, defined as data points with a value above 3 s for all participants and all regions (representing 3.1% over all responses). They were treated as missing values. Log-transformed response times and accuracy proportions were analyzed with (generalized) linear mixed-effects regression models with random intercepts for participants and items (Baayen et al., 2008), using the statistical software R (R Development Core Team, 2013). Estimates, *t*-values (for LME), *z*-values (for GLME) and *p*-values for the fixed factors and interactions were obtained via the lmerTest package, which provides *p*-values calculated based on Satterthwaite's approximation. Significant interactions were further explored with (generalized) linear mixed-effects regression models separately on each of the two modalities of one of the variables involved in the interaction.

Results

Reading Times

The distribution of reading times across the different experimental conditions in the different regions is reported in **Figure 1**.

Region 2 (main verb). A marginal effect of structure was found with slower response times on the main verb of the relative clause condition (581 ms) than on the main verb of the complement clause (561 ms; $\beta = -0.033$, $t = -1.76$, $p = 0.081$). There was no effect of number and no interaction ($ts < 1$).

Region 3 (object NP). No main effect or interaction was significant (all $ts < 1$).

TABLE 1 | Example of item in the four experimental conditions of Experiment 1.

Relative clause	Match	Jérôme/parle/à la prisonnière/que/le gardien/sort/parfois/dans la cour. *Jérôme/speaks/to the prisoner-SG/that/the guard-SG/takes-SG out/sometimes/in the yard.*
	Mismatch	Jérôme/parle/aux prisonnières/que/le gardien/sort/parfois/dans la cour. *Jérôme/speaks/to the prisoners-PL/that/the guard-SG/takes-SG out/sometimes/in the yard.*
Sentence complement	Match	Jérôme/dit/à la prisonnière/que/le gardien/sort/parfois / dans la cour. *Jérôme/tells/the prisoner-SG/that/the guard-SG/goes-SG out/sometimes/in the yard.*
	Mismatch	Jérôme/dit/aux prisonnières/que/le gardien/sort/parfois/dans la cour. *Jérôme/tells/the prisoners-PL/that/the guard-SG/goes-SG out/sometimes/in the yard.*

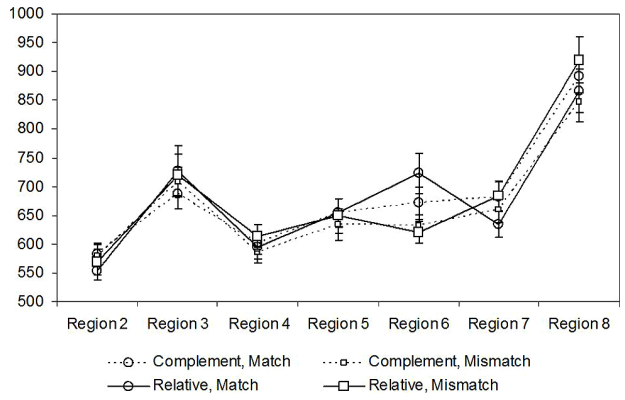


FIGURE 1 | Distribution of reading times (in ms) in the four experimental conditions of the different regions of Experiment 1. The critical region containing the test verb is Region 6.

Region 4 (complementizer). No main effect or interaction was significant (all $ts < 1$).

Region 5 (subject NP). No main effect or interaction was significant (all $ts < 1$).

Region 6 (target verb). A significant effect of number was found, with slower response times for singular objects (695 ms) than for plural objects (623 ms; $\beta = -0.078$, $t = -2.73$, $p = 0.006$). The main effect of structure was not significant ($t < 1$), but it entered into an interaction with number ($\beta = -0.013$, $t = -1.94$, $p = 0.053$). Subsequent models exploring the interaction showed that whereas number significantly affected reading times in the relative clause condition ($\beta = -0.115$, $t = -2.23$, $p = 0.027$), it failed to affect them in the complement clause condition ($t < 1$).

Region 7 (adverb). A trend toward slower response times in the complement clause condition (671 ms) than in the relative clause condition (657 ms) was found ($\beta = -0.029$, $t = -1.34$, $p = 0.180$). There was no effect of number and no interaction ($ts < 1$).

Region 8 (locative). A trend toward an interaction between number and structure was found ($\beta = 0.055$, $t = 1.17$, $p = 0.24$). There was no effect of number or structure ($ts < 1$).

Accuracy

The distribution of mean accuracy scores in the four experimental conditions is illustrated in **Figure 2**. Generalized linear mixed effect analysis showed that accuracy was significantly higher in the complement clause condition (0.97) than in the relative clause condition (0.89; $p < 0.001$; $\beta = -1.467$, $z = -3.89$, $p < 0.001$). The interaction and the number effect were not significant ($ts < 1$).

Discussion

Experiment 1 shows that the object's number feature influences processing speed at the verb in the object relative clause condition, but not in the corresponding complement clause condition, despite the surface similarity between the two structures. The observed effect shows that a feature mismatch between the extracted object and the subject of the relative clause speeds up reading of the verb segment. As mentioned in the Introduction, the experimentally elicited production of agreement is made more error-prone by the presence of a feature mismatch. Under the hypothesis that number mismatch influences the computation of agreement similarly in production and comprehension, one would have expected to find that it slows down sentence comprehension processes. Rather, it appears that a feature mismatch makes production of the relative clause verb error-prone, but reading of the relative clause verb faster. Why does a featural mismatch trigger opposite results in production and comprehension? Three observations suggest an answer, which is that the comprehension task did not tap into the mechanism of agreement computation, but rather in a mechanism of chain resolution responsible for linking a moved element to its trace.

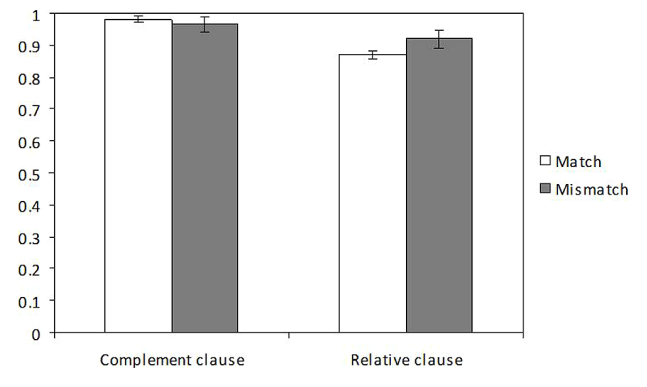


FIGURE 2 | Accuracy proportions in the comprehension questions of Experiment 1.

First, the present experiment shows an influence of the object's number in grammatical object relative clauses, which contrasts with the study by Wagers et al. (2009) who only found an effect in object relatives that contained an agreement error. Wagers et al. (2009) suggested that the interference effect they reported reflects reanalysis: only if the verb feature conflicts with the predicted feature would a cue-based retrieval process be deployed to actively retrieve the matching feature in the parsed tree. Our finding of object interference in the context of naturally reading a grammatical sentence cannot be accounted for by this view that the mismatch effect only arises as part of a second-pass process of agreement 'rechecking.'

Second, our finding that participants were faster in the presence of a plural feature mismatching the singular subject than in the presence of a singular feature matching the singular subject also contrasts with various comprehension studies showing a detrimental effect associated with a plural mismatching feature, whether it is on a subject modifier or a moved object, and whether comprehension is tested by way of self-paced reading or more indirect experimental procedures like maze tasks, classification tasks, grammaticality judgment or two-choice verb selection (Nicol et al., 1997; Clifton et al., 1999; Pearlmutter et al., 1999; Pearlmutter, 2000; Staub, 2009, 2010; Häussler, unpublished).

Third, although the direction of the interference effect in Experiment 1 may, at first glance, appear to be in line with that reported by Wagers et al. (2009) who also found faster reading times with plural mismatching objects, the two effects fundamentally differ since we found the effect in grammatical sentences whereas Wagers and colleagues found it in ungrammatical sentences. In the latter case, the parser is lured by the presence of a plural feature on the object creating an illusion of grammaticality. Hence, the finding that participants were faster in the plural mismatch condition also argues against a structure-based feature spreading mechanism like the one assumed to take place in production (Nicol et al., 1997; Vigliocco and Nicol, 1998; Franck et al., 2002; Eberhard et al., 2005). This raises the intriguing possibility that a different mechanism underlies the mismatch effect found here.

How do these different aspects of the data inform us about the mechanism underlying the number effect found in the present experiment? Directly relevant to the present study are the recent reports in the acquisition literature of intervention effects in the comprehension of object relatives. Adani et al. (2010) and Adani

(unpublished) found that both English and Italian speaking children showed better performance in a sentence-picture matching task when the object and the subject of the object relative clause mismatched in number (e.g., *Show me the elephant that the lions are washing* is better understood than *Show me the lion that the elephant is washing*). Using a similar task Belletti et al. (2012) reported better performance in Hebrew-speaking children for sentences involving a gender mismatch between the subject and the object relative. Empirical evidence suggests that only features that play an active role as triggers of syntactic movement have the potential to influence comprehension. This conclusion was reached on the basis of cross-linguistic evidence showing that in contrast to Hebrew children, Italian children failed to show a comparable sensitivity to gender mismatch in their comprehension of object relative clauses: this property was related to the different syntactic status of gender agreement in Italian.

According to the version of Relativized Minimality (Rizzi, 1990, 2004) assumed in the references quoted (along the lines developed in Friedmann et al., 2009), what makes the relevant kind of object relatives problematic for children is the intervention of the subject DP in the path connecting the relative head and its trace in object position. In particular, the difficulty is attributed to the set-theoretic relation of inclusion that characterizes the feature make-up of the object and that of the subject. When both the object and the subject are singular, the object is endowed with features [+R, +N, +Sg] (where +R is the feature designating the relative head) whereas the subject is endowed with features [+N, +Sg]: hence, the featural make-up of the intervener is included in the one of the antecedent. Friedmann et al. (2009) proposed that inclusion is problematic for children to explain their difficulty with making the required connection between the object and its trace. However, if the subject is plural, the number mismatch creates an intersection set, the object carrying [+R, +N, +Sg] and the subject [+N, +Pl]. Intersection is higher than inclusion in a natural scale of distinctness, a relation that is assumed to be accessible to the child's system in Belletti et al. (2012). Transposing this approach to the adult data collected in Experiment 1, the slowing down of reading time at the verb in the match condition as compared to the mismatch condition may be interpreted as an indication of the same gradation observed in children, inclusion being more difficult than intersection.

In this view, there is no contradiction between the number effect found in the production and comprehension studies of object relatives: whereas the production experiments directly tap into the agreement process, requiring the choice of a properly agreeing form, an operation which is penalized by the presence of a mismatching intervener in the immediate vicinity, the reading of the sentences primarily reflects the process of structure building, and in particular of the building of an appropriate A'-chain across an intervener, an operation which is enhanced by number mismatch. Hence the seemingly opposite consequences of mismatch in production and comprehension may be seen as a byproduct of the specific demands of the experimental tasks. If self-paced reading, as it is used in Experiment 1, mostly reflects the time taken by the parser to build the sentence structure and resolve the A'-dependency, it does not directly bear on agreement computation; one direct prediction of that account would be that the same effect of feature mismatch should be observed in sentences that do not involve an agreement configuration in their structure. We are currently exploring that possibility.

If self-paced reading does not tap into agreement processes, at least when complex sentences involving movement are involved, it may be relevant to identify a task that taps into the component of agreement processing in sentence comprehension. Experiment 2 uses a speeded grammaticality judgment task with sentences involving ungrammatical agreement, such that participants were forced to process agreement. If the same computational principles of agreement are at play in this task as in production, Experiment 2 should uncover the same structure-dependent attraction effects as found in sentence production.

Experiment 2: Object Attraction in Speeded Grammaticality Judgment

Experiment 2 tests the same experimental conditions as Experiment 1, contrasting object relatives and sentence complements, but this time with a speeded grammaticality judgment task in which agreement computation is explicitly assessed. Hence, in this task, agreement markers cannot be used for the structure building process; rather, agreement can only be computed once the hierarchical structure has been built. If the grammaticality judgment task allows tapping specifically into agreement processing in sentence comprehension, and if this process shows the signature of intervention effects as reported in sentence production, interference is expected to show up selectively in the condition where the object intervenes on the agreement dependency, i.e., in object relatives. In contrast, no interference is expected from the object of the main verb in sentence complements. Moreover, interference should take the shape of an attraction effect, with slower judgment times in the presence of plural mismatching objects.

Method

Participants

Thirty students of the University of Geneva, different from Experiment 1 and aged between 18 and 40, took part in the Experiment. They received credits for their participation. The experiment was approved by the ethics committee of the Department of Psychology of the University of Geneva, and informed consent was obtained from all participants.

Materials

Materials consisted of the same 24 test items as Experiment 1 without the last two windows, such that all sentences ended with the target verb. All sentences were ungrammatical with respect to subject-verb agreement: the verb in the subordinate clause was plural with a singular subject head noun. In addition to the test items, 120 filler items were built. Forty-eight of them were of the same structure as the experimental items; 16 correct with a singular subject (half with a singular object), 16 correct with a plural subject (half with a singular object) and 16 incorrect with a plural subject (half with a singular object). The 72 remaining

items had a different structure, with a subject modifier intervening linearly between the subject and the verb. Half of the modifiers consisted of subject relative clauses (e.g., Jean parle au gardien des bâtiments qui dort), the other half consisted of complement clauses (e.g., Jean dit que le programme des expériences fonctionne). Thirty-six (half) of these sentences were correct, the other half were incorrect. Half had a singular subject, the other half had a plural subject. Examples of test items are presented in **Table 2**.

Procedure

Materials were presented on a computer screen using the e-prime software. Sentences were split in windows corresponding to phrases (content word + grammatical word if present). Windows were presented for a fixed period of 500 ms, except at the verb, i.e., the final word of the sentence. These rather long presentation times were selected in order to minimize judgment errors, and avoid a possible trade-off between speed and accuracy. Grammaticality judgment times were measured at the verb onset. Participants were asked to judge the grammaticality of the sentences as quickly as possible and press on the corresponding response button. Pressing the button made the next window appear, such that a sustained rhythm was imposed.

Data Analyses

Incorrect grammaticality judgments representing 7.8% of the data were removed from the response times analyses. Analyses of response times were run both on the full dataset as well as on the data trimmed for outliers, defined as responses slower than 3 s (representing 7.9% of the data). Since both models provided similar outputs, the model of the complete data set is reported. Log-transformed response times and accuracy proportions were analyzed by way of (generalized) linear mixed-effects regression models with random intercepts for participants and items (Baayen et al., 2008), following the same procedure as for Experiment 1.

TABLE 2 | Example of an item in the four experimental conditions of Experiment 2.

Relative clause	Match	* Jérôme parle à la prisonnière que le gardien sortent. * *Jérôme speaks to the prisoner-SG that the guard-SG take out-PL.*
	Mismatch	* Jérôme parle aux prisonnières que le gardien sortent. * *Jérôme speaks to the prisoners-PL that the guard-SG take out-PL.*
Sentence complement	Match	* Jérôme dit à la prisonnière que le gardien sortent. * *Jérôme tells the prisoner-SG that the guard-SG go out-PL.*
	Mismatch	* Jérôme dit aux prisonnières que le gardien sortent. * *Jérôme tells the prisoners-PL that the guard-SG go out-PL.*

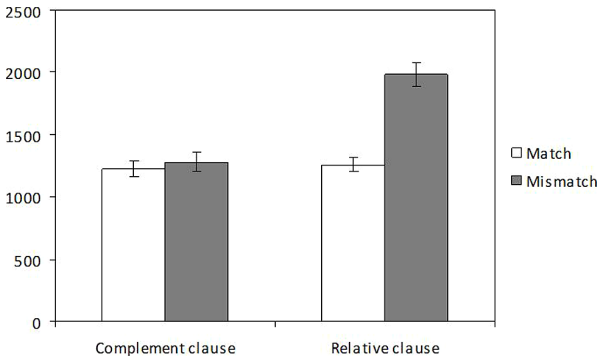


FIGURE 3 | Distribution of RTs (ms) for correct judgments at the target verb in the speeded grammaticality judgment task of Experiment 2.

Results

Response Times

The distribution of response times is illustrated in **Figure 3**. Mixed models revealed a main effect of number, with slower RTs with plural objects (1619 ms) than with singular ones (1242 ms; $\beta = 0.186$, $t = 4.449$, $p < 0.001$) as well as a main effect of structure with slower times for judging the grammaticality of object relatives (1609 ms) than for judging complement clauses (1253 ms; $\beta = 0.179$, $t = 4.277$, $p < 0.001$). The model showed a significant interaction between structure and number ($\beta = 0.273$, $t = 3.185$, $p = 0.002$): whereas number significantly affected response times in the relative clause condition ($\beta = 0.302$, $t = 4.463$, $p < 0.001$), no effect of number was found in the complement clause condition ($t < 1$).

Accuracy

Mean accuracy scores are reported in **Figure 4**. Accuracy was significantly affected by structure with better scores in the complement clause condition (0.97) than in the relative clause condition (0.90; $\beta = -1.379$, $z = -2.18$, $p = 0.03$). Number was marginally significant with higher scores for singular objects (0.97) than for plural ones (0.91; $\beta = -1.180$, $z = -1.863$, $p = 0.06$). The interaction was not significant ($t < 1$).

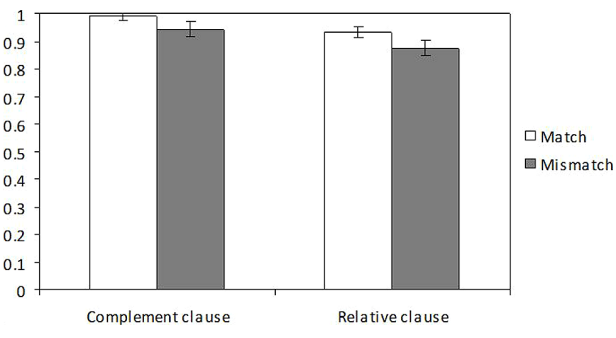


FIGURE 4 | Accuracy proportions in the speeded grammaticality judgment task of Experiment 2.

Discussion

In line with the production data reported on the same materials (Franck et al., 2010), participants were disturbed by the presence of a plural object in object relative clauses when performing a grammaticality judgment task bearing on the verbal agreement morphology: they were significantly slower to judge that the sentence was ungrammatical when the object was plural than when it was singular. By contrast, the plural feature in the sentence complement structure generated no or at least significantly reduced interference. The parallelism with production reports finds a natural explanation under the hypothesis that the same mechanism of agreement computation is at play in both tasks. This mechanism is sensitive to the hierarchical intervention of the intermediate trace of the object on the subject-verb dependency, which may have as processing consequence the local reactivation of the object, leading to interference in the processing of agreement, as argued in Franck et al. (2010). Under this hypothesis, the data provide evidence that agreement computation in sentence comprehension operates on the same syntactic representations as in sentence production (e.g., Nicol et al., 1997; Pearlmutter et al., 1999; Thornton and MacDonald, 2003; Hartsuiker, 2006; Badecker and Kuminiak, 2007).

Experiment 2 differs from Experiment 1 in the direction of the number effect: whereas faster response times were observed in the number mismatch condition in the self-paced reading Experiment 1, number mismatch slowed down grammaticality judgments in Experiment 2, in line with attraction effects found in sentence production. One could argue that the opposite direction of the effect found in Experiment 2 as compared to Experiment 1 is due to the fact that whereas Experiment 1 involved grammatical sentences, Experiment 2 involved ungrammatical sentences. Experiment 3 tests whether grammaticality affects the direction of the number interference effect in a grammaticality judgment task. If the effect reported in Experiment 2 merely reflects properties of syntactic representations, grammaticality should not affect performance since the same hierarchical structure underlies grammatical and ungrammatical sentences.

Experiment 3: The Role of C-Command

Findings in sentence production suggested that c-commanding interveners have a greater potential to trigger interference effects than preceding interveners (Franck et al., 2006, 2010).

Experiment 3 contrasts two conditions involving wh-movement of a complex objects, which potentially interferes with subject-verb agreement while transiting in a vP peripheral position, along the lines illustrated in the Introduction. The property that varies is where, in the complex object, the plural feature is expressed. In (3a) the DP head (hence, the whole object DP) is plural (*quelles patientes du médecin*), while in (3b) (*le chirurgien de quelles patientes*) only the embedded DP within a PP modifier is plural (here the embedded DP pied-pipes the whole object DP triggering its movement to the left periphery).

(3) a. Quelles **patientes** du médecin dis-tu que le juriste défend/*défendent?
*Which-PL patients-PL of the doctor do you say that the lawyer-SG defends-SG/*defend-PL?*

b. Le chirurgien de quelles **patientes** dis-tu que le juriste défend/*défendent?
*The surgeon of which-PL patients-PL do you say that the lawyer-SG defends-SG/*defend-PL?*
Which patients' surgeon do you say that the lawyer defends?

The crucial point here is that when the complex object DP transits through the vP peripheral position, intervening in the agreement process between the verbal inflection and the subject, the DP with plural marking intervenes in terms of c-command in (3a) (a hierarchical property), while it only intervenes in terms of precedence in (3b), where it is buried within the PP modifier. Under the hypothesis that the same guiding principles operate in sentence production and in the agreement checking process assumed to take place in grammaticality judgment, the plural feature on the c-commanding element 'patientes' in (3a) is expected to generate stronger attraction than the plural feature on the preceding element 'patientes' in (3b). While Experiment 1 tested grammatical sentences and Experiment 2 tested ungrammatical ones, Experiment 3 manipulates grammaticality in order to provide a systematic assessment of its role on agreement processing in grammaticality judgment.

Method

Participants

Twenty-six students of the University of Geneva different from Experiments 1 and 2 and aged between 18 and 40, took part in the Experiment. They received credits for their participation. The experiment was approved by the ethics committee of the Department of Psychology of the University of Geneva, and informed consent was obtained from all participants.

Materials

Materials consisted of 24 sets of 8 items. The variables manipulated include the number of the attractor (singular vs. plural), the position of the attractor with respect to the subject in its base position (c-command vs. precedence), and the grammaticality of subject-verb agreement (grammatical vs. ungrammatical). Questions, rather than declarative object relative clauses, were used to avoid attachment ambiguity (a relative clause could either be attached to the higher DP or to the DP embedded in the modifying PP). The position of the wh-marked element *quelles* was always on the target DP such that it was on the head in the c-commanding condition and on the PP modifier in the precedence condition. In this design, the plural DP was the wh-DP in both the c-command and precedence condition, so that the crucially varying DP would have the same role of wh-operator at Logical Form. As a result, the finding of an effect of structure would attest to a syntactic position effect, not of a semantic/logical form effect. All DPs were animate. An example of item across the eight experimental conditions is presented in **Table 3** (the full list of items is available in the Supplementary Materials).

Thirty-two filler grammatical items with the same structure as the test items but with plural subjects were created. An additional

TABLE 3 | Example of an item in the eight experimental conditions of Experiment 3.

C-command	Match	Quelle patiente du médecin dis-tu que le juriste défend/*défendent? *Which patient of the doctor do you say that the lawyer defends/*defend?*
	Mismatch	Quelles patientes du médecin dis-tu que le juriste défend/*défendent? *Which patients of the doctor do you say that the lawyer defends/*defend?*
Precedence	Match	Le chirurgien de quelle patiente dis-tu que le juriste défend/*défendent? The surgeon of which patient do you say that the lawyer defends/**defend*?
	Mismatch	Le chirurgien de quelles patientes dis-tu que le juriste défend/*défendent? The surgeon of which patients do you say that the lawyer defends/**defend*?

set of 40 fillers (half with singular subjects, half grammatical) was added. These consisted of object relatives (16), subject relatives (16) and PP modifiers in simple structures (8). Items were spread across four experimental lists that each contained 48 test items and the 72 fillers. Each list contained both the grammatical and the ungrammatical version of a test item, presented in two separate blocks with a short pause in between. Each block contained the same number of items in the eight conditions, presented in random order.

Procedure and Data Analyses

The same procedure and data analyses as Experiment 2 were adopted. Incorrect grammaticality judgments representing 17.3% of the data were removed from the response times analyses. Again, since the models with and without data trimming were identical, the models reported are those without data trimming. The number of incorrect judgments (216 data points) is small and their distribution is too complex to allow conclusions, nevertheless, the analysis is available in the Supplementary Materials.

Results

Response Times

The distribution of response times is illustrated in **Figure 5**. The model showed a main effect of number ($\beta = -100.73$, $t = -2.213$, $p = 0.027$), with slower response times in the condition with plural objects (766 ms) as compared to the condition with singular objects (658 ms). There was no effect of structure, and no effect of grammaticality ($ts < 1$). The predicted interaction between number and structure was significant ($\beta = 212.50$, $t = 2.341$, $p = 0.019$), and failed to interact with grammaticality, as attested by the non-significant three-way interaction ($t < 1$). Models run separately on each structure showed a significant effect of number of the c-commanding element ($\beta = -213.31$, $t = -3.173$, $p = 0.002$), with slower RTs for plural c-commanding elements (790 ms) than for singular ones (619 ms), but no significant effect of number of the preceding element (742 vs. 696 ms, $t < 1$). Grammaticality played no role in these models ($ts < 1$).

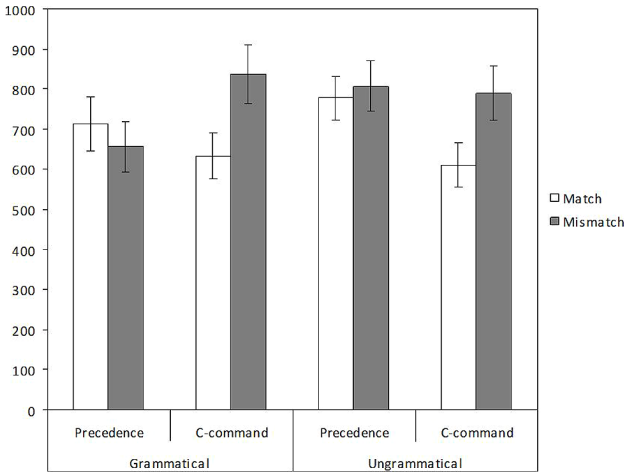


FIGURE 5 | Distribution of RTs (ms) for correct judgments in the speeded grammaticality judgment of Experiment 3.

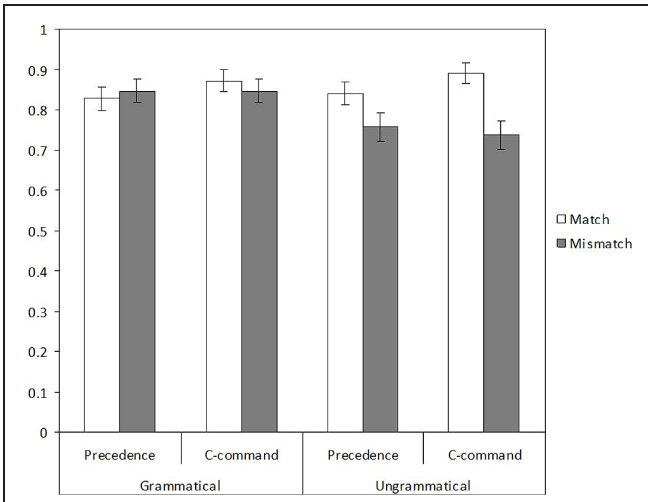


FIGURE 6 | Accuracy proportions in the speeded grammaticality judgment of Experiment 3.

Accuracy

Mean accuracy scores are reported in **Figure 6**. The model showed a main effect of number ($\beta = 0.442$, $z = 2.77$, $p = 0.006$), with lower accuracy rates with plural attractors (0.80) than with singular ones (0.86), as well as an interaction between number and grammaticality ($\beta = 0.817$, $z = 2.567$, $p = 0.010$). The interaction between number and structure failed to reach significance level ($\beta = -0.471$, $z = -1.48$, $p = 0.138$). Splitting the interaction into two separate models showed that whereas number significantly affected accuracy in ungrammatical sentences ($\beta = 0.817$, $z = 3.926$, $p < 0.001$), with better scores in the number match condition (0.87) than in the mismatch condition (0.75), it did not affect it in grammatical sentences ($z < 1$). Number and structure failed to interact significantly in the two models ($z < 1$).

Discussion

Experiment 3 brings two new findings. First, we found an effect of the structural variable manipulated: whereas the plural mismatching feature on the DP intervening in terms of c-command on agreement significantly contributed to slowing grammaticality judgments as compared to the singular matching feature, the plural feature on the DP intervening in terms of precedence failed to significantly affect response times. The finding that c-command intervention creates stronger interference than precedence intervention replicates previous reports in sentence production, with yet different constructions. Data from sentence production showed that the accusative clitic object pronoun situated pre-verbally creates more interference than a PP modifier situated in the same linear position (Franck et al., 2006) and more interference than the preverbal dative clitic (Franck et al., 2010). Both the PP modifier and the dative can be argued to intervene on the agreement dependency in terms of precedence, being embedded in a prepositional layer, whereas the accusative clitic intervenes in terms of c-command.

Second, grammaticality does not impact on performance in a speeded grammaticality judgment task: the same interference effect is found independently of whether the sentence is grammatical or not (see Häussler, unpublished, for a similar finding in German). This finding suggests that the differences in the direction of the number interference effect between Experiment 1 (self-paced reading), showing similarity-based interference, and Experiment 2 (speeded grammaticality judgment), showing attraction, is not due to the fact that that the former tested grammatical sentences whereas the latter tested ungrammatical sentences. Indeed, Experiment 3 shows that interference always shows up as an attraction effect in speeded grammaticality judgment. Hence, what seems critical is the process that the task taps into: whereas self-paced reading taps into the process of structure building and resolution of an A'-dependency, facilitated by the presence of a number mismatching feature on the subject intervening on the object-gap dependency, grammaticality judgment taps into the process of agreement checking, penalized by the presence of a number mismatching feature on the object intervening on the subject-verb dependency.

Finally, response times in Experiment 3 are faster than in Experiment 2, which showed particularly slow responses. The two experiments also differ in their overall error rates: in Experiment 2 where response times were between 1 and 2 s, the error rate was smaller than 10% overall; in contrast, the error rate in Experiment 3 where response times were between 600 and 850 ms was between 15 and 25%. It may therefore be the case that participants granted a privileged status to accuracy in Experiment 2. Nevertheless, it is important to note that there was no trade-off between response times and error rates across conditions in Experiment 3: conditions that were slower were also those that generated more errors.

General Discussion

We reported three studies exploring the consequences of a number featural mismatch in the comprehension of structures involving intervention configurations. The structures manipulated shared similar superficial characteristics, but critically differed in their hierarchical configurations. Experiments 1 and 2 contrasted object relatives clauses, involving an intermediate position created by movement of the object and intervening on the subject-verb agreement relation, and sentence complement clauses in which the object fails to intervene on the agreement relation at any point in the derivation of the hierarchical structure. Experiment 3 contrasted two structures involving complex objects also intervening on agreement in the object's intermediate position, but differing in the hierarchical position of the number mismatching feature situated either in a position of intervention in terms of c-command on agreement, or in terms of linear precedence.

The comparison between the first two experiments conducted on the same materials shows that self-paced reading (Experiment 1) and grammaticality judgment (Experiment 2) tap into distinct processes differently sensitive to intervention. The combination of the last two grammaticality judgment experiments illustrates the role of fine aspects of the hierarchical structure in agreement processing in sentence comprehension. Task-dependency and structure-dependency of number interference effects are discussed in turn.

Task-Dependent Interference in the Process of Structure Building

Experiment 1 using a self-paced reading procedure with grammatical sentences showed that participants read the verb significantly faster in the presence of a mismatching plural object in the relative clause, while no effect of number was found in the complement clause. Experiment 2 using a grammaticality judgment task specifically focusing on subject-verb agreement with the use of ungrammatical sentences also found an effect of the object's number restricted to relatives. However, this effect was reversed, with slower grammaticality judgments in the presence of a plural object, in line with attraction effects found in sentence production (Bock and Miller, 1991; Franck et al., 2006, 2010).

What is the mechanism underlying interference in the two experiments? We have suggested that both experiments reflect intervention effects on the hierarchical structure; however, the two experiments, because of the different techniques used, tap into two distinct processes, highlighting two different kinds of intervention effects. Experiment 1 reflects subject intervention on the object A'-dependency, Experiment 2 reflects object intervention on the subject-verb agreement dependency. More particularly, we have argued that self-paced reading taps into the process of structure building, in which the parser needs to resolve the required antecedent-gap dependency and assign the appropriate theta-roles to the arguments of the verb. The data of Experiment 1 are in line with recent developmental research attesting to children's better understanding of object relatives when the subject and object mismatch in number (Adani et al., 2010; Adani, unpublished). Similarly, mismatches in other features have also been found to facilitate object relative clause comprehension: gender mismatch (Belletti et al., 2012), animacy mismatch (e.g., Mak et al., 2002, 2006; Traxler et al., 2002) or mismatch in the

NP-type (DP, pronoun, proper name; e.g., Gordon et al., 2001, 2004; Warren and Gibson, 2002, 2005; Grillo, 2009; Belletti and Rizzi, 2013). Capitalizing on the theory of Relativized Minimality (Rizzi, 1990, 2004), Friedmann et al. (2009) suggested that the difficulty in building A'-dependencies in object relatives stems from the intervention of the subject DP in the path connecting the relative head and its trace in object position. Critically, the difficulty is hypothesized to be a function of the degree of overlap in syntactic features between the relative head and the intervener. According to this set-theoretic approach, the minimal degree of distinctness, identity, excludes the configuration from the grammar, while the maximal degree of distinctness, disjunction, makes the configuration fully accessible to both children and adults. The intermediate cases of inclusion and intersection would respectively engender stronger and weaker difficulty, the former manifesting itself in terms of the failure to build the A'-dependency in children and of a significant slowing down of processing in adults. In this framework, the facilitating role of number mismatch is captured in terms of the set theoretic relation in featural specification of the intervener with respect to the target.

The approach we have assumed expresses the intervention effect and the amelioration observed with feature mismatch directly in terms of a grammatical constraint, Relativized Minimality. Alternative approaches rooted in the psycholinguistic tradition do not appeal to a particular grammatical constraint and directly focus on the process of retrieving the object from memory when the verb is reached in parsing. Memory retrieval models assume that retrieval in long-distance dependencies involves a cue-based mechanism operating on content-addressable memory representations (e.g., McElree et al., 2003; Van Dyke and Lewis, 2003; Lewis and Vasishth, 2005; Van Dyke and McElree, 2006). These models grant a key role to similarity-based interference, which arises when memory units other than the retrieval target partially overlap with it in terms of their syntactic or semantic make-up. Although these models capture various interference effects reported in the literature (e.g., Lewis et al., 2006) only few attempts have tried to understand how the memory mechanisms posited can capture complex relational syntactic constraints (e.g., Vasishth et al., 2008; Dillon et al., 2013; Alcocer and Phillips, unpublished). One possible way in which our interpretation based on Relativized Minimality and more standard psycholinguistic approaches in terms of cue-based object retrieval may differ concerns the locus of the interference effect. If our interpretation is on the right track, the observed faster reading in the mismatch condition in sentences like (1a) has no direct relation with subject-verb agreement on the verb: it simply has to do with the resolution of an (object) A'-dependency across a partially matching intervener (the subject). If this is correct, we would expect the mismatch effect to enhance reading times at the verb (when the object trace is postulated and the A'-dependency resolved) even if the verb is uninflected, as for example in a sentence with a modal in English *(*e.g., *John talked to the patient(s) that the medicine can cure*), or in a sentence with an infinitival verb. If, as assumed in cue-based retrieval models, number on the verb serves as a linking address to memory units, the number effect should disappear with uninflected verbs. We intend to test this prediction in future work.

If a cue-based retrieval mechanism is at play in Experiment 1, it is, in any case, of a different type from the one assumed by Wagers et al. (2009) who tied it to a process of agreement 'rechecking,' triggered by the unpredicted number feature on the erroneously agreed verb. The number effect in Experiment 1 was found on grammatical sentences and in the verb region, while Wagers et al. (2009) found it in ungrammatical sentences and in the post-verbal region. These differences in the data suggest that if memory retrieval is responsible for the effect here, it must be tied to an early process of structure building and not to a late process of rechecking after the structure has been built, as proposed by Wagers et al. (2009). One could then wonder why Wagers et al. (2009) failed to find number interference in grammatical sentences in their work. The two studies differ in at least two respects. First, whereas Wagers et al. (2009) tested both grammatical and ungrammatical sentences, our materials only involved grammatical sentences. The presence of agreement errors in the English materials may have contributed to artificially disqualify number as a relevant cue to parsing, therefore explaining the lack of an effect in grammatical sentences. Second, our materials involved a mix of superficially similar object relatives and complement clauses; one cannot exclude the possibility that having to switch from one structure to the other increased the processing burden on structure building. The two factors may have played a cumulative role in the differences observed between the two studies.

The finding that the number effect in object relatives was reversed when measured in the grammaticality judgment task of Experiment 2, and turned into an attraction effect similar to the one found in sentence production, was taken as evidence that the task tapped into a different process. The grammaticality judgment task indeed forces the parser to first build the hierarchical structure over which agreement can be calculated, and can therefore be reasonably thought of as tapping into agreement computation proper. The finding that number interference arises as an attraction effect, and that the effect is restricted to object relatives and fails to manifest in complement clauses, suggests that the same mechanism underlies agreement computation in comprehension and production. In the next section, we describe our view of that mechanism.

Structure-Dependent Attraction in Agreement in Sentence Comprehension

In contrast to Experiment 1, Experiments 2 and 3 both showed that the presence of a number mismatching feature in the sentence significantly penalizes grammaticality judgments. Although Experiment 2 only tested ungrammatical sentences, Experiment 3 showed that the number effect arose independently of whether the sentence is grammatical or ungrammatical.

Results of Experiment 2 showed that attraction arises only in object relatives, when the attractor is the moved object of the target verb in the relative, but not when it is the object of the main verb in sentence complements, despite the superficial similarity of the two structures. Experiment 3 showed that a mismatching feature in a moved complex object intervening by c-command on an agreement configuration generates more attraction than one intervening by precedence. These two findings replicate our

previous reports in sentence production (Franck et al., 2006, 2010), arguing in favor of identical syntactic representations over which agreement takes place in production and comprehension (e.g., Nicol et al., 1997; Pearlmutter et al., 1999; Thornton and MacDonald, 2003; Hartsuiker, 2006; Badecker and Kuminiak, 2007).

What are the operating principles of agreement computation? In our production research, we suggested that attraction arises because of the intervention, on the subject-verb (AGREE) dependency, of the object transiting in its intermediate position at the periphery of the vP (Franck et al., 2006, 2010). Intervention by the intermediate object trace created by object movement, was argued to set the necessary condition for interference to arise. On that view, attraction results from the incorrect feature passing from the object to the verb via AGREE. One could nevertheless entertain a different scenario to account for the report of attraction in object relatives but not in sentence complements. A vast literature suggests that the parser reactivates the moved object when reaching the verb that it is an argument of (e.g., Stowe, 1986; McElree et al., 2003; Fedorenko et al., 2013). Hence, one cannot exclude the possibility that interference arises because the object is active during the same time window as the subject. Against this hypothesis, experimental evidence from sentence production shows that attraction arises even for moved objects that are not arguments of the target verb, as in (4).

(4) Voici les otages que le journaliste *apprennent qu'on a blessés. *Here are the hostages-PL that the journalist-SG *learn-PL that someone injured.*

Moreover, the strength of the attraction effect in this context is identical to attraction from the verbal argument tested in the context of object relatives (*John speaks to the patients that the medicine *cure*; Franck et al., 2010, Experiment 4). Hence, in order for a blind memory reactivation account to capture this report, one would need to assume that the parser reactivates all noun phrases from the parsed tree at the target verb to the same extent: *hostages* in (4) should be reactivated at the verb *learn*, of which it is not an argument, to a similar extent as the argument *patients* is reactivated at *cure* in the object relative clause. Even though retrieval is indeed known to be sensitive to interference from non-target elements sharing cues with the target (e.g., Van Dyke and McElree, 2006), it is marginal since, in the vast majority of the cases, the correct target is retrieved. Thus, a simple memory activation model fails to capture our finding that a moved object that is not part of the argument structure of the target verb triggers similar attraction to a moved verbal argument. The critical explanatory factor in interference rather appears to be the intervention of a moved DP in the hierarchical subject-verb dependency, a configuration which is identical whether the moved DP is an argument of the critical verb or not.

Results of Experiment 3 bring further support to our previous finding in agreement production that c-commanding interveners are more prone to trigger attraction than preceding ones. C-command has played a key role in syntactic theory, ever since work by Reinhart (unpublished), and has pervasive consequences on various morphosyntactic and interpretive processes like the binding of anaphors and the proper scope interpretation of quantifiers. The AGREE operation by which the subject's features are copied onto the agreement node in the functional layer of the clause also takes place under the constraint of c-command. Experiment 3 shows that an intervening object trace disrupts the processing of subject-verb agreement when a mismatching number feature intervenes between the subject and the verbal inflection in the hierarchical terms of c-command, and does so significantly more than when it intervenes in terms of mere precedence. This result parallels previous results on the stronger interference triggered by a c-commanding intervener in sentence production (Franck et al., 2010). In conclusion, both production and comprehension systems show a parallel sensitivity to the hierarchical relation of c-command, which thus has a central role both in grammar and performance.

Tanner et al. (2014) proposed that maintaining a unified account of agreement in production and comprehension minimally requires (1) that the same factors that modulate attraction in production also modulate attraction in comprehension, and (2) that interference in comprehension is symmetrical, as in production, meaning that attraction is expected to manifest independently of whether the sentence is grammatical or not in comprehension, or whether the correct verb form is ultimately chosen or not in production (a slowing down has been observed in the presence of a plural attractor in production even if correct agreement was used on the verb, Staub, 2009, 2010; Brehm and Bock, 2013). Results from the speeded grammaticality judgment Experiments 2 and 3 meet these requirements, suggesting that even though number interference effects may arise from different causes depending on the task used to measure sentence comprehension, the mechanism of agreement computation itself is the same in production and comprehension. This mechanism appears to operate under fine constraints as defined in formal syntax, including movement, intermediate traces and c-command.

Conclusion

The finding that the same structural effects as those found in sentence production are found in sentence comprehension is relevant both at the theoretical level and at the methodological levels. At the theoretical level, it argues in favor of a common syntactic component shared by production and comprehension, in spite of the obvious differences due to the intrinsically anticipatory nature of the parser. We suggested that the common component of agreement shows up when the comprehension task allows it to, as is the case when participants are required to judge the grammaticality of the sentence under time constraints. At the methodological level, grammaticality judgment is much easier to use than elicitation tasks, which often produce a very small range of errors that are problematic to analyze statistically. Speeded grammaticality judgment allows measuring not only errors but also response times, hence providing a finer measure necessary for subtle syntactic variables to show up in

an otherwise noisy performance. It therefore offers an ideal tool for the future exploration of the core syntactic components of agreement computation.

Acknowledgments

This work was supported by grant 100014-126924 from the Swiss National Fund for Scientific Research to Julie Franck. We wish to thank Brian Dillon, Akira Omaki, Whit Tabor and Matt Wagers for enriching discussions, and Maria Lucia Calí for data collection. We take complete responsibility for the content of the paper.

Supplementary Material

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fpsyg.2015.00349/abstract

References

Adani, F., van der Lely, H. K. J., Forgiarini, M., and Guasti, M. T. (2010). Grammatical feature dissimilarities make relative clauses easier: a comprehension study with Italian children. *Lingua* 120, 2148–2166. doi: 10.1016/j.lingua.2010.03.018

Baayen, R. H., Davidson, D. J., and Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *J. Mem. Lang.* 59, 390–412. doi: 10.1016/j.jml.2007.12.005

Badecker, W., and Kuminiak, F. (2007). Morphology, agreement and working memory retrieval in sentence production: evidence from gender and case in Slovak. *J. Mem. Lang.* 56, 65–85. doi: 10.1016/j.jml.2006.08.004

Belletti, A., Friedmann, N., Brunato, D., and Rizzi, L. (2012). Does gender make a difference? Comparing the effect of gender on children's comprehension of relative clauses in Hebrew and Italian. *Lingua* 122, 1053–1069. doi: 10.1016/j.lingua.2012.02.007

Belletti, A., and Rizzi, L. (2013). "Intervention in grammar and processing," in *From Grammar to Meaning*, eds I. Caponigro and C. Cecchetto (Cambridge: Cambridge University Press), 294–311. doi: 10.1017/CBO9781139519328.015

Bock, K., and Cutting, C. (1992). Regulating mental energy: performance units in language production. *J. Mem. Lang.* 31, 99–127. doi: 10.1016/0749-596X(92)90007-K

Bock, K., and Miller, C. (1991). Broken agreement. *Cogn. Psychol.* 23, 35–43. doi: 10.1016/0010-0285(91)90003-7

Brehm, L., and Bock, K. (2013). What counts in grammatical number agreement? *Cognition* 128, 149–169. doi: 10.1016/j.cognition.2013.03.009

Caplan, D. (2010). Task effects on BOLD signal correlates of implicit syntactic Processing. *Lang. Cogn. Process.* 25, 866–901. doi: 10.1080/01690961003672447

Caplan, D., Chen, E., and Waters, G. (2008). Task-dependent and task-independent neurovascular responses to syntactic processing. *Cortex* 44, 257–275. doi: 10.1016/j.cortex.2006.06.005

Chesi, C. (2005). *Phases and Complexity in Phrase Structure Building*. The 15th Meeting of Computational Linguistics in the Netherlands.

Chomsky, N. (2001). "Derivation by phase," in *Ken Hale: A Life in Language,* ed. M. Kenstowicz (Cambridge, MA: MIT Press), 1–52.

Chung, S. (1998). *The Design of Agreement: Evidence from Chamorro*. Chicago: Chicago University Press.

Clifton, C. Jr., Frazier, L., and Deevy, P. (1999). Feature manipulation in sentence comprehension. *Rivisti di Linguistica* 11, 11–39.

Dillon, B., Mishler, A., Slogett, S., and Phillips, C. (2013). Contrasting intrusion profiles for agreement and anaphora: experimental and modeling evidence. *J. Mem. Lang.* 69, 85–103. doi: 10.1016/j.jml.2013.04.003

Eberhard, K. M., Cutting, J. C., and Bock, K. (2005). Making syntax of sense: number agreement in sentence production. *Psychol. Rev.* 112, 531–559. doi: 10.1037/0033-295X.112.3.531

Fayol, M., Largy, P., and Lemaire, P. (1994). When cognitive overload enhances subject-verb agreement errors. A study in French written language. *Q. J. Exp. Psychol.* 47, 437–464. doi: 10.1080/14640749408401119

Fedorenko, E., Woodbury, R., and Gibson, E. (2013). Direct evidence of memory retrieval as a source of difficulty in long-distance structural dependencies in language. *Cogn. Sci.* 37, 378–394. doi: 10.1111/cogs.12021

Franck, J., Cronel-Ohayon, S., Chillier, L., Frauenfelder, U., Hamann, C., Rizzi, L., et al. (2004). Normal and pathological development of subject-verb agreement in speech production: a study on French children. *J. Neurolinguist.* 17, 147–180. doi: 10.1016/S0911-6044(03)00057-5

Franck, J., Lassi, G., Frauenfelder, U. H., and Rizzi, L. (2006). Agreement and movement: a syntactic analysis of attraction. *Cognition* 101, 173–216. doi: 10.1016/j.cognition.2005.10.003

Franck, J., Soare, G., Frauenfelder, U. H., and Rizzi, L. (2010). Object interference: the role of intermediate traces of movement. *J. Mem. Lang.* 62, 166–182. doi: 10.1016/j.jml.2009.11.001

Franck, J., Vigliocco, G., and Nicol, J. (2002). Subject–verb agreement errors in French and English: the role of syntactic hierarchy. *Lang. Cogn. Process.* 17, 371–404. doi: 10.1080/01690960143000254

Friedmann, N., Belletti, A., and Rizzi, L. (2009). Relativized relatives. Types of intervention in the acquisition of a-bar dependencies. *Lingua* 119, 67–88. doi: 10.1016/j.lingua.2008.09.002

Gibson, E., and Warren, T. (2004). Reading time evidence for intermediate linguistic structure in long-distance dependencies. *Syntax* 7, 55–78. doi: 10.1111/j.1368-0005.2004.00065.x

Gordon, P. C., Hendrick, R., and Johnson, M. (2001). Memory interference during language processing. *J. Exp. Psychol.* 27, 1411–1423.

Gordon, P. C., Hendrick, R., and Johnson, M. (2004). Effects of noun phrase type on sentence complexity. *J. Mem. Lang.* 51, 97–114. doi: 10.1016/j.jml.2004.02.003

Grillo, N. (2009). Generalized Minimality: feature impoverishment and comprehension deficits in agrammatism. *Lingua* 119, 1426–1443. doi: 10.1016/j.lingua.2008.04.003

Hale, J. (2001). "A probabilistic earley parser as a psycholinguistic model," in *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics,* Pittsburgh, PA, 159–166.

Hartsuiker, R. J. (2006). Are speech error patterns affected by a monitoring bias? *Lang. Cogn. Process.* 21, 856–891. doi: 10.1080/01690960600824138

Hartsuiker, R. J., Anton-Mendez, I., and Van Zee, M. (2001). Object attraction in subject-verb agreement construction. *J. Mem. Lang.* 45, 546–572. doi: 10.1006/jmla.2000.2787

Häussler, J. and Bader, M. (2009). Agreement checking and number attraction in sentence comprehension: Insights from German relative clauses. *Trav. Cercle Linguist. Prague* 7.

Häussler, J., Bader, M., and Bayer, J. (2003). "Number attraction in sentence comprehension," in *Poster Presented at the 9th Annual Conference on Architecture and Mechanisms for Language Processing*, eds R. Alterman and D. Krish, University of Glasgow, Glasgow.

Hemforth, B., and Konieczny, L. (2003). "Proximity in agreement errors," in *Proceedings of the 25th Annual Conference of the Cognitive Science Society*. Boston, MA: Cognitive Science Society.

Just, M. A., Carpenter, P. A., and Wooley, J. D. (1982). Paradigms and processes in reading comprehension. *J. Exp. Psychol. Gen.* 111, 228–238. doi: 10.1037/0096-3445.111.2.228

Kaan, E. (2002). Investigating the effects of distance and number interference in processing subject-verb dependencies: an ERP study. *J. Psycholinguist. Res.* 31, 165–193. doi: 10.1023/A:1014978917769

Kayne, R. (1989). Romance clitics, verb movement and PRO. *Linguist. Inquiry* 22, 647–686.

Kayne, R. (2000). *Parameters and Universals*. New York: Oxford University Press.

Kempen, G., Olsthoorn, N., and Sprenger, S. (2012). Grammatical workspace sharing during language production and language comprehension: evidence from grammatical multitasking. *Lang. Cogn. Process.* 27, 345–380. doi: 10.1080/01690965.2010.544583

Konieczny, L., Schimke, S., and Hemforth, B. (2004). "An activation-based model of agreement errors in production and comprehension," in *Proceedings of the 26th Annual Meeting of the Cognitive Science Society*, Chicago.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition* 106, 1126–1177. doi: 10.1016/j.cognition.2007.05.006

Lewis, R. L., and Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cogn. Sci.* 29, 1–45. doi: 10.1207/s15516709cog0000_25

Lewis, R. L., Vasishth, S., and Van Dyke, J. A. (2006). Computational principles of working memory in sentence comprehension. v*Trends Cogn. Sci.* 10, 447–454. doi: 10.1016/j.tics.2006.08.007

Mak, W. M., Vonk, W., and Schriefers, H. (2002). The influence of animacy on relative clause processing. *J. Mem. Lang.* 47, 50–68. doi: 10.1006/jmla.2001.2837

Mak, W. M., Vonk, W., and Schriefers, H. (2006). Animacy in processing relative clauses: the hikers that rocks crush. *J. Mem. Lang.* 54, 466–490. doi: 10.1016/j.jml.2006.01.001

McElree, B., Foraker, S., and Dyer, L. (2003). Memory structures that subserve sentence comprehension. *J. Mem. Lang.* 48, 67–91. doi: 10.1016/S0749-596X(02)00515-6

Nicol, J. L., Forster, K. I., and Veres, C. (1997). Subject-verb agreement processes in comprehension. *J. Mem. Lang.* 36, 569–587. doi: 10.1006/jmla.1996.2497

Pearlmutter, N. J. (2000). Linear versus hierarchical agreement feature processing in comprehension. *J. Psycholinguist. Res.* 29, 89–98. doi: 10.1023/A:1005128624716

Pearlmutter, N. J., Garnsey, S. M., and Bock, K. (1999). Agreement processes in sentence comprehension. *J. Mem. Lang.* 41, 427–456. doi: 10.1006/jmla.1999.2653

R Development Core Team. (2013). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rizzi, L. (1990). *Relativized Minimality*. Cambridge, MA: MIT Press.

Rizzi, L. (2004). "Locality and the left periphery," in *Structures and Beyond: The Cartography of Syntactic Structures*, ed. A. Belletti (Oxford: Oxford University Press), 223–251.

Salverda, A. P., Brown, M., and Tanenhaus, M. K. (2010). A goal-based perspective on eye movements in visual world studies. *Acta Psychol.* 137, 172–180. doi: 10.1016/j.actpsy.2010.09.010

Staub, A. (2009). On the interpretation of the number attraction effect: response time evidence. *J. Mem. Lang.* 60, 308–327. doi: 10.1016/j.jml.2008.11.002

Staub, A. (2010). Response time distributional evidence for distinct varieties of number attraction. *Cognition* 114, 447–454. doi: 10.1016/j.cognition.2009.11.003

Stowe, L. A. (1986). Parsing WH-constructions: evidence for on-line gap location. *Lang. Cogn. Process.* 1, 227–245. doi: 10.1080/01690968608407062

Tanner, D., Nicol, J., and Brehm, L. (2014). The time course of feature interference in agreement comprehension: multiple mechanisms and asymmetrical attraction. *J. Mem. Lang.* 76, 195–215. doi: 10.1016/j.jml.2014.07.003

Thornton, R., and MacDonald, M. C. (2003). Plausibility and grammatical agreement. *J. Mem. Lang.* 48, 740–759. doi: 10.1016/S0749-596X(03)00003-2

Traxler, M. J., Morris, R. K., and Seely, R. E. (2002). Processing subject and object relative clauses: evidence from eye movements. *J. Mem. Lang.* 47, 69–90. doi: 10.1006/jmla.2001.2836

Van Dyke, J. A., and Lewis, R. L. (2003). Distinguishing effects of structure and decay on attachment and repair: a retrieval interference theory of recovery from misanalyzed ambiguities. *J. Mem. Lang.* 49, 285–316. doi: 10.1016/S0749-596X(03)00081-0

Van Dyke, J. A., and McElree, B. (2006). Retrieval interference in sentence processing. *J. Mem. Lang.* 55, 157–166. doi: 10.1016/j.jml.2006.03.007

Vasishth, S., Lewis, R. L., Brüssow, S., and Drenhaus, H. (2008). Processing polarity: how the ungrammatical intrudes on the grammatical. C*ogn. Sci.* 32, 685–712. doi: 10.1080/03640210802066865

Vigliocco, G., and Nicol, J. (1998). Separating hierarchical relations and word order in language production. Is proximity concord syntactic or linear? *Cognition* 68, 13–29. doi: 10.1016/S0010-0277(98)00041-9

Wagers, M. W., Lau, E. F., and Phillips, C. (2009). Agreement attraction in comprehension: representations and processes. *J. Mem. Lang.* 61, 206–237. doi: 10.1016/j.jml.2009.04.002

Warren, T., and Gibson, E. (2002). The influence of referential processing on sentence complexity. *Cognition* 85, 79–112. doi: 10.1016/S0010-0277(02)00087-2

Warren, T., and Gibson, E. (2005). Effects of NP type in reading cleft sentences in English. *Lang. Cogn. Process.* 20, 751–767. doi: 10.1080/01690960500051055

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2015 Franck, Colonna and Rizzi. This is an open-access article distributed under the terms of the **Creative Commons Attribution License (CC BY).** *The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.*

frontiers in Psychology

GENERAL COMMENTARY
published: 09 June 2015
doi: 10.3389/fpsyg.2015.00807

Corrigendum: Task-dependency and structure-dependency in number interference effects in sentence comprehension

Julie Franck[1]*, *Saveria Colonna*[2] *and Luigi Rizzi*[3,4]

[1] Laboratoire de Psycholinguistique, University of Geneva, Geneva, Switzerland, [2] Centre National de la Recherche Scientifique – University of Paris 8, Paris, France, [3] Department of Linguistics, University of Geneva, Geneva, Switzerland, [4] Interdepartmental Centre for Cognitive Studies of Language, University of Siena, Siena, Italy

Keywords: number, agreement, attraction, intervention, intermediate traces, c-command, cue-based retrieval, comprehension

A corrigendum on

Task-dependency and structure-dependency in number interference effects in sentence comprehension

by Franck, J., Colonna, S., and Rizzi, L. (2015). Front. Psychol. 6:349. doi: 10.3389/fpsyg.2015.00349

The reference of the following sentence should be Adani et al. (2014) rather than Adani (unpublished).

"Adani et al. (2010) and Adani (unpublished) found that both English and Italian speaking children showed better performance in a sentence-picture matching task when the object and the subject of the object relative clause mismatched in number (e.g., Show me the elephant that the lions are washing is better understood than Show me the lion that the elephant is washing)."

Adani et al. (2014) should thus be added to the References.

References

Adani, F., Forgiarini, M., Guasti, M. T., and van der Lely, H. J. K. (2014). Number dissimilarities facilitate the comprehension of relative clause in children affected by (Grammatical) Specific Language Impairment. *J. Child Lang.* 41, 811–841. doi: 10.1017/S0305000913000184

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2015 Franck, Colonna and Rizzi. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

OPEN ACCESS

Edited and reviewed by:
Claudia Felser, University of Potsdam, Germany

***Correspondence:**
Julie Franck, julie.franck@unige.ch

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: *21 May 2015*
Accepted: *28 May 2015*
Published: *09 June 2015*

Citation:
Franck J, Colonna S and Rizzi L (2015) Corrigendum: Task-dependency and structure-dependency in number interference effects in sentence comprehension. Front. Psychol. 6:807. doi: 10.3389/fpsyg.2015.00807

frontiers in Psychology

ORIGINAL RESEARCH
published: 22 April 2016
doi: 10.3389/fpsyg.2016.00549

Syntactic Constraints and Individual Differences in Native and Non-Native Processing of Wh-Movement

Adrienne Johnson[1, 2, 3]*, *Robert Fiorentino*[2] *and Alison Gabriele*[3]

[1] Department of Education, Missouri Western State University, St. Joseph, MO, USA, [2] Neurolinguistics and Language Processing Laboratory, Department of Linguistics, University of Kansas, Lawrence, KS, USA, [3] Second Language Acquisition and Processing Laboratory, Department of Linguistics, University of Kansas, Lawrence, KS, USA

There is a debate as to whether second language (L2) learners show qualitatively similar processing profiles as native speakers or whether L2 learners are restricted in their ability to use syntactic information during online processing. In the realm of wh-dependency resolution, research has examined whether learners, similar to native speakers, attempt to resolve wh-dependencies in grammatically licensed contexts but avoid positing gaps in illicit contexts such as islands. Also at issue is whether the avoidance of gap filling in islands is due to adherence to syntactic constraints or whether islands simply present processing bottlenecks. One approach has been to examine the relationship between processing abilities and the establishment of wh-dependencies in islands. Grammatical accounts of islands do not predict such a relationship as the parser should simply not predict gaps in illicit contexts. In contrast, a pattern of results showing that individuals with more processing resources are better able to establish wh-dependencies in islands could conceivably be compatible with certain processing accounts. In a self-paced reading experiment which examines the processing of wh-dependencies, we address both questions, examining whether native English speakers and Korean learners of English show qualitatively similar patterns and whether there is a relationship between working memory, as measured by counting span and reading span, and processing in both island and non-island contexts. The results of the self-paced reading experiment suggest that learners can use syntactic information on the same timecourse as native speakers, showing qualitative similarity between the two groups. Results of regression analyses did not reveal a significant relationship between working memory and the establishment of wh-dependencies in islands but we did observe significant relationships between working memory and the processing of licit wh-dependencies. As the contexts in which these relationships emerged differed for learners and native speakers, our results call for further research examining individual differences in dependency resolution in both populations.

Keywords: wh-dependencies, individual differences, self-paced reading, second language processing, counting span, reading span, islands, working memory

OPEN ACCESS

Edited by:
Claudia Felser,
University of Potsdam, Germany

Reviewed by:
Robert Kluender,
University of California,
San Diego, USA
Akira Omaki,
Johns Hopkins University, USA

***Correspondence:**
Adrienne Johnson
ajohnson76@missouriwestern.edu

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: *16 July 2015*
Accepted: *01 April 2016*
Published: *22 April 2016*

Citation:
Johnson A, Fiorentino R and Gabriele A (2016) Syntactic Constraints and Individual Differences in Native and Non-Native Processing of Wh-Movement. Front. Psychol. 7:549. doi: 10.3389/fpsyg.2016.00549

INTRODUCTION

Research on the processing of wh-dependencies has found evidence that both native speakers and second language (L2) learners are able to utilize abstract syntactic information in the course of online processing (e.g., Aldwayan et al., 2010; Omaki and Schulz, 2011; Kim et al., 2015). The focus of these studies has been whether island constraints, which constrain the type of structures from which wh-extraction is possible (Ross, 1967; Chomsky, 1973, 1986), are respected in real time. For example, building on the seminal work of Stowe (1986), Aldwayan et al. (2010) examined whether L2 learners, similar to native speakers, would attempt to resolve wh-dependencies only in grammatically licensed positions: evidence of a reading time slowdown in (1b) at either the filled subject position (*Barbara*) or the filled object position (*us*) as compared to the same positions in the declarative sentence in (1a) would suggest that the L2 parser actively posits gaps in licit positions, while a lack of slowdown in the prepositional object position in (2b) as compared to (2a) would suggest avoidance of positing gaps within grammatically unlicensed positions, such as within the Complex Noun Phrase (NP) island (*the boring comments about John's used car*).

(1a) My brother asked if **Barbara** will photograph **us** beside Mom at the graduation.
(1b) My brother asked who **Barbara** will photograph **us** beside ___ at the graduation.
(2a) My sister wondered if the boring comments about **John's** used car were intended to entertain the group.
(2b) My sister wondered who the boring comments about **John's** used car were intended to entertain ___.

The results of a self-paced reading experiment with native speakers of English and Najdi Arabic learners of English indeed showed this pattern: there was a clear reading time slowdown or "filled-gap effect" for both learners and natives at the licit verbal object position (1a, 1b) but not at the prepositional object position within the complex NP island (2a, 2b). In a follow-up study, Canales (2012) revised the stimuli in (2), embedding the critical object within a relative clause island as in (3) so that the critical position in both the licit (1) and illicit (3) contexts followed a verb.

(3a) My brother questioned if the journalist that followed **Henry** last Saturday provoked the guard at the store.
(3b) My brother questioned who the journalist that followed **Henry** last Saturday provoked _____at the store.

Canales (2012) found converging evidence in a study testing Spanish-speaking learners of English, showing evidence of a filled-gap effect at the direct object position (*us*) in (1) but no difference in reading times at the direct object position within the relative clause island (*Henry*) in (3a,b). The presence of the object filled-gap effects across studies suggests that the lack of a reading time slowdown within the island conditions in (2) and (3) is not due to, for example, a lack of statistical power. Both Aldwayan et al. (2010) and Canales (2012) also found limited evidence of subject filled-gap effects (e.g., a reading time slowdown at *Barbara* in 1b as compared to 1a) in both experiments, suggesting that the parser can actively generate a prediction for a gap immediately following the wh-element. While these results were not consistent across experiments or participant groups, both native and learner groups showed evidence of subject filled-gap effects in at least one experiment in each study. The inconsistent emergence of subject filled-gap effects in these studies is not surprising as subject filled-gap effects did not emerge in Stowe's original study, testing English native speakers (see Stowe, 1986; Gibson et al., 1994; Lee, 2004; Johnson, 2015 for further discussion). Overall, the results of the studies discussed above suggest that the L2 parser is guided by syntactic constraints, attempting to resolve wh-dependencies only in licit positions. Using a different paradigm, Omaki and Schulz (2011) and Kim et al. (2015) also provide evidence that Spanish-speaking learners of English actively posit gaps in licit positions but avoid positing gaps in islands. These recent results are in line with several earlier, behavioral studies that showed that L2 learners at very high levels of proficiency are able to show native-like levels of performance on a grammaticality judgment task with respect to the rejection of ungrammatical island violations (e.g., Martohardjono, 1993; White and Juffs, 1998; see review in Belikova and White, 2009).

However, there is a debate as to whether islands are indeed grammatically unlicensed structures and are thus a relevant test case for investigating the recruitment of syntactic knowledge during processing or whether islands are simply processing bottlenecks (e.g., Kluender and Kutas, 1993; Hofmeister and Sag, 2010; Sprouse et al., 2012). It has been proposed that the parser may avoid positing gaps within islands, not due to adherence to syntactic constraints as was suggested above, but because the complex structure inherent to islands simply overwhelms an individual's processing capacities (Kluender and Kutas, 1993; Kluender, 2004; Hofmeister and Sag, 2010).

GRAMMATICAL VS. PROCESSING ACCOUNTS OF ISLANDS

Under grammatical accounts, gap-filling inside islands is avoided due to constraints on wh-extraction; under these views, both the avoidance of gap-filling in islands during sentence processing and the low acceptability ratings that island-violating sentences incur in acceptability judgment tasks are due to the utilization of syntactic knowledge (e.g., Sprouse et al., 2012). On the other hand, according to recent processing accounts, at least some islands are not the result of grammatical constraints (e.g., Hofmeister and Sag, 2010). Instead, the appearance of island sensitivity during processing and the elicitation of low ratings for island-violating sentences in judgment tasks are a consequence of processing pressures which are argued to increase difficulty in resolving wh-dependencies. On these accounts, island effects emerge when various processing burdens combine to render processing particularly difficult, leaving few resources to resolve dependencies. These processing burdens arise from a range of factors that are not unique to islands. They include the presence of the filler-gap dependency itself, which is argued to incur a processing cost that may increase as distance increases, the

processing of additional referents along the path between the filler and gap, processing clause boundaries, and how complex or semantically rich the filler phrase is, among other factors (e.g., Kluender, 1992, 1998, 2004; Hofmeister and Sag, 2010; Hofmeister et al., 2013; see also Cinque, 1990). Since these factors are hypothesized to lead to island effects, manipulating them in order to ease processing difficulty is expected to ameliorate or remove island effects. This expectation is arguably not shared by grammatical accounts, which hold that the parser should not predict a gap within an island context, regardless of such factors. In support of the processing accounts, several studies have put forth evidence that manipulating one or more non-structural factors leads to both improved judgments and reduced processing difficulty as indexed by self-paced reading times. For example, Hofmeister and Sag (2010, Experiment 2) used self-paced reading to investigate whether the processing difficulty and low acceptability ratings of wh-islands such as (4b) below would be ameliorated by simply replacing a bare wh-element (*Who*) with a more complex, semantically rich wh-phrase (*Which employee*). Participants first read a lead-in sentence like (4a) below, and then a second sentence with either a bare wh-phrase in a wh-island construction (4b), a which phrase in a wh-island construction (4c), or a baseline condition involving no island violation (4d).

(4a) Albert learned that the managers dismissed the employee with poor sales after the annual performance review.
(4b) Bare Wh-phrase: Who did Albert learn whether they dismissed after the annual performance review?
(4c) Which phrase: Which employee did Albert learn whether they dismissed after the annual performance review?
(4d) Grammatical Baseline: Who did Albert learn that they dismissed after the annual performance review?

Reading times at the three regions following the embedded verb (*dismissed* in 4 above) were significantly faster for the Which phrase condition (4c) than the Bare Wh-phrase condition (4b); indeed, reading times for the Which phrase condition did not differ from the grammatical baseline condition (4d). These results suggest that manipulation of the semantic complexity of the filler phrase may reduce the difficulty of processing the wh-island. In a follow-up where bare wh- and which phrase sentences (presented in embedded questions) were rated for acceptability, the which-phrase sentences received higher acceptability ratings. That the manipulation of this factor both eased processing difficulty following the gap site and improved acceptability ratings was taken to suggest that processing pressures contribute to island effects.

Individual differences in processing resources (in particular, working memory) constitute another factor that may affect whether island effects emerge (e.g., Kluender, 1992; Hofmeister et al., 2012a,b). As Hofmeister and Sag (2010) point out, "Notably, some individuals seem fairly accepting of island violations, while others reject the same tokens. This type of variation in acceptability judgments, both within and across subjects emerges naturally on the processing account of islands. Individuals are known to differ significantly from one another in terms of working memory capacity (Daneman and Carpenter, 1980; King and Just, 1991; Just and Carpenter, 1992), and the same individual may have more or fewer resources available, depending upon factors such as fatigue, distractions, or other concurrent tasks" (p. 403). However, in cases of extreme processing difficulty, such individual differences may not emerge (e.g., Hofmeister et al., 2014).

In a series of large-scales studies, Sprouse et al. (2012) examined whether individual differences in processing resources modulate the acceptability of islands using off-line acceptability rating tasks and two measures of working memory capacity. The acceptability judgment tasks, testing four types of islands, used a factorial design manipulating the presence or absence of an island structure, as well as the position of the gap (in the matrix clause or in the embedded clause) as in (5). Sprouse et al. examined whether the combined effect of the presence of an island structure and extraction as in (5d) was "superadditive," yielding lower acceptability ratings than would be expected by the addition of the two individual factors.

(5a) Who ____ claimed that John bought the car?
NON-ISLAND/MATRIX
(5b) What did you claim that John bought ____?
NON-ISLAND/EMBEDDED
(5c) Who ___ made the claim that John bought a car?
ISLAND/MATRIX
(5d) *What did you make the claim that John bought ____?
ISLAND/EMBEDDED

According to Sprouse et al. (2012), processing accounts should further predict that effects of superadditivity or sensitivity to island violations would be reduced in those with superior processing resources. Grammatical accounts predict no such relationship. Sprouse et al. (2012) argue that their results showed no meaningful relationship between working memory and the "superadditive" effect on acceptability judgments that they observed, taking these findings to support the grammatical accounts of islands.

However, in response, Hofmeister et al. (2012a,b) point out that the lack of a relationship between acceptability ratings and processing resources in Sprouse et al. (2012) could be due to the nature of the tasks used. Hofmeister et al. argue that the stimuli tested, which included decontextualized questions with bare wh-fillers may have been particularly difficult to process, not allowing the variability in acceptability judgments that would allow a correlation to emerge. They also claim that the working memory tasks (n-back and serial recall) used in Sprouse et al. (2012) may assess short term memory, as opposed to working memory (see also Conway et al., 2005), and have not been shown to capture variability in sentence processing in other contexts.

Aldosari (2015) modified Sprouse et al.'s stimuli in order to address some of the concerns raised by Hofmeister et al. (2012a,b). In the acceptability judgment task, Aldosari included a context sentence which preceded the wh-question so as to not present decontextualized questions. The wh-questions themselves were revised to include lexical wh-fillers as opposed to bare wh-words (Hofmeister and Sag, 2010; see also Goodall,

2015). The goal of these revisions was to potentially decrease the processing difficulty in order to allow more room for variability to emerge in the judgments. Finally, a complex span task (operation span) was used. The results for both native speakers and Najdi Arabic learners of English showed, in line with Sprouse et al., clear effects of superadditivity, with low acceptance of the island sentences. In addition, the results revealed no significant relationship between working memory and sensitivity to island violations for either native speakers or L2 learners.

In addition to these recent studies examining the relationship between processing abilities and acceptability judgments, previous studies have also utilized reading-time measures in order to test whether islands can be indeed be explained via processing limitations, without recourse to grammatical constraints. The approach in these studies (e.g., Phillips, 2006; Wagers and Phillips, 2009) has been to examine whether gaps are posited in linguistic contexts that typically constitute islands, but can under some circumstances be rescued by later material in the sentence. For example, extraction from complex subjects (e.g., 6a below) is typically prohibited; however, extraction from subject islands is acceptable in "parasitic gap" constructions in which the wh-element is associated with two different gaps, one within the subject island and a second, object gap, which can "rescue" the violation (as is shown in 6b). Note, however, that a second gap cannot rescue the first if the verb is finite, as in (6c).

(6a) *What did [the plan to build *t*] ultimately destroy the house?
(6b) What did [the plan to build *t*] ultimately destroy *t* ?
(6c) *What did [the group that built *t*] ultimately destroy *t* ?

Phillips (2006) reasoned that, if the subject island in (6) results from processing pressures which simply make it too difficult to resolve the dependency there, then the possibility that extraction from that position may be rescued by the presence of a subsequent gap (as is true in non-finite structures like 6b) should not matter; a gap should never be posited in that position.

Phillips (2006) provided reading-time evidence that a gap was indeed posited within a subject island when the structure was non-finite, and thus potentially rescuable by a subsequent gap, but not when the structure was finite (see Ross, 1967). These results were taken to be consistent with grammatical accounts of islands. However, Hofmeister and colleagues challenge this interpretation, pointing out that, under their view, islands are positions that are difficult rather than impossible to extract from, and that factors like verb finiteness may indeed modulate how difficult it is to process the clause, thus rendering gap filling within the subject island more vs. less likely (Kluender, 2004; Hofmeister et al., 2013).

Using eye-tracking, Boxell and Felser (2013) replicated the results of Phillips (2006) with a group of native speakers but showed a somewhat different pattern for native German learners of English. According to first-pass reading measures, while native speakers posited gaps in islands only when such gaps might ultimately be rescued, L2 learners initially posited gaps in islands across the board (see also Kim et al., 2015). The L2 learners did however show a native-like pattern at the critical region in rereading time, a measure that includes all fixations in a region after it has been exited.

The distinct pattern that emerges in the early reading measures for the native speakers and L2 learners leads Boxell and Felser (2013) to propose that L2 processing differs significantly from native processing: while native speakers are immediately constrained by island restrictions, L2 learners' sensitivity to island constraints is delayed. A recent study by Felser et al. (2012) also argues that native and non-native processing differ in terms of the type of information that is prioritized at different stages of processing. Felser et al. (2012) conducted two experiments, one with a filled-gap paradigm and the other with a plausibility mismatch paradigm, both examining whether learners and natives would attempt to resolve wh-dependencies in non-island contexts but avoid positing gaps in relative clause islands. In the filled-gap experiment, a filled-gap effect emerged for natives at the critical region and for learners at the spillover region. Neither group attempted to resolve wh-dependencies in islands. In the plausibility mismatch experiment, it was the L2 learners who showed an immediate plausibility mismatch effect only for the non-island structures; the same effect for natives emerged in re-reading measures, also at the critical region. The results of the Boxell and Felser (2013) and Felser et al. (2012) study differ critically in that the learners in the Felser et al. study do not attempt to resolve wh-dependencies in islands at any point while the results of Boxell and Felser (2013) suggest an initial insensitivity to islands. Boxell and Felser speculate that differences in the processing complexity of the two different types of islands (subject islands vs. relative clause islands) may account for the differences in the two studies. The present study will further address whether L2 learners demonstrate island sensitivity similarly to native speakers; in addition, we bring together two strands of research discussed above by examining whether there is a relationship between individual differences and the online processing of wh-dependencies in both island and non-island contexts.

PRESENT STUDY

In the current study, we examine the relationship between working memory and filled-gap effects in both native speakers and L2 learners[1]. To our knowledge, no previous study has directly examined the relationship between processing abilities and filled-gap effects in islands, which provide an online measure of the processing of wh-dependencies. However, this approach may be advantageous as it is possible that offline measures of acceptability do not capture variability that may emerge in the course of processing the island itself. Grammatical accounts do not predict a relationship between working memory and the establishment of wh-dependencies in island contexts as the parser should simply not attempt to resolve dependencies in

[1]A reviewer suggests that it would have been beneficial to include an offline measure of acceptability. Note that this would primarily be a concern if learners do show filled-gap effects inside islands. However, even if an offline grammaticality judgment task had been included, native-like performance on this type of task would not necessarily indicate a native-like grammar of wh-dependencies (see Aldwayan et al., 2010).

grammatically unlicensed positions. In contrast, a processing account such as that proposed by Hofmeister and colleagues predicts that working memory and island sensitivity may be related (e.g., Hofmeister and Sag, 2010). Thus, as suggested by Sprouse et al. (2012), results showing that individuals with better processing abilities are better able to establish wh-dependencies in complex structures such as islands would be consistent with this kind of account, a claim that Hofmeister et al. (2012b) acknowledge to be broadly in line with their proposal.

On the other hand, finding a relationship between working memory and the processing of grammatically licensed wh-dependencies would be consistent with both proposals. One possibility is that lower working memory may lead to greater filled-gap effects in licit positions. A number of models highlight effects of distance on dependency resolution, pointing out that the resolution of wh-dependencies becomes more difficult at a greater distance (e.g., Gibson, 1998) although the reasons for these effects and the specific circumstances under which increased distance indeed engenders processing burden remain a matter of investigation (e.g., Wagers and Phillips, 2014; Nicenboim et al., 2015). Considering that wh-dependency resolution may generally become more burdensome as distance increases (thus leading the parser to resolve the dependency as soon as possible; e.g., Frazier, 1987), perhaps those participants with low working memory will show greater eagerness to quickly resolve the dependency, and thus yield greater evidence of active gap-filling than those with high working memory.

However, there is also reason to speculate that higher working memory would lead to greater filled gap-effects in licit positions. Resolving wh-dependencies involves a range of processes, from initially encoding the wh-dependency, which is argued to involve generating predictions for upcoming gap sites in advance of unambiguous bottom-up evidence (e.g., Nakano et al., 2002; Lee, 2004; Omaki et al., 2015), to maintaining and/or retrieving dependency-related information while also processing bottom-up information, monitoring for conflicts among expected and encountered material, and ultimately resolving the dependency. All of these processes have been argued to make recourse to working memory or other resources related to attentional control (e.g., Daneman and Carpenter, 1983; Engle, 2002; Hutchison, 2007; Slevc and Novick, 2013). It may thus be those with greater resources who are more likely to successfully engage these processes.

Some evidence suggesting that higher working memory may lead to greater gap-filling effects comes from Nakano et al. (2002), who examined pre-verbal gap filling in Japanese using the cross-modal lexical priming paradigm. Nakano et al. examined whether evidence for pre-verbal gap filling depended on working memory, finding that only those participants with high working memory showed evidence of pre-verbal trace reactivation (see also Roberts et al., 2007). While there remains a paucity of studies directly examining individual differences in working memory/attentional control in wh-dependency resolution (Nicenboim et al., 2015), the above evidence is consistent with the prediction that those with higher working memory may show greater filled-gap effects in licit positions.

Examples of the target stimuli in our experiment, which were adapted from Canales (2012), are given below in (7) and (8). Our first comparison involves sentences that do not contain an island structure, the *Non-Island* sentences in (7a-b). The comparison of reading times for *Non-Island* sentences that do (7a) and do not (7b) involve wh-extraction allows us to probe for filled-gap effects in licit, filled subject (*Chris*) and filled object (*Tom*) positions. Our second comparison involves sentences that contain a relative clause island, the *Island* sentences in (8a,b). The comparison of *Island* sentences that do (8a) and do not (8b) involve wh-extraction allows us to probe for filled-gap effects both in the licit filled subject position (*the actress*) and a filled object position within the relative clause island (*Tyler*).

Non-Island, No extraction

(7a) The instructor wondered if **Chris** will film **Tom** with Susan at the reception.

Non-Island, Wh-extraction

(7b) The instructor wondered who **Chris** will film **Tom** with ____ at the reception.

Island, No extraction

(8a) My father asked if **the** actress that married **Tyler** last summer kissed the director during$_{15}$ the$_{16}$ rehearsal$_{17}$.

Island, Wh-extraction

(8b) My father asked who **the** actress that married **Tyler** last summer kissed ____ during the rehearsal.

The present study examines both native speakers and native Korean learners of English in order to better understand the nature of the processing of wh-dependencies in both native and learner populations. Previous studies have shown that Korean learners may not abide by island constraints during online processing (Kim et al., 2015), which they suggest may be due to the fact that Korean is a wh-*in situ* language which does not exhibit overt wh-movement (Sohn, 1980, 1999). However, as Kim et al. (2015) acknowledge, some recent papers have suggested that wh-*in situ* languages, such as Korean, do block extraction from relative clauses, just as in English (Han and Kim, 2004; Phillips, 2013). Our previous work with Najdi Arabic learners of English has also shown that is possible for native speakers of a wh-*in situ* language to abide by island constraints during processing. Thus, we include both native speakers and Korean learners of English to compare native and non-native processing broadly, but not necessarily to examine potential effects of L1 transfer.

The present study examines whether native speakers and learners show qualitatively similar patterns, as has been shown in some studies (e.g., Aldwayan et al., 2010; Omaki and Schulz, 2011), or whether learners are unable to use syntactic information on the same timecourse as native speakers (Felser et al., 2012; Boxell and Felser, 2013). If the two groups show qualitatively similar patterns, a filled-gap effect should emerge at the grammatically licensed direct object position in the *Non-Island* sentences (*Tom* in 7b) but not within the relative clause island in the *Island* sentences (*Tyler* in 8b) for both groups. In contrast, if learners are unable to prioritize syntactic information and use it in the earliest stages of processing (Felser et al., 2012), then learners should either show filled-gap effects at the direct

object positions in both the *Non-Island* and *Island* sentences, suggesting an attempt to resolve wh-dependencies within islands (Boxell and Felser, 2013; Kim et al., 2015) or they should show sensitivity to island contexts only at a delay. With respect to the second possibility, learners may, for example, pattern similarly to native speakers, positing a gap in (7b) and avoiding positing a gap within the island (no difference between 8b) but this pattern should emerge on a different timecourse from native speakers, perhaps emerging at a region later in the sentence as has been observed in previous studies (e.g., Felser et al., 2012). We will also examine effects at the licit filled subject positions in both Non-Island and Island sentences (*Chris* in 7b; *the actress* in 8b). However, as discussed above, the inconsistency of subject filled-gap effects in both native speakers and L2 learners in previous experiments using this same design (Stowe, 1986; Lee, 2004; Aldwayan et al., 2010; Canales, 2012) does not allow us to make strong predictions regarding similarities and differences between learners and native speakers. We will return to this issue in the discussion.

The study also examines the nature of islands, investigating whether there is a relationship between working memory and the processing of wh-dependencies in islands. No such relationship is predicted by the grammatical accounts. A positive relationship between working memory and the size of the filled-gap effect in the object position within the relative clause island (*Tyler* in 8a,b) would be consistent with Hofmeister and colleagues' versions of the processing account (e.g., Hofmeister et al. 2012a,b, 2014). Any significant relationships that emerge between working memory and the grammatically licensed potential gap sites (subject positions in 7b and 8b, object position in 7b) would be consistent with both proposals[2].

MATERIALS AND METHODS

Participants

Forty-nine advanced Korean learners of English and 54 native English speakers participated in the study. The Korean learners (mean age = 28.41; 28 females) were recruited from the University of Kansas and its surrounding community; their mean age of arrival was 22.89 years old. All learners reported no significant exposure to English before age of 12, and no learner reported significant exposure to any wh-movement language other than English. The learners' English proficiency was assessed using the University of Michigan Listening Comprehension Test, a 45 question test which covers various aspects of English grammar (mean proficiency score = 39.39). Eight additional Korean learners and eight additional native English speakers also participated in the study, but were identified as outliers with respect to magnitude of their filled-gap effects and excluded from the final analysis of the data, and one additional English speaker also participated but was excluded from the final analysis for showing exceptionally fast reading times (faster than 250 ms) across regions, as described in the *Data Analysis* section below. The Korean learners of English were provided with payment for their participation, and the native English speakers (mean age = 21.15; 41 females), who were all students at the University of Kansas, completed the study for extra credit. This study was approved by the Institutional Review Board of the University of Kansas and all participants provided their written informed consent before participating.

Stimuli

Non-Island stimuli

The *Non-Island* stimuli included 20 pairs of sentences, with each pair consisting of a control sentence with no extraction (9a) and a matched wh-extraction sentence (9b); the region number for each word is indicated by the subscripts in (9). A full list of stimuli is provided as Supplementary Material.

Non-Island, No extraction

(9a) The$_1$ instructor$_2$ wondered$_3$ if$_4$ **Chris**$_5$ will$_6$ film$_7$ **Tom**$_8$ with$_9$ Susan$_{10}$ at$_{11}$ the$_{12}$ reception$_{13}$.

Non-Island, Wh-extraction

(9b) The$_1$ instructor$_2$ wondered$_3$ who$_4$ **Chris**$_5$ will$_6$ film$_7$ **Tom**$_8$ with$_9$ _____$_{10}$ at$_{11}$ the$_{12}$ reception$_{13}$.

The wh-structure in (9b) involves extraction from the grammatically licit prepositional object position (region 10). Preceding this position are two grammatically licit potential gap positions that are filled with lexical material: the embedded subject position (region 5) which is filled with the subject *Chris*, and the post-verbal direct object position (region 8) which is filled with the object *Tom*; these positions are bolded in example (9) above. These two regions and their spillover regions (region 6 and 9, respectively) serve as critical regions to test for filled-gap effects in positions from which wh-extraction is grammatically licit.

The embedded verbs used in region 7 were all transitive verbs. The prepositional objects (region 8) were all proper names that were three letters long, and the embedded subjects (region 5) were all proper names as well (mean length = 5.4 letters, range 4–11 letters).

Island stimuli

The *Island* stimuli included 20 additional pairs of sentences, with each pair consisting of a control sentence with no extraction (10a) and a matched wh-extraction sentence (10b).

Island, No extraction

(10a) My$_1$ father$_2$ asked$_3$ if$_4$ **the**$_5$ actress$_6$ that$_7$ married$_8$ **Tyler**$_9$ last$_{10}$ summer$_{11}$ kissed$_{12}$ the$_{13}$ director$_{14}$ during$_{15}$ the$_{16}$ rehearsal$_{17}$.

Island, Wh-extraction

[2]A reviewer asked us to specify the predictions for the relationship between working memory and the size of the filled-gap effect if the Korean learners of English do not in fact show knowledge of syntactic constraints. If the Korean learners of English do not pattern similarly to the native speakers, then the specific nature of the relationship between working memory and the size of the filled-gap effect would need to be examined. For example, if Korean learners of English are found to establish wh-dependencies in islands, and further, if those learners with higher working memory resources showed larger filled-gap effects within islands, then the results would support a processing account. Note, however that, in our results, the Korean learners of English do show similar island sensitivity to the native speakers.

(10b) My$_1$ father$_2$ asked$_3$ who$_4$ **the**$_5$ actress$_6$ that$_7$ married$_8$ **Tyler**$_9$ last$_{10}$ summer$_{11}$ kissed$_{12}$ ____$_{13-14}$ during$_{15}$ the$_{16}$ rehearsal$_{17}$.

The wh-structure in (10b) involves extraction from the grammatically licit object position (regions 13–14). Crucially, preceding this position is a relative clause island, from which wh-extraction is illicit. While the relative island contains a post-verbal object position (region 9) filled with a proper name (*Tyler*, in 10b above), extraction from this position is not grammatically licensed. Thus, region 9 and its spillover region (region 10) serve as critical regions to probe for filled-gap effects in a grammatically illicit position (within a relative clause island). Region 5 and its spillover region (region 6) constitute the filled embedded subject position, a grammatically licit site for extraction. Like the embedded subject position in the *Non-Island* sentences, the embedded subject region 5 and its spillover region (region 6) serve as critical regions to test for filled-gap effects in the grammatically licit, subject position.

The verbs inside the relative clause island (region 8) were all transitive verbs. The post-verbal object position within the island (region 9) was always filled with a proper name that was five letters long. An adverbial phrase (e.g., *last summer*) always followed region 9 (e.g., *Tyler*) in order to provide a spillover region following the post-verbal object position that would precede the verb that licenses the actual gap position in the wh-extraction sentence (e.g., *kissed*). The embedded subject position from which extraction is grammatically licensed (region 5 and its spillover region, region 6) were comprised of a determiner-noun combination; the determiner in region 5 was always the three-letter-long determiner "the."

The 20 the *Non-Island* sentences, the 20 *Island* sentences, and 80 filler sentences were presented together, yielding a 1:2 target-to-filler ratio. Two Latin-square lists were created, such that every participant was presented with either the extraction or no-extraction version of every sentence, but no participant read more than one version of a given sentence. The sentences were presented in different randomized order for each participant.

Procedure

All participants completed a background questionnaire, the self-paced reading task, and then two working memory tasks (the reading span task and the counting span task) which are described below; the order of the two working memory tasks was counterbalanced across participants. Korean learners of English also completed the University of Michigan Listening Comprehension Test (1972), after completing all other tasks. The self-paced reading task, working memory tasks and proficiency test were all administered using *Paradigm* presentation software (Tagliaferri, 2005).

Self-Paced Reading Task

Each sentence was presented word-by-word in a non-cumulative moving window self-paced reading paradigm (Just et al., 1982). At the beginning of each trial, each word of the sentence was masked by a series of dashes; this masking included words and punctuation, but did not include the spaces between words. Each time the participant clicked a mouse button to advance through the sentence, the next word was unmasked, and the previous word was masked again. After the last word of each sentence, the sentence was then presented again in full, but with one word missing (e.g., "My _____ asked if the actress that married Tyler last summer kissed the director during the rehearsal."). Participants selected the missing word from among two options (e.g., "father" and "sister") which were presented on the screen, by pressing the appropriate key on the computer keyboard (either the key labeled "L" for the word on the left of the screen, or that labeled "R" for the word on the right of the screen). Prior to the experiment, participants completed a practice session consisting of five practice sentences. Participants were instructed to read the sentences naturally for comprehension, and to answer the end-of-sentence question as accurately as possible. Breaks were provided after 40 and 80 trials.

Working Memory Tasks

Participants completed a verbal measure of working memory, the reading span task (Daneman and Carpenter, 1980), and a non-verbal measure of working memory, the counting span task (Case et al., 1982). These tasks are argued to reflect working memory rather than short-term memory, as they involve both a memory component and a processing component, which interferes with rehearsal. Both tasks were presented to the native English speakers and the Korean learners of English in their native language, as it has been argued that measures of working memory capacity which are given in the second language are affected by the second language learners' English proficiency (e.g., Harrington and Sawyer, 1992; Juffs and Harrington, 2011).

In the reading span task, following the protocol in Conway et al. (2005), participants were asked to read sentences out loud and make sensicality judgments, while remembering random letters of the alphabet which followed each sentence (Kim, 2008). On each trial, the participant read the sentence out loud into a microphone, provided the sensicality judgment, and then said the letter that followed the sentence out loud, which triggered the next sentence in the series to immediately appear. After a series of 2–5 sentences, the participant was shown a screen prompting them to enter the letters that followed the previous set of sentences. Participants entered the recalled letters into boxes on this screen and were instructed to use a period (.) as a placeholder for letters that they could not recall.

The counting span task required participants to count target visual stimuli mixed in with distractor stimuli in a series of successive displays, while remembering the total number of target stimuli for each individual display (Conway et al., 2005). In each trial, the participant was presented with an array of target objects (dark blue circles) and distractor objects (light green circles); upon presentation of this array, the participant counted the number of target stimuli out loud, repeating the total, at which point the experimenter immediately entered the total using a computer keyboard, which triggered the next trial to begin. After a series of 2–6 trials, the participant was shown a screen prompting them to enter the total number of target objects from each of the previous arrays they had been presented. Participants entered the totals that they recalled into boxes on this screen, and

entered a period (.) as a placeholder for any totals that they could not recall.

For both the reading span task and the counting span task, participants were instructed to respond as quickly and accurately as possible. Stimuli within each task were presented in a randomized order. The entire testing session, including all of the above-mentioned tasks, took ∼60 min for native English speakers and 75 min for the Korean learners of English.

Data Analysis

As mentioned in the *Participants* section above, in addition to the 49 advanced Korean learners of English (mean age = 28.41; range 18–48 years old) and 54 native English speakers (mean age = 21.09; range 17–65 years old) reported in the current study, eight additional Korean learners of English and eight native English speakers were initially tested but identified as outliers and excluded from the final analysis, since their filled-gap effects were >3 standard deviations from the mean effect size of the dataset as a whole. Using filled-gap effect size as a value for identifying outliers is motivated by the fact that filled-gap effect size is a primary variable of interest in the regression analyses reported below. While these outliers are of most concern for the regression analyses, in order to keep the participant groups identical in the ANOVA analyses reported below (which probe for the presence of filled-gap effects in grammatically licit positions and for the avoidance of gap-filling inside islands) and in the regression analyses (which examine the relationships between individuals' filled-gap effect size and working memory), these participants were removed from both types of analysis. One additional native English speaker was also removed prior to analysis as this participant read at an extremely fast rate (faster than 250 ms) across regions and conditions.

For the dataset reported here, overall mean accuracy rate for the end-of-sentence question was 96.3% for native speakers and 93.4% for Korean learners of English; no participant in either group performed at <80% accuracy. Only those trials for which the end-of-sentence question was answered correctly were carried forward for statistical analysis. For *Non-Island* sentences, this resulted in exclusion of 3.43% of the data for native English speakers and 6.43% of the data for the Korean learners of English. For *Island* sentences, this resulted in exclusion of 3.8% of the data for the native English speakers and 6.73% of the data for the Korean learners of English.

Residual reading times were calculated by subtracting the raw reading time from the reading time predicted given a word's length by a regression equation that was constructed separately for each participant. Residual reading times beyond 2 standard deviations from the participant's mean for a given condition in a given region were excluded from the analysis (Ratcliff, 1993). For *Non-Island* sentences, this resulted in exclusion of 3.88% of the data for the native English speakers and 3.85% of the data for the Korean learners of English. For *Island* sentences, this resulted in exclusion of 4.1% of the data for the native English speakers and 4.03% of the data for the Korean learners of English.

2 × 2 mixed repeated-measures ANOVAs were performed on the remaining data, both by participants (F_1) and by items (F_2). For both the *Non-island* and the *Island* comparisons, the between-subjects factor was Group (native vs. learner) and the within-subjects factor was Condition (wh-extraction vs. no extraction). The critical regions for *Non-Island* sentences were region 8 (object filled-gap) and its spillover region (region 9), as well as region 5 (subject filled-gap) and its spillover region (region 6). The critical regions for *Island* sentences were region 9 (illicit object filled-gap within the relative clause island) and its spillover region (region 10), as well as region 5 (subject filled-gap) and its spillover region (region 6).

We also conducted a regression analysis to examine the relationship between filled-gap effect size and working memory both in grammatically licit positions and inside islands. For this analysis, we calculated for each individual the difference in mean reading times between the no extraction and wh-extraction conditions (subtracting the no extraction from the wh-extraction condition) in a given critical region; this measure, which we refer to throughout as Filled-gap Effect Size, serves as the dependent variable for the regression analyses. In order to obtain an independent variable reflecting working memory, we averaged for each individual their scores on the reading span task and the counting span task to create a Combined Working Memory Score. We use this score rather than the separate scores for each of the two working memory measures because the scores on these two measures are highly correlated ($r = 0.528$, $p < 0.001$). Because cognitive functioning, which includes working memory, declines with age (e.g., Hess, 2005; Oberauer, 2005; McArdle et al., 2007; Nettelbeck and Burns, 2010; Wass et al., 2012), we also control for age in our regression models. For both the ANOVA analyses and the regression analyses, we interpret $p < 0.05$ as significant and p-values between 0.05 and 0.10 as marginal.

Working memory score was calculated as a score from 1 to 100 based on percent of letters (for the reading span task) or numbers (for the counting span task) that were accurately recalled. Korean learners of English scored an average of 61.95% (range of 33.46–93.75%) on this composite measure of working memory, as compared to 59.73% (range of 29.25–88.93%) for native speakers of English. We used partial-unit scoring such that participants were given credit for each letter or number recalled in the correct position within a given trial. Performance on the processing tasks was not included in the working memory score, following the protocol outlined in Conway et al. (2005), who discuss the fact that accuracy on the processing tasks often correlates with the recall accuracy of the target items[3].

To address whether higher working memory capacity facilitates gap filling within or outside islands, we completed a sequential regression analysis for each critical and spillover region, while controlling for the effects of age. Filled-gap Effect

[3]Following a reviewer's suggestion, we examined the relationship between the recall and processing components of the working memory tasks. Recall and accuracy scores for the counting span task were significantly correlated ($r = 0.300$, $p = 0.002$) and those for the reading span task were moderately correlated ($r = 0.181$, $p = 0.067$). While there was a small but positive correlation between these components, we ran new regression models using a new composite working memory score which incorporated participants' performance on both the recall and processing components for each task (using an average of the processing and recall scores), in line with Waters and Caplan (1996). The pattern of results remains unchanged in these new analyses.

Size at each region was regressed on age, centered scores of Combined Working Memory, and Group (native = 0, L2 learner = 1) in the first block of a sequential regression. The cross-product of the centered Combined Working Memory scores and Group was then added in the second block and ΔR^2 was examined to determine if an interaction between groups was present. Follow-up analyses were conducted for those regions showing an interaction.

RESULTS

Filled-Gap Effects

Object Filled-Gap Effects

In the *Non-Island* comparison, the results of the mixed repeated measures ANOVA for region 8, the critical post-verbal object position from which extraction is grammatically licit, did not reveal main effects of Group [$F_{1(1,\ 101)} = 0.18$, $p = 0.67$; $F_{2(1,\ 38)} = 0.004$, $p = 0.95$] or Condition [$F_{1(1,\ 101)} = 0.802$, $p = 0.372$; $F_{2(1,\ 38)} = 2.160$, $p = 0.15$]. Furthermore, there was no interaction between these factors [$F_{1(1,\ 101)} = 0.308$, $p = 0.58$; $F_{2(1,\ 38)} = 2.750$, $p = 0.11$]. However, a main effect of Condition emerged at region 9 [$F_{1(1,\ 101)} = 6.032$, $p < 0.05$; $F_{2(1,\ 38)} = 13.967$, $p < 0.01$], reflecting a reading time slowdown in the wh-extraction condition as compared to the no extraction condition. There was no main effect of Group [$F_{1(1,\ 101)} = 0.657$, $p = 0.419$; $F_{2(1,\ 38)} = 1.909$, $p = 0.18$] nor was there an interaction at region 9 between Group and Condition [$F_{1(1,\ 101)} = 1.609$, $p = 0.207$; $F_{2(1,\ 38)} = 0.160$, $p = 0.69$]. Mean reading times for native English speakers in the *Non-Island* sentences are shown in **Figure 1**, and those for Korean learners of English are shown in **Figure 2**.

For the *Island* comparison, no main effects of Group [$F_{1(1,\ 101)} = 0.208$, $p = 0.65$; $F_{2(1,\ 38)} = 1.479$, $p = 0.23$] or Condition [$F_{1(1,\ 101)} = 0.779$, $p = 0.38$; $F_{2(1,\ 38)} = 1.269$, $p = 0.27$] emerged at the critical region 9, the post-verbal object position within the relative clause island. There was a marginal Group by Condition interaction at region 9 in the by-participants analysis [$F_{1(1,\ 101)} = 3.008$, $p = 0.09$; $F_{2(1,\ 38)} = 2.177$, $p = 0.148$]. However, *post-hoc* *t*-tests revealed that the reading time difference between the wh-extraction condition and the no extraction condition was not significant for either native English speakers [$t_{(53)} = -0.729$, $p = 0.47$, two-tailed paired *t*-test] or Korean learners of English [$t_{(48)} = 1.578$, $p = 0.12$, two-tailed paired *t*-test]. At the spillover region, region 10, there was a main effect of Group in the by-items analysis only [$F_{1(1,\ 101)} = 0.365$, $p = 0.55$; $F_{2(1,\ 38)} = 4.173$, $p < 0.05$]. This effect reflected the fact that residual reading times were slower overall for Korean learners of English than for native English speakers. Additionally, there was an effect of Condition in region 10 which reached significance only in the by-items analysis [$F_{1(1,\ 101)} = 2.605$, $p = 0.11$; $F_{2(1,\ 38)} = 5.240$, $p < 0.05$]. However, this effect was in the opposite direction of what would be expected if a filled-gap effect were to emerge; participants read faster in the wh-extraction condition as compared to the no extraction condition. There was also no interaction between Group and Condition at region 10 [$F_{1(1,\ 101)} = 0.096$, $p = 0.76$; $F_{2(1,\ 38)} = 0.484$, $p = 0.49$]. Overall, the results from the *Non-Island* comparison indicate that, although numerically small, a significant filled-gap effect emerged for both groups at the spillover region of the filled direct object position. In contrast, as evidenced by the results from the *Island* comparison, neither native English speakers, nor Korean learners of English show a filled-gap effect within the relative clause island. Mean reading times for the *Island* sentences for native English speakers are illustrated in **Figure 3**, and those for Korean learners of English are shown in **Figure 4**.

Subject Filled-Gap Effects

In addition to examining whether native English speakers and Korean learners of English showed evidence of object filled-gap effects, we also examined the critical region 5 and spillover region 6 in both the *Non-Island* and the *Island* sentences for possible subject filled-gap effects. Recall that for both sentence types, the subject gap positions are licit positions for wh-extraction.

In the *Non-Island* comparison, there was no effect of Condition in either region 5 [$F_{1(1,\ 101)} = 1.612$, $p = 0.21$; $F_{2(1,\ 38)} = 2.523$, $p = 0.12$] or region 6 [$F_{1(1,\ 101)} = 0.926$, $p = 0.34$; $F_{2(1,\ 38)} = 2.082$, $p = 0.16$]. There was an effect of Group in region 5 [$F_{1(1,\ 101)} = 17.157$, $p < 0.001$; $F_{2(1,\ 38)} = 4.990$, $p < 0.05$], reflecting the fact that Korean learners of English yielded slower residual reading times overall compared to native English speakers. There was no effect of Group in region 6 [$F_{1(1,\ 101)} = 0.05$, $p = 0.82$; $F_{2(1,\ 38)} = 0.010$, $p = 0.92$]. There was no interaction between Group and Condition in either the critical region 5 [$F_{1(1,\ 101)} = 0.309$, $p = 0.579$; $F_{2(1,\ 38)} = 0.069$, $p = 0.79$] or the spillover region 6 [$F_{1(1,\ 101)} = 0.257$, $p = 0.61$; $F_{2(1,\ 38)} = 1.465$, $p = 0.23$].

In the *Island* comparison, there was a main effect of Condition at region 5 [$F_{1(1,\ 101)} = 7.308$, $p < 0.01$; $F_{2(1,\ 38)} = 6.769$, $p < 0.05$] reflecting that participants showed a reading time slowdown in the wh-extraction condition as compared to the no extraction condition. There was no effect of Group at region 5 [$F_{1(1,\ 101)} = 0.029$, $p = 0.87$; $F_{2(1,\ 38)} = 0.173$, $p = 0.68$]. There was a marginal interaction in the by-participants analysis, and a significant interaction in the by-items analysis between Group and Condition at region 5 [$F_{1(1,\ 101)} = 3.759$, $p = 0.055$; $F_{2(1,\ 38)} = 9.826$, $p < 0.01$]. *Post-hoc* *t*-tests revealed that native English speakers showed a significant slowdown in the wh-extraction condition as compared to the no extraction condition [$t_{1(53)} = -3.507$, $p < 0.01$, two-tailed paired *t*-test; $t_{2(19)} = -5.284$, $p < 0.01$, two-tailed paired *t*-test]. However, the effect of Condition for Korean learners of English at region 5 was not significant [$t_{1(48)} = -0.506$, $p = 0.62$, two-tailed paired *t*-test; $t_{2(19)} = 0.317$, $p = 0.75$, two-tailed paired *t*-test]. At the spillover region 6, there was no main effect of Condition [$F_{1(1,\ 101)} = 0.492$, $p = 0.49$; $F_{2(1,\ 38)} = 0.067$, $p = 0.80$]. There was an effect of Group [$F_{1(1,\ 101)} = 13.117$, $p < 0.001$; $F_{2(1,\ 38)} = 2.696$, $p = 0.11$] reflecting slower residual reading times overall for Korean learners of English than for native English speakers. There was no interaction between Group and Condition at region 6 [$F_{1(1,\ 101)} = 0.016$, $p = 0.90$; $F_{2(1,\ 38)} = 0.745$, $p = 0.39$]. Thus, subject filled-gap effects emerged only for native English speakers, and only at the critical region in the *Island* comparison.

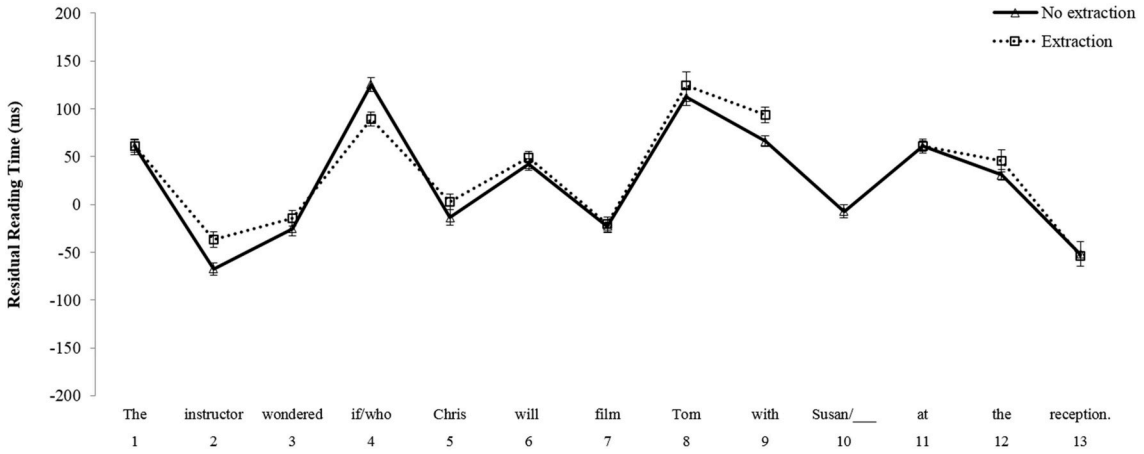


FIGURE 1 | Mean reading times by participants for Non-Island sentences, Native English speakers.

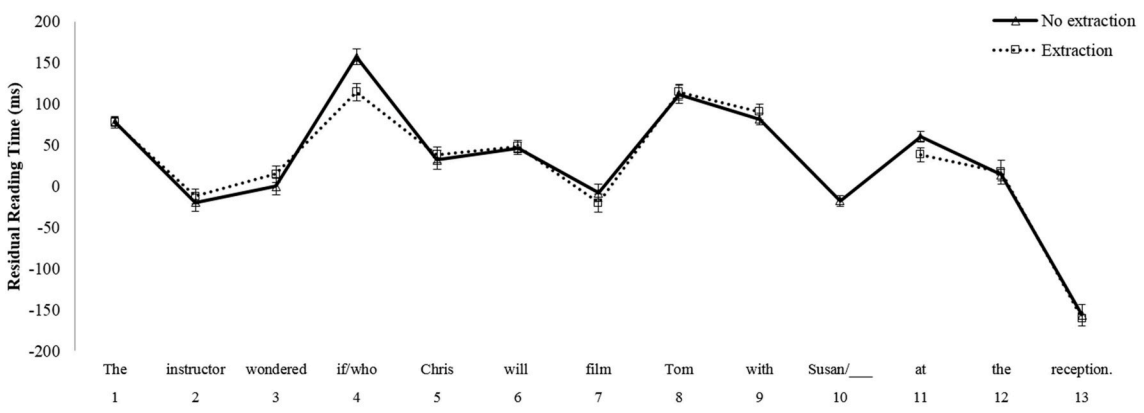


FIGURE 2 | Mean reading times by participants for Non-Island sentences, Korean learners of English.

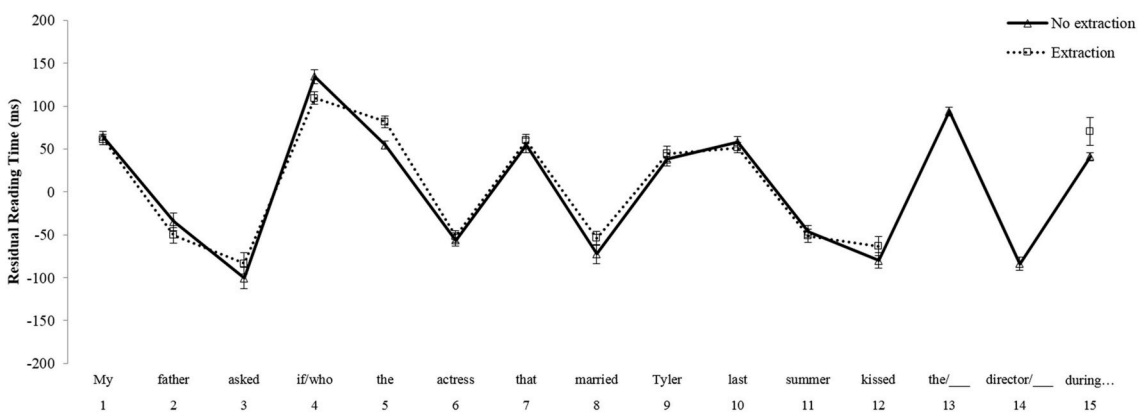


FIGURE 3 | Mean reading times by participants for Island sentences, Native English speakers.

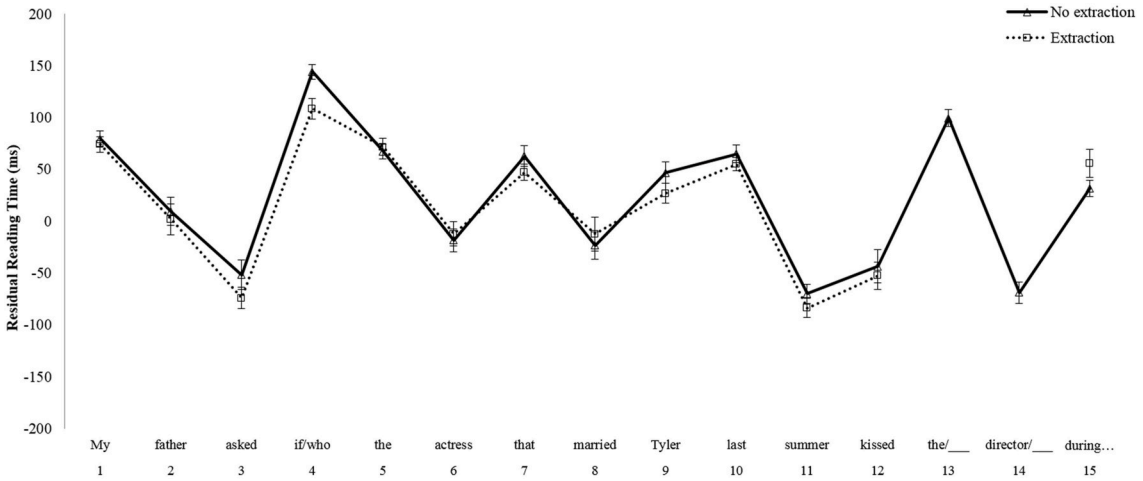


FIGURE 4 | Mean reading times by participants for Island sentences, Korean learners of English.

Results: Effects of Working Memory on Filled-Gap Effect Size

Gap-Filling within Islands

Regression models for the critical and spillover filled-gap regions (regions 9 and 10) within the relative clause island in *Island* sentences were not significant. For region 9, the first block of the sequential regression was not significant [adjusted $R^2 = 0.002$, $F_{(3, 99)} = 1.055$, $p = 0.372$]. The addition of the cross-product of Combined Working Memory and Group in the second block did not significantly increase the variance explained by the model [$\Delta R^2 = 0.000$, adjusted $R^2 = -0.009$, $F_{(1, 98)} = 0.001$, $p = 0.982$]. Similarly, for region 10 the first block of the sequential regression was not significant [adjusted $R^2 = 0.008$, $F_{(3, 99)} = 1.267$, $p = 0.29$]. The addition of the cross-product of Combined Working Memory and Group in the second block did not significantly increase the variance explained by the model [$\Delta R^2 = 0.000$, adjusted $R^2 = -0.002$, $F_{(1, 98)} = 0.000$, $p = 1.00$]. Thus, working memory does not predict gap-filling in positions which are subject to island constraints.

Gap-Filling in Grammatically Licit Positions

As individual differences in working memory may affect the resolution of wh-dependencies in grammatically licensed positions, we also examined whether working memory modulated the magnitude of filled-gap effects in the following positions: the filled object position in *Non-Island* sentences, and the filled subject position in both *Non-Island* and *Island* sentences.

Object Filled-Gap: Non-Island Sentences

No significant effect of Working Memory on Filled-gap Effect Size was found at the critical region 8 for the object filled-gap. In region 8, the first block of the sequential regression was not significant [adjusted $R^2 = 0.016$, $F_{(3, 99)} = 1.539$, $p = 0.209$]. The addition of the cross-product of Combined Working Memory and Group in the second block did not significantly increase the variance explained by the model [$\Delta R^2 = 0.003$, adjusted $R^2 = 0.008$, $F_{(1, 98)} = 0.285$, $p = 0.595$]. A significant effect of Working Memory on Filled-gap Effect Size was found at the spillover region for the object filled-gap. For region 9 the first block of the sequential regression was not significant [adjusted $R^2 = -0.001$, $F_{(3, 99)} = 0.949$, $p = 0.42$]. However, the addition of the cross-product of Combined Working Memory and Group in the second block significantly increased the variance explained by the model [$\Delta R^2 = 0.043$, adjusted $R^2 = 0.034$, $F_{(1, 98)} = 4.584$, $p < 0.05$]. Thus, the effect of working memory on object filled-gap effects in the spillover region depends on group membership. In follow-up analyses, the regression slopes were plotted separately by Group (**Figure 5**). To examine the differences in slope for the two groups, follow-up regression analyses were performed separately for native speakers and learners. The results show that the regression of Working Memory on Filled-gap Effect Size for native speakers, when controlling for age, was not significant [adjusted $R^2 = 0.002$, $F_{(2, 51)} = 1.049$, $p = 0.358$]. The regression of Working Memory and Age on Filled-gap Effect Size for Koreans was significant [adjusted $R^2 = 0.131$, $F_{(2, 46)} = 4.626$, $p < 0.02$]. When controlling for age, working memory had a moderately significant effect on Filled-gap Effect Size. For every one standard deviation increase in working memory score, Filled-gap Effect Size decreased by 0.268 standard deviations [$b = -1.40$, $t_{(46)} = -1.96$, $p = 0.056$, $\beta = -0.268$, 95% CI (-2.84 -0.036)]. Thus, the data shows a trend suggesting that working memory predicts the degree of Filled-gap Effect Size at the spillover object filled-gap region for Korean learners of English, but not native English speakers. Specifically, an increase in working memory predicts a reduced Filled-gap Effect Size, and thus decreased filled-gap effects, at the spillover region 9 in Korean learners of English.

Subject Filled-Gap: Non-Island Sentences

For the critical subject filled-gap region in the *Non-Island* sentences (region 5), the first block of the sequential regression

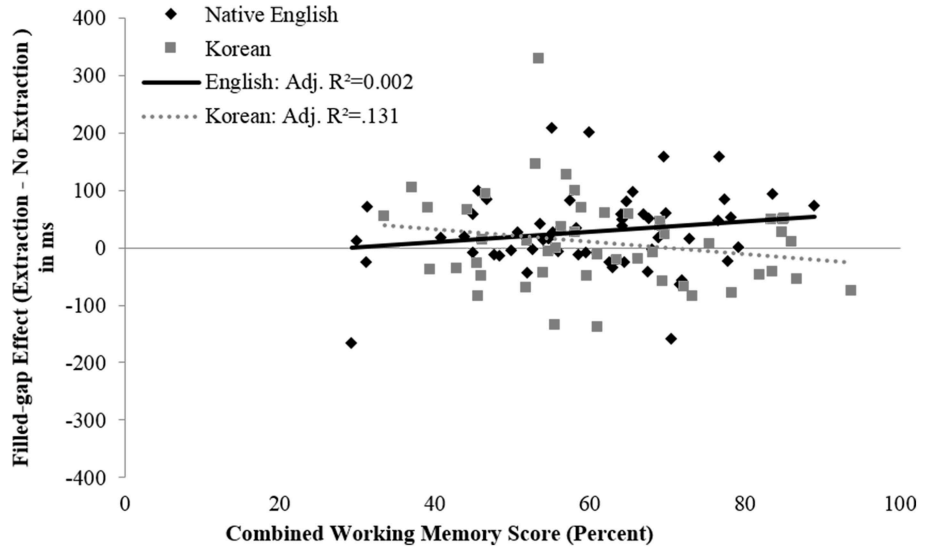


FIGURE 5 | Filled-gap effect size as a function of combined working memory score for the spillover region (region 9) of the licit object filled-gap position in Experiment 1. Linear trend lines (least-squares method) are illustrated in the figure; adjusted R^2-values are provided in the legend.

was not significant [adjusted $R^2 = 0.016$, $F_{(3, 99)} = 1.55$, $p = 0.206$]. However, the addition of the cross-product of Combined Working Memory and Group in the second block significantly increased the variance explained by the model [$\Delta R^2 = 0.084$, adjusted $R^2 = 0.093$, $F_{(1, 98)} = 9.392$, $p < 0.01$]. Thus, the effect of working memory on gap-filling, when controlling for age, depends on Group.

To better understand the nature of the moderation, the regression slopes were plotted separately by Group (**Figure 6**). To examine the differences in slope for the two groups at region 5, follow-up regression analyses were performed separately for the two groups. The results show that the regression of Working Memory on Filled-gap Effect Size for native speakers was significant [adjusted $R^2 = 0.2$, $F_{(2, 51)} = 7.621$, $p < 0.01$]. For every one standard deviation increase in working memory score, Filled-gap Effect Size increased by 0.433 standard deviations, when controlling for age [$b = 3.029$, $t_{(51)} = 3.506$, $p < 0.01$, $\beta = 0.433$, 95% CI (1.30–4.76)]. However, the effect of Working Memory on Filled-gap Effect Size for Korean learners of English was not significant [adjusted $R^2 = -0.017$, $F_{(2, 46)} = 0.602$, $p = 0.552$]. Thus, the data suggests that working memory does predict reading times at the subject filled-gap region for native English speakers, but not for Korean learners of English. Specifically, an increase in working memory predicts an increased slowdown, or filled-gap effect, at the filled subject gap region 5 in native speakers of English.

In the spillover region 6 for the subject filled-gap in the Non-Island sentences, the first block of the sequential regression was not significant [adjusted $R^2 = -0.022$, $F_{(3, 99)} = 0.253$, $p = 0.859$]. The addition of the cross-product of Combined Working Memory and Group in the second block did not significantly increase the variance explained by the model [$\Delta R^2 = 0.015$, adjusted $R^2 = -0.017$, $F_{(1, 98)} = 1.493$, $p = 0.225$].

Subject Filled-Gap: Island Sentences

For region 5 in the *Island* sentences, the first block of the sequential regression was not significant [adjusted $R^2 = 0.012$, $F_{(3, 99)} = 1.425$, $p = 0.24$]. The addition of the cross-product of Combined Working Memory and Group in the second block increased the variance explained by the model by a significant amount [$\Delta R^2 = 0.048$, adjusted $R^2 = 0.052$, $F_{(1, 98)} = 5.186$, $p < 0.05$]. However, follow-up regression analyses performed separately for the two groups found that the regression of Working Memory on Filled-gap Effect Size for Native Speakers, when controlling for age, was not significant [adjusted $R^2 = 0.016$, $F_{(2, 51)} = 1.418$, $p = 0.252$]. The regression of Working Memory on Filled-gap Effect Size for Korean learners of English, when controlling for age, was also not significant [adjusted $R^2 = 0.018$, $F_{(2, 46)} = 1.434$, $p = 0.249$]. There were no significant effects in the spillover region 6 in the *Island* sentences. The first block of the sequential regression was not significant [adjusted $R^2 = -0.021$, $F_{(3, 99)} = 0.302$, $p = 0.824$]. The addition of the cross-product of Combined Working Memory and Group in the second block did not significantly increase the variance explained by the model [$\Delta R^2 = 0.002$, adjusted $R^2 = -0.030$, $F_{(1, 98)} = 0.160$, $p = 0.690$].

DISCUSSION

The present study examined whether native speakers and L2 learners show qualitatively similar patterns in the processing of wh-dependencies in both licit and illicit contexts. Previous studies have shown that native speakers attempt to resolve wh-dependencies in grammatically licensed positions but avoid positing gaps in islands (Stowe, 1986; Traxler and Pickering, 1996). In the present study, we replicated this pattern for native English speakers and showed the same pattern of results for advanced Korean learners of English as well. In the non-island

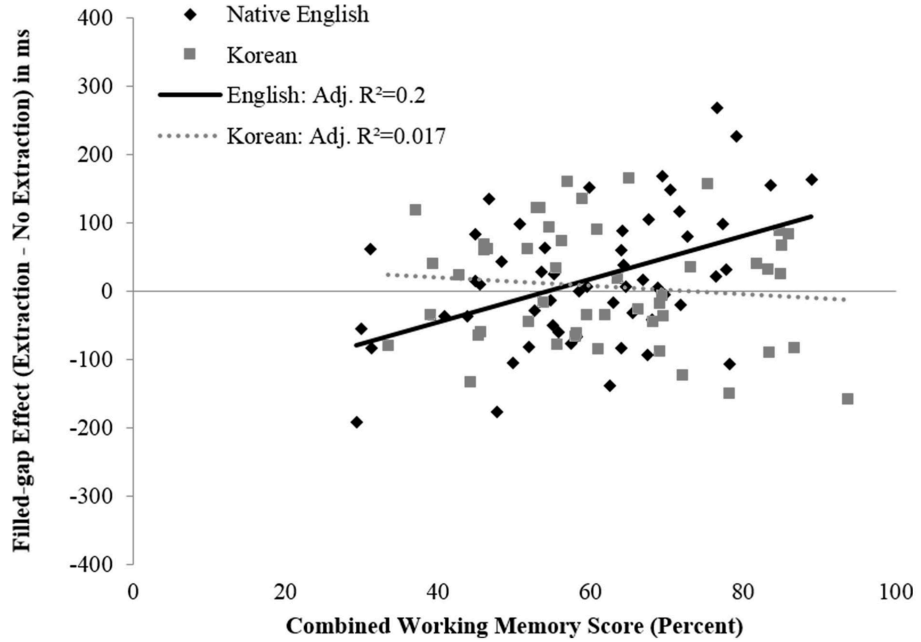


FIGURE 6 | Filled-gap effect size as a function of combined working memory score for the licit subject filled-gap position (region 5) in Experiment 1. Linear trend lines (least-squares method) are illustrated in the figure; adjusted R^2-values are provided in the legend.

sentences, a significant filled-gap effect emerged in the spillover region following the direct object of the verb. A significant interaction with group did not emerge, demonstrating qualitative similarity between the two groups. In Felser et al. (2012), evidence of a filled-gap effect emerged for L2 learners in a later region than the region in which the effect emerged for native speakers, a result which supported their proposal that learners cannot use syntactic information on the same timecourse as native speakers[4]. In contrast, the results of the present study are in line with our previous work, which also showed the same pattern for L2 learners and natives (Aldwayan et al., 2010). While it is true that self-paced reading does not allow the same range of dependent measures as eye-tracking in terms of characterizing the timecourse of processing, it is important to point out that in the Felser et al. (2012) study, the filled-gap effects for natives and learners emerged in distinct regions, not in different dependent measures within the same region.

In contrast to the non-island sentences, where significant object filled-gap effects emerged for both groups, there were no object filled-gap effects in island sentences, in which the critical region was embedded within a relative clause island. Our results are in line with several previous studies which have examined relative clause islands and have shown that learners avoid attempting to resolve wh-dependencies in grammatically unlicensed contexts (Aldwayan et al., 2010; Omaki and Schulz, 2011; Felser et al., 2012; Kim et al., 2015 for Spanish natives)[5]. In the current literature, both studies which showed evidence of gap-filling in islands by L2 learners used a plausibility mismatch paradigm (Boxell and Felser, 2013 in first pass reading measures; Kim et al., 2015 for Korean natives). However, it is important to point out that in the Felser et al. (2012) study, which also used a plausibility mismatch paradigm, learners showed effects of plausibility even earlier than natives but at no point did they show evidence of attempting to resolve wh-dependencies in islands.

Our examination of the subject position yielded a significant subject filled-gap effect for native speakers, but only in island sentences[6]. As we discussed above, this inconsistency across experiments and groups is in line with previous studies. There

[4]A reviewer suggests that the results of the present study may differ from the results of previous studies which showed differences between learners and native speakers because of differences in proficiency levels. The learners in the present study scored between intermediate and advanced levels on the proficiency test and were immersed in the L2 environment. However, the Felser et al. (2012) study also included intermediate-advanced learners who were immersed in an English-speaking environment. As different proficiency measures were used across studies, it is hard to directly compare proficiency levels.

[5]As a reviewer pointed out, the lexical items in the *Non-Island* and the *Island* sentences are not the same, questioning whether this may complicate interpretation of the findings. Our overall finding of a filled-gap effect in the licit object position, and avoidance of a filled-gap effect in object position within an island converges with a range of previous studies using either filled-gap or plausibility manipulations and more closely matched lexical material (e.g., Stowe, 1986; Traxler and Pickering, 1996; Omaki and Schulz, 2011; Felser et al., 2012). However, we agree that a future extension of the current study, with closely matched lexical items across the *Non-Island* and *Island* conditions would be ideal for addressing this open question.

[6]A reviewer pointed out that lexical differences may have played in a role in the distribution of the subject filled-gap effect, which emerged only in the Non-Island comparison. Indeed, as the lexical subject was a proper name in the first comparison but a determiner-noun sequence in the second comparison, this is a possibility. Although very few studies report significant subject filled-gap effects, across studies, significant effects have been reported for both determiner-noun sequences (e.g., Aldwayan et al., 2010) and proper names (e.g., Lee, 2004; Johnson, 2015). Moreover, not all studies which include a proper name in subject position

is an extensive literature discussing why evidence for filled-gap effects in subject position is mixed (Stowe, 1986; Clifton and Frazier, 1989; Clifton and De Vincenzi, 1990; De Vincenzi, 1991; Gibson et al., 1994; Lee, 2004; Johnson, 2015): several researchers have proposed that the adjacency of the wh-filler and the subject position may not provide sufficient time to either generate or commit to a prediction for a subject gap. This proposal would suggest that allowing more time, in terms of the distance between the wh-filler and subject position, may yield different results (see Lee, 2004). Also related to this proposal, one might also expect that individuals with greater processing resources would be more likely to be able to immediately generate a prediction for a subject gap; we will return to this point below in our discussion of individual differences.

The present study also examined the nature of islands by investigating the relationship between working memory and filled-gap effects in both native speakers and L2 learners. A pattern of results showing that individuals with more processing resources are better able to establish wh-dependencies in islands would be compatible with the processing account proposed by Hofmeister and colleagues (e.g., Hofmeister and Sag, 2010). In contrast, grammatical accounts do not predict such a relationship within islands as the parser should simply not predict a gap within island contexts. Note however that a pattern of results that shows no relationship between working memory and filled-gap effects within islands is also potentially compatible with the processing accounts as null results may be explained by a range of factors including, as discussed by Hofmeister et al. (2012a,b, 2013, 2014) inappropriate choice of working memory measures and selection of stimuli that are simply too complex for individual differences to emerge. As our results showed that there was indeed no significant relationship between working memory and filled-gap effects in island contexts for either native speakers or learners, we will consider this range of possibilities as related to our study. In the present study, the lack of a relationship between working memory and filled-gap effects in islands is unlikely to be due to the selection of an inappropriate measure of working memory or lack of statistical power as significant relationships between working memory and filled-gap effects emerged within licit contexts for both learners and natives (although the patterns for the two groups differed). Although the interpretation of these findings is complex, they do suggest, in line with previous studies, that our working memory measure is one that can indeed capture variability in linguistic processing (Daneman and Carpenter, 1980; King and Just, 1991; Just and Carpenter, 1992; Hofmeister et al., 2014; see Hofmeister et al., 2012a for discussion).

Next, we consider whether the difference between the licit and illicit island contexts is simply the result of differences in processing load: if the island sentences simply overwhelmed the parser, perhaps a significant relationship with working memory did not emerge because of a lack of variability. For example, Hofmeister et al. (2014) fail to show a relationship between reading span scores and acceptability judgments for sentences of extreme processing difficulty although significant relationships did emerge for less complex structures. We think that this explanation is unlikely due to the comparability of the stimuli which targeted licit and illicit gap sites (see 9, 10). In both the non-island and island conditions, the target sentences were all grammatical, indirect questions which allowed us to avoid presenting direct wh-questions in isolation, which Hofmeister et al. (2012a) have argued is unnatural. In terms of the comparison between the licit and illicit object positions, it is important to note that these potential gap sites occur at similar points in the sentence (region 8, region 9) and at similar distances from the wh-filler (three and four words after the filler). In addition, in both sentence types, the wh-filler is followed by a single animate noun phrase and a tensed verb. These similarities serve to minimize the differences in processing difficulty of the licit and illicit object gap sites. Thus, while it is difficult to argue categorically in support of or against either account on the basis of a lack of a relationship between working memory and filled-gap effects in islands, we believe the design of the present study can potentially be defended against some of the criticisms raised in the literature by Hofmeister et al. (2012a,b, 2014). In addition, we believe there is merit to the approach we have taken in examining the relationship between individual differences and processing-based dependent measures across both island and non-island contexts. Indeed, it would be interesting to examine whether the results of the current study would be replicated in an experiment testing sentences that include linguistic properties that have been shown to ease the processing of wh-dependencies, such as complex wh-fillers (e.g., Hofmeister and Sag, 2010; Goodall, 2015). Such an experiment would provide an ideal way to address the potential concern that the lack of variability in gap-filling inside islands in the current study could be because the processing of those island structures is simply beyond the reach of all participants, even those with high working memory.

As we discussed earlier, any significant relationships that emerge at the licit gap sites are consistent with both the processing and grammatical accounts of islands but we believe that our findings raise very interesting questions as to the nature of the relationship between working memory and the processing of wh-dependencies in both learners and native speakers. In the non-island sentences, a positive correlation emerged between working memory and the filled-gap effect size at the subject position; this effect was significant only for native speakers. As we discussed above, one possible explanation is that participants with greater processing resources are better able to immediately generate a prediction for a potential gap (e.g., Hutchison, 2007; Slevc and Novick, 2013; Johnson, 2015) and thus show a greater filled-gap effect. The question remains why this relationship did not emerge at the licit subject position in both non-island and island contexts or in the L2 learner group. In a recent study in our lab, Johnson (2015) conducted a large scale study of native speakers ($n = 110$) and intermediate and advanced Korean learners of English ($n = 100$). The self-paced reading experiment included sentences similar to the ones tested in the present study. The results showed that significant subject filled-gap effects emerged for both groups. All participants also completed

show significant subject filled-gap effects (e.g., Stowe, 1986). Thus, the effect of the structure of the noun phrase remains an interesting open question.

measures of cognitive abilities including working memory (counting span) and attentional control (number Stroop). The size of the subject filled-gap effect in both natives and L2 learners was significantly related to attentional control, which Hutchison (2007) has argued to be a key component in the ability to generate and maintain predictions. Taken together, these results show that there is even variability in the processing of wh-dependencies that are relatively simple in terms of structure but demanding in terms of the need to automatically generate a prediction for an upcoming gap. This variability may lead to a need for large sample sizes, such as those in Johnson (2015), in order for robust effects to emerge. In addition, in an effort to better understand the cognitive abilities that underlie this variability in both natives and learners, future studies should include a wider range of measures, allowing for a more precise examination of whether the cognitive abilities that underlie variability in native speakers are similar or different to the abilities that underlie the variability in learners.

Our results for the non-island sentences also showed a relationship between working memory and the size of the licit object filled-gap effect in the spillover region but this effect emerged only for the L2 learners. Unexpectedly, the results showed that an increase in working memory predicted a reduced reading time slowdown or a smaller object filled-gap effect. One possible explanation is that the learners with greater processing resources may have recovered more easily from encountering the filled-gap, resulting in a reduced filled-gap effect at the spillover region. To explore this possibility, we separated the Korean learners of English into high ($n = 22$) and low ($n = 27$) working memory groups, based on whether they scored above or below the mean for the group (62) and then compared the size of the filled-gap effects at both the critical region (region 8) and the spillover region (region 9), where the relationship with working memory emerged (see **Figure 7**).

This comparison demonstrates that the high working memory group showed a numerical slowdown in the predicted direction only at the critical region. Thus, it is at least possible that learners with higher working memory showed a reduced filled-gap effect at the spillover region because they had already recovered from encountering the lexical material in the preceding region. As this comparison is exploratory, we present this numerical pattern in the learner data in order to suggest a direction for future research, one that may also benefit from an increased sample size, as in Johnson (2015), which may allow a wider range of variability to emerge in both learners and native speakers. An alternative method such as eye-tracking may also allow a more precise characterization of the dynamics of attempting to resolve wh-dependencies, including the initial detection of a filled potential gap site and recovery from this mis-analysis.

Although the results of our individual differences analyses raise many open questions, they suggest that processing resources do modulate the processing of wh-dependencies in certain grammatically licensed contexts. Why different relationships with working memory arise for the learners and native speakers is a very interesting question for future research. Further study is needed to examine whether similar or different cognitive abilities facilitate processing at different points for the two populations.

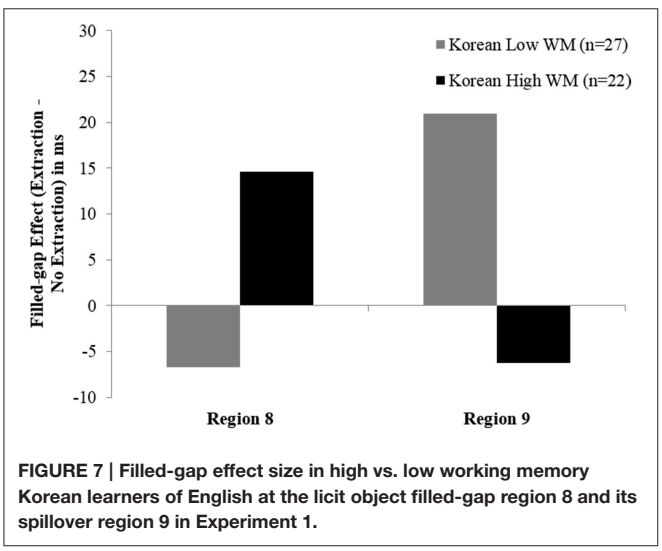


FIGURE 7 | Filled-gap effect size in high vs. low working memory Korean learners of English at the licit object filled-gap region 8 and its spillover region 9 in Experiment 1.

CONCLUSION

In the current study, we investigated the processing of wh-dependencies in both native speakers and L2 learners, examining whether the two groups show qualitatively similar patterns in processing and whether there is a relationship between working memory and filled-gap effects in both island and non-island contexts. The results showed that both native and non-native speakers posit gaps in grammatically licensed contexts but avoid positing gaps in islands. The processing profile of natives and L2 learners was qualitatively similar, showing no evidence of a delay in the use of syntactic knowledge as has been argued in recent proposals (Felser et al., 2012; Boxell and Felser, 2013). Our individual differences analyses showed no relationship between working memory and filled-gap effects within islands but we did observe significant relationships between working memory and the processing of licit wh-dependencies. As the contexts in which these relationships emerged differed for learners and native speakers, our results call for further research examining individual differences in dependency resolution in the two populations.

AUTHOR CONTRIBUTIONS

All authors listed, have made substantial, direct and intellectual contribution to the work, and approved it for publication.

ACKNOWLEDGMENTS

We are thankful to Goun Lee and Ji Yeon Lee for their help in preparing our experimental materials in Korean and in recruiting. We would also like to thank JoAnn Doll for assistance with testing participants, Alonso Canales for assistance with the stimuli construction, and Bruno Tagliaferri for his help with Python scripting. We thank the members of the Research in

Acquisition and Processing Seminar at the University of Kansas and the audience at the 38th *Boston University Conference on Language Development* for their feedback. We would also like to thank the reviewers for their feedback which helped to improve the paper.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpsyg.2016.00549

REFERENCES

Aldosari, S. (2015). *The Role of Individual Differences in the Acceptability of Island Violations in Native and Non-native Speakers*. Doctoral Dissertation, University of Kansas.

Aldwayan, S., Fiorentino, R., and Gabriele, A. (2010). "Evidence of syntactic constraints in the processing of wh-movement: a study of Najdi Arabic learners of English," in *Research in Second Language Processing and Parsing*, eds B. Van Patten and J. Jegerski (Amsterdam: John Benjamins Publishing Company), 65–86.

Belikova, A., and White, L. (2009). Evidence for the fundamental difference hypothesis or not? Island constraints revisited. *Stud. Second Lang. Acquis.* 31, 199–223. doi: 10.1017/S0272263109090287

Boxell, O., and Felser, C. (2013). "*The processing of subject islands containing parasitic gaps: evidence from eye-tracking during reading in native and non-native speakers,*" in *Slide Presentation at Linguistic Evidence* (Berlin).

Canales, A. (2012). *Online Processing of wh-Dependencies in English by Native Speakers of Spanish*. Doctoral dissertation, University of Kansas. ProQuest Dissertations and Theses.

Case, R., Kurland, M. D., and Goldberg, J. (1982). Operational efficiency and the growth of short-term memory span. *J. Exp. Child Psychol.* 33, 386–404. doi: 10.1016/0022-0965(82)90054-6

Chomsky, N. (1973). "Conditions on transformations," in *A festschrift for Morris Halle,* eds S. R. Anderson and P. Kiparsky (New York, NY: Holt, Rinehart and Winston), 232–286.

Chomsky, N. (1986). *Barriers*. Cambridge, MA: MIT Press.

Cinque, G. (1990). *Types of ÄĀ-Dependencies*. Cambridge, MA: MIT press.

Clifton, C. Jr., and De Vincenzi, M. (1990). "Comprehending sentences with empty elements," in *Comprehension Processes in Reading*, eds D. A. Balota, G. B. Flores d'Arcais, and K. Rayner (Hillsdale, NJ: Erlbaum), 265–283.

Clifton, C. Jr., and Frazier, L. (1989). "Comprehending sentences with long-distance dependencies," in *Linguistic Structure in Language Processing,* eds G. Carlson, and M. Tanenhaus (Dordrecht: Kluwer), 94–128.

Conway, A., Kane, M., Bunting, M., Hambrick, D., Wilhelm, O., and Engle, R. (2005). Working memory span tasks: a methodological review and user's guide. *Psychon. Bull. Rev.* 12, 769–786. doi: 10.3758/BF03196772

Daneman, M., and Carpenter, P. A. (1980). Individual differences in working memory and reading. *J. Verbal Learning Verbal Behav.* 19, 450–466. doi: 10.1016/S0022-5371(80)90312-6

Daneman, M., and Carpenter, P. A. (1983). Individual differences in working memory and reading. *J. Exp. Psychol. Learn. Mem. Cogn.* 9, 561–584. doi: 10.1037/0278-7393.9.4.561

De Vincenzi, M. (1991). *Syntactic Parsing Strategies in Italian*. Dordrecht: Kluwer. doi: 10.1007/978-94-011-3184-1

Engle, R. W. (2002). Working memory capacity as executive attention. *Curr. Dir. Psychol. Sci.* 11, 19–23. doi: 10.1111/1467-8721.00160

Felser, C., Cunnings, I., Batterham, C., and Clahsen, H. (2012). The timing of island effects in nonnative sentence processing. *Stud. Second Lang. Acquis.* 34, 67–98. doi: 10.1017/S0272263111000507

Frazier, L. (1987). Syntactic processing: evidence from Dutch. *Nat. Lang. Linguist. Theory* 5, 519–559. doi: 10.1007/BF00138988

Gibson, E. (1998). Linguistic complexity: locality of syntactic dependencies. *Cognition* 68, 1–76. doi: 10.1016/S0010-0277(98)00034-1

Gibson, E., Hickok, G., and Schütze, C. T. (1994). Processing empty categories: a parallel approach. *J. Psycholinguist. Res.* 23, 381–405. doi: 10.1007/BF02143946

Goodall, G. (2015). The D-linking effect on extraction from islands and non-islands. *Front. Psychol.* 5:1493. doi: 10.3389/fpsyg.2014.01493

Han, C. H., and Kim, J. B. (2004). Are there "double relative clauses" in Korean? *Linguis. Inq.* 35, 315–337. doi: 10.1162/002438904323019101

Harrington, M., and Sawyer, M. (1992). L2 working memory capacity and L2 reading skills. *Stud. Second Lang. Acquis.* 14, 25–38. doi: 10.1017/S0272263100010457

Hess, T. M. (2005). Memory and gaining in context. *Psychol. Bull.* 131, 383–406. doi: 10.1037/0033-2909.131.3.383

Hofmeister, P., Casasanto, L. S., and Sag, I. A. (2012a). Misapplying working memory tests: a reductio *ad absurdum*. *Language (Baltim.)* 88, 408–409. doi: 10.1353/lan.2012.0033

Hofmeister, P., Casasanto, L. S., and Sag, I. A. (2012b). How do individual cognitive differences relate to acceptability judgments?: a reply to Sprouse, Wagers, and Phillips. *Language (Baltim.)* 88, 390–400. doi: 10.1353/lan.2012.0025

Hofmeister, P., Casasanto, L. S., and Sag, I. A. (2013). "Islands in the grammar? Standards of evidence," in *Experimental Syntax and Island Effects,* eds J. Sprouse and N. Hornstein (Cambridge: Cambridge University Press), 42–63. doi: 10.1017/CBO9781139035309.004

Hofmeister, P., Casasanto, L. S., and Sag, I. A. (2014). Processing effects in linguistic judgment data: (Super-)additivity and reading span scores. *Lang. Cogn.* 6, 111–145. doi: 10.1017/langcog.2013.7

Hofmeister, P., and Sag, I. (2010). Cognitive constraints and island effects. *Language (Baltim.)* 86, 366–415. doi: 10.1353/lan.0.0223

Hutchison, K. (2007). Attentional control and the relatedness proportion effect in semantic priming. *J. Exp. Psychol.* 33, 645–662. doi: 10.1037/0278-7393.33.4.645

Johnson, A. (2015). *Individual Differences in Predictive Processing: Evidence from Subject Filled-Gap Effects in Native and Non-native Speakers of English*. Doctoral dissertation, University of Kansas.

Juffs, A., and Harrington, M. (2011). Aspects of working memory in L2 learning. *Lang. Teach.* 44, 137–166. doi: 10.1017/S0261444810000509

Just, M. A., Carpenter, P. A., and Wooley, J. D. (1982). Paradigms and processes in reading comprehension. *J. Exp. Psychol. Gen.* 3, 228–238. doi: 10.1037/0096-3445.111.2.228

Just, M., and Carpenter, P. A. (1992). A capacity theory of comprehension: individual differences in working memory. *Psychol. Rev.* 99, 122–149. doi: 10.1037/0033-295X.99.1.122

Kim, E., Baek, S., and Tremblay, A. (2015). The role of island constraints in second language sentence processing. *Lang. Acquis.* 00, 1–33. doi: 10.1080/10489223.2015.1115051

Kim, J. H. (2008). *Working Memory Effects on Bilingual Sentence Processing*. Doctoral dissertation. ProQuest Dissertations and Theses.

King, J., and Just, M. (1991). Individual differences in syntactic processing: the role of working memory. *J. Mem. Lang.* 30, 580–602. doi: 10.1016/0749-596X(91)90027-H

Kluender, R. (1992). Deriving island constraints from principles of predication. *Island Constraints* 30, 223–258. doi: 10.1007/978-94-017-1980-3_8

Kluender, R. (1998). On the distinction between strong and weak islands: a processing perspective. *Syntax Semant.* 29, 241–279.

Kluender, R. (2004). "Are subject islands subject to a processing account?" in *Proceedings of West Coast Conference on Formal Linguistics,* Vol. 23, eds A. Rodríguez, V. Chand, A. Kelleher, and B. Scheiser (Somerville, MA: Cascadilla Press), 475–499.

Kluender, R., and Kutas, M. (1993). Subjacency as a processing phenomenon. *Lang. Cogn. Process.* 8, 573–633. doi: 10.1080/01690969308407588

Lee, M. W. (2004). Another look at the role of empty categories in sentence processing (and grammar). *J. Psycholinguist. Res.* 33, 51–73. doi: 10.1023/B:JOPR.0000010514.50468.30

Listening Comprehension Test, LCT (1972). Listening Comprehension Test. Ann Arbor, MI: English Language Institute, The University of Michigan.

Martohardjono, G. (1993). *Wh-Movement in the Acquisition of a Second Language: A Crosslinguistic Study of Three Languages with and without Overt Movement*. Doctoral Dissertation, Cornell University, Ithaca, NY.

McArdle, J. J., Fisher, G. G., and Kadlec, K. M. (2007). Latent variable analyses of age trends of cognition in the health and retirement study, 1992-2004. *Psychol. Aging* 22, 525–545. doi: 10.1037/0882-7974.22.3.525

Nakano, Y., Felser, C., and Clahsen, H. (2002). Antecedent priming at trace positions in Japanese long-distance scrambling. *J. Psycholinguist. Res.* 31, 531–571. doi: 10.1023/A:1021260920232

Nettelbeck, T., and Burns, N. R. (2010). Processing speed, working memory and reasoning ability from childhood to old age. *Pers. Individ. Dif.* 48, 379–384. doi: 10.1016/j.paid.2009.10.032

Nicenboim, B., Vasishth, S., Gattei, C., Sigman, M., and Kliegl, R. (2015). Working memory differences in long-distance dependency resolution. *Front. Psychol.* 6:312. doi: 10.3389/fpsyg.2015.00312

Oberauer, K. (2005). Control of the contents of working memory-a comparison of two paradigms and two age groups. *J. Exp. Psychol. Learn. Mem. Cogn.* 31, 714–728. doi: 10.1037/0278-7393.31.4.714

Omaki, A., Lau, E. F., Davidson White, I., Dakan, M. L., Apple, A., and Phillips, C. (2015). Hyper-active gap filling. *Front. Psychol.* 6:384. doi: 10.3389/fpsyg.2015.00384

Omaki, A., and Schulz, B. (2011). Filler-gap dependencies and island constraints in second language sentence processing. *Stud. Second Lang. Acquis.* 33, 563–588. doi: 10.1017/S0272263111000313

Phillips, C. (2006). The real-time status of island phenomena. *Language (Baltim).* 82, 795–823. doi: 10.1353/lan.2006.0217

Phillips, C. (2013). Some arguments and nonarguments for reductionist accounts of syntactic phenomena. *Lang. Cogn. Process.* 28, 156–187. doi: 10.1080/01690965.2010.530960

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychol. Bull.* 114, 510–532. doi: 10.1037/0033-2909.114.3.510

Roberts, L., Marinis, T., Felser, C., and Clahsen, H. (2007). Antecedent priming at trace positions in children's sentence processing. *J. Psycholinguist. Res.* 36, 175–188. doi: 10.1007/s10936-006-9038-3

Ross, J. (1967). *Constraints on Variables in Syntax*. Doctoral dissertation, MIT.

Slevc, L. R., and Novick, J. M. (2013). Memory and cognitive control in an integrated theory of language processing. *Behav. Brain Sci.* 36, 373–374. doi: 10.1017/S0140525X12002683

Sohn, H. (1980). Theme prominence in Korean. *Korean Linguist.* 2, 2–19. doi: 10.1075/kl.2.01hms

Sohn, H. M. (1999). *The Korean Language*. Cambridge: Cambridge University Press.

Sprouse, J., Wagers, M., and Phillips, C. (2012). A test of the relation between working-memory capacity and syntactic island effects. *Language (Baltim.)* 88, 82–123. doi: 10.1353/lan.2012.0004

Stowe, L. (1986). Parsing wh-constructions: evidence for online gap location. *Lang. Cogn. Process.* 1, 227–245. doi: 10.1080/01690968608407062

Tagliaferri, B. (2005). *Paradigm: Perception Research Systems, Inc.* [Software]. Available online at: http://www.perceptionresearchsystems.com

Traxler, M. J., and Pickering, M. J. (1996). Plausibility and the processing of unbounded dependencies: an eye-tracking study. *J. Mem. Lang.* 35, 454–475. doi: 10.1006/jmla.1996.0025

Wagers, M. W., and Phillips, C. (2009). Multiple dependencies and the role of the grammar in real-time comprehension. *J. Linguis.* 45, 395–433. doi: 10.1017/S0022226709005726

Wagers, M. W., and Phillips, C. (2014). Going the distance: memory and control processes in active dependency construction. *Q. J. Exp. Psychol.* 67, 1274–1304. doi: 10.1080/17470218.2013.858363

Wass, S. V., Scerif, G., and Johnson, M. H. (2012). Training attentional control and working memory-is younger, better? *Dev. Rev.* 32, 360–387. doi: 10.1016/j.dr.2012.07.001

Waters, G., and Caplan, D. (1996). The measurement of verbal working memory capacity and its relation to reading comprehension. *Q. J. Exp. Psychol.* 49A, 51–79. doi: 10.1080/713755607

White, L., and Juffs, A. (1998). "Constraints on wh-movement in two different contexts of nonnative language acquisition: competence and processing," in *The Generative Study of Second Language Acquisition*, eds S. Flynn, G. Martohardjono, and W. O'Neil (Mahwah, NJ: Lawrence Erlbaum), 111–129.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2016 Johnson, Fiorentino and Gabriele. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 12 January 2016
doi: 10.3389/fpsyg.2015.02048

Similarity of *wh*-Phrases and Acceptability Variation in *wh*-Islands

Emily Atkinson, Aaron Apple, Kyle Rawlins and Akira Omaki*

Department of Cognitive Science, The Johns Hopkins University, Baltimore, MD, USA

In *wh*-questions that form a syntactic dependency between the fronted *wh*-phrase and its thematic position, acceptability is severely degraded when the dependency crosses another *wh*-phrase. It is well known that the acceptability degradation in *wh*-island violation ameliorates in certain contexts, but the source of this variation remains poorly understood. In the syntax literature, an influential theory – Featural Relativized Minimality – has argued that the *wh*-island effect is modulated exclusively by the distinctness of morpho-syntactic features in the two *wh*-phrases, but psycholinguistic theories of memory encoding and retrieval mechanisms predict that semantic properties of *wh*-phrases should also contribute to *wh*-island amelioration. We report four acceptability judgment experiments that systematically investigate the role of morpho-syntactic and semantic features in *wh*-island violations. The results indicate that the distribution of *wh*-island amelioration is best explained by an account that incorporates the distinctness of morpho-syntactic features as well as the semantic denotation of the *wh*-phrases. We argue that an integration of syntactic theories and perspectives from psycholinguistics can enrich our understanding of acceptability variation in *wh*-dependencies.

Keywords: relativized minimality, *wh*-island, D-linking, acceptability judgment, amelioration, similarity interference

OPEN ACCESS

Edited by:
Claudia Felser, University of Potsdam, Germany

Reviewed by:
Dario Leander Jim Felix Paape, University of Potsdam, Germany
Ankelien Schippers, Carl von Ossietzky Universität Oldenburg, Germany

****Correspondence:***
Emily Atkinson
atkinson@cogsci.jhu.edu

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: *31 August 2015*
Accepted: *23 December 2015*
Published: *12 January 2016*

Citation:
Atkinson E, Apple A, Rawlins K and Omaki A (2016) Similarity of wh-Phrases and Acceptability Variation in wh-Islands. Front. Psychol. 6:2048. doi: 10.3389/fpsyg.2015.02048

INTRODUCTION

Much work in syntax has investigated the acceptability of English sentences that involve multiple *wh*-phrases, as in (1):

(1) a. **Who** __ wondered **who** bought the car?
b. ***What** did you wonder **who** bought __ ?

Despite the superficial resemblance of sentences in (1), native speakers of English perceive (1a) as a more acceptable sentence of English than (1b). This example illustrates the so-called *wh*-island constraint (Chomsky, 1964, 1977; cf. Ross, 1967): the grammar disallows dependency formation between the fronted *wh*-phrase (e.g., *what*) and its thematic position when there is another intervening *wh*-phrase (*who*). The discovery of this constraint raised a number of empirical and theoretical questions that remain unresolved: what types of representational or derivational constraints underlie the *wh*-island phenomenon? Are all *wh*-islands created equal, such that they all produce a similar degree of degradation? If not, what types of linguistic or cognitive factors affect the acceptability variation in *wh*-island violation?

The present paper aims to shed light on these questions through experimental tests of a recent, influential theory of *wh*-islands, called Featural Relativized Minimality (henceforth Featural RM;

Friedmann et al., 2009; Belletti et al., 2012; Rizzi, 2013; for related proposals, see also Starke, 2001; Boeckx and Jeong, 2003). As the review below illustrates, there are two reasons why this theory deserves ample attention from syntacticians and psycholinguists. First, unlike many syntactic theories that only distinguish grammatical from ungrammatical sentences, Featural RM predicts fine variations in acceptability across different types of *wh*-islands, in particular, how the acceptability of *wh*-island violations can *ameliorate* depending on the similarity of *wh*-phrases. Second, as noted by Rizzi (2013), Featural RM resembles memory constraints on sentence processing, where the similarity of competing words in the sentence often predicts comprehension difficulties. As such, empirical investigations of *wh*-island amelioration effects provide a unique opportunity to explore the link between Featural RM and memory constraints in parsing. We report 4 experiments that explore the empirical predictions of Featural RM, and demonstrate that the theory needs refinement by incorporating aspects of memory encoding and retrieval constraints that guide the real-time computation of syntactic representations.

Featural Relativized Minimality and Similarity Interference in Parsing

The definition of the Featural RM constraint can be summarized as in (2), which is slightly modified from Rizzi (2013) for expository purposes:

(2) In the configuration [. . . X . . . Z . . . Y . . .], X and Y cannot form a dependency if Z c-commands Y, and Z is the same structural type as X.

The syntactic condition as stated in (2) ensures that a *wh*-dependency cannot be established when there is a competing intervener [Z in (2)] that is structurally closer to the thematic position (Y) than the fronted *wh*-phrase (X). In Featural RM, the definition of the *structural type* that constitutes a violation of RM is stated in terms of morpho-syntactic features of those constituents.

A critical empirical observation that led to the use of morpho-syntactic features in Featural RM is the amelioration of *wh*-island violations with a D(iscourse)-linked *wh*-phrase (Pesetsky, 1987). While D-linked *wh*-phrases have been intuitively characterized as linked to previous discourse in some way, we will primarily use it here as a cover-term for *which*-phrases that denote a set of individuals. In the syntax literature, it has been reported that extracting the bare *wh*-phrase *what* from the *wh*-island, as in (3a), results in an ungrammatical sentence, but the extraction of the D-linked *wh*-phrase *which problem* in (3b) is considered marginally grammatical. This suggests that the *wh*-island violation in (3b) is somewhat ameliorated, though its acceptability is still degraded compared to the grammatical *wh*-extraction in (3c).

(3) a. *__What__ do you wonder __who__ solved __?
b. ?__Which problem__ do you wonder __who__ solved __?
c. __Which problem__ do you think that John solved __?

Assuming the acceptability pattern indicated in (3), Rizzi and colleagues proposed that the degree of overlap in morpho-syntactic features of *wh*-phrases accounts for the acceptability variation (Friedmann et al., 2009; Belletti et al., 2012; Rizzi, 2013). For example, the feature relation between the two *wh*-phrases can be characterized as identity (3a), inclusion (3b), and disjunction (3c). In (3a), the extracted constituent and the intervener both contain only a [+Q(uestion)] feature, and hence the feature sets are identical. This *identity* relation results in a severe degradation in acceptability. In (3b), the intervener only contains [+Q], whereas the feature set for the D-linked *wh*-phrase contains [+Q] as well as [+N(oun)], the latter of which represents the "referential status" of the D-linked *wh*-phrase (see Cinque, 1990). This configuration is called an *inclusion* configuration, as the extracted constituent is more richly specified, and its feature set is a superset of that of the intervener. This inclusion relation leads to a less severe degradation in acceptability, and the *wh*-island effect is ameliorated relative to (3a), but the sentence is not necessarily judged as fully acceptable. Finally, in (3c) the embedded clause contains no [+Q] feature, and hence the feature specifications for the extracted constituent and the (potential) intervener are distinct. This is termed a *disjunction* configuration, which leads to no violation of Featural RM. These three feature set relations and their well-formedness statuses are summarized in **Table 1**.

In summary, a key property of Featural RM is that it is concerned with the similarity of the fronted constituent and intervener in terms of morpho-syntactic features: the overlap of features causes degradation, and amelioration is observed when the extracted constituent has a richer or distinct set of morpho-syntactic features than the intervener.

The data discussed above concern the acceptability of sentences, but related observations have been made in adult and child sentence processing research on comprehension of filler-gap dependencies. For example, children experience greater comprehension difficulties with object *wh*-questions like *Which dog did the cat bite __ ?* than *Who did the cat bite __ ?*, possibly due to the overlap of [+N] feature in the fronted *wh*-phrase *which dog* and the intervening NP *the cat* (Friedmann et al., 2009; Belletti et al., 2012; for counter-arguments, see Goodluck, 2010; Bentea and Durrleman, 2014). In adult sentence processing, object relative clauses with two definite Noun Phrases (NPs) like *The banker that the barber praised __* pose greater comprehension difficulties than sentences in which the intervening NP is replaced by a pronoun or a name, as in *The banker that you/John praised__* (Gordon et al., 2001, 2002, 2004, 2006; Warren and Gibson, 2002, 2005). This adult finding may be compatible with Featural RM if we expand the relevant morpho-syntactic features to

TABLE 1 | Taxonomy of feature set and well-formedness in Featural RM.

X *Fronted phrase*	Z *Intervener*	Y *Thematic position*	Well-formedness	Type
+A	+A	<+A>	Ungrammatical (*)	Identity
+A, +B	+A	<+A, +B>	Marginal (?)	Inclusion
+A	+B	<+A>	Grammatical (√)	Disjunction

include features that distinguish definite NPs from pronouns or names.

An alternative explanation, which has received much support from sentence processing as well as domain-general working memory research, is that these observations reflect constraints on memory encoding and retrieval mechanisms, which are subject to so called *similarity-based interference* (Lewis and Vasishth, 2005; for a review, see Van Dyke and Johns, 2012). There are two ways in which similarity-based interference could occur. The first and more well-known type of similarity-based interference is *retrieval interference*. Comprehension of relative clauses or *wh*-questions requires the parser to retrieve the fronted *wh*-phrase and relate it to its thematic position. According to these memory accounts, this retrieval mechanism uses a cue-based search process, and activates all NPs that meet (some of) the search cues. The retrieval competition among candidates with similar features results in comprehension difficulties. The second type is called *encoding interference*. This type of interference is observed when the parser encounters words or phrases that are similar to one another, and the process of encoding and storing them as distinct items in memory is disrupted. The resulting representations that are stored in memory may be less precise or robust, and may require more cognitive resources to retrieve later in the sentence (see Gordon et al., 2002).

This raises questions about whether the variation of acceptability judgments in (3) may also be an instance of similarity-based interference: the identity relation in (3a) causes greater similarity-based interference than the inclusion configuration in (3b), which in turn causes more interference than (3c). In fact, it may even be possible to reduce Featural RM (**Table 1**) to constraints on working memory. However, as noted by Rizzi (2013), one key difference between Featural RM and memory retrieval accounts is that Featural RM is strictly concerned with the overlap of morpho-syntactic features, whereas similarity-based interference is typically sensitive to a variety of similarities, including semantic features (Van Dyke and McElree, 2006; Hofmeister, 2011; Hofmeister and Vasishth, 2014; Kush et al., 2015). Thus, further investigations of the role of semantic overlap in *wh*-island amelioration could shed light on the link between Featural RM and similarity-based interference.

The Present Study

The present study uses acceptability judgment experiments to explore the role of morpho-syntactic and semantic features in amelioration of *wh*-island violations. Specifically, we will explore the acceptability of the inclusion configuration (4a), and how it compares to the acceptability of the D-linked identity configuration (4b).[1]

(4) a. **Which athlete** did she wonder **who** would recruit __? (Inclusion)
b. **Which athlete** did she wonder **which coach** would recruit __? (D-linked identity)

In (4a) the extracted *wh*-phrase is D-linked and the intervener is a bare *wh*-phrase, whereas in (4b), both the extracted *wh*-phrase and the intervener *wh*-phrase are D-linked. Under Featural RM, the dependency in (4b) should be classified as an identity configuration, since both *wh*-phrases have features [+Q, +N]. We will refer to this configuration as *D-linked identity,* to distinguish it from the typical identity configuration [e.g., (3a)] that only includes bare *wh*-phrases. The dependency in (4a) is an inclusion configuration, since the intervening *wh*-phrase only has the feature [+Q]. Given these assumptions about the morpho-syntactic features, Featural RM predicts that (4b) should be less acceptable than (4a). On the other hand, both *wh*-phrases in the D-linked identity configuration (4b) are semantically more specific, as they characterize distinct sets of individuals: a set of athletes and a set of coaches. The *wh*-phrases in (4a) are less distinct because they do not denote distinct sets: the set of athletes is a proper subset of the set of people denoted by *who*. Thus, if semantic distinctness plays a role in dependency formation, the D-linked identity configuration (4b) may cause less similarity-based interference and lead to *wh*-island amelioration, possibly more so than in the inclusion condition (4a).

Informal judgment data reported in the syntax literature (Pesetsky, 1987, 2000; Comorovski, 1996; Shields, 2008) suggest that the D-linked configuration in (4b) should be more acceptable than the inclusion configuration in (4a); in fact, Pesetsky originally annotated them as fully grammatical, in contrast to non-D-linked identity examples. This may challenge the predictions of Featural RM, but it may reflect the fact that differences such as (4a) vs. (4b) are extremely subtle, and the reliability of the data in (4) may be in question. Although D-linked *wh*-phrases are reported to ameliorate *wh*-island violations, those sentences are still often described as unacceptable or ungrammatical to some degree. In other words, sentences like (4a) differ from non-D-linked identity sentences only in the severity of degradation, which is not guaranteed to be readily distinguishable in informal judgments. While D-linked identity examples are often (but not uniformly) annotated as fully grammatical in the linguistics literature, there is evidence that they have a different status than non-D-linked identity examples (Pesetsky, 2000; Shields, 2008). For example, Pesetsky (2000) demonstrates that they, unlike regular grammatical multiple-*wh* examples, e.g., (1a), show intervention effects, e.g., **Which book didn't which person read?* Because the contrasts are empirically subtle and complex, we will use acceptability judgment experiments with a 7-point scale that provide a quantitative measure of acceptability variation. Such experiments have proven useful for a variety of syntactic phenomena that involve subtle contrasts in acceptability intuitions (e.g., McDaniel and Cowart, 1999; Featherston, 2005; Alexopoulou and Keller, 2007; Hofmeister and Sag, 2010; Sprouse et al., 2012; Sprouse and Hornstein, 2013).

In fact, several experimental studies have provided preliminary evidence that semantic information may indeed play a role in island amelioration (Alexopoulou and Keller, 2013; Goodall, 2015; see also Fanselow et al., 2011). Alexopoulou and Keller (2013) investigated the acceptability of extraction out of *whether*-islands (e.g., *What does Claire wonder whether we will watch __ at the cinema?*) while manipulating the animacy

[1]For a related study in French, see Villata et al. (in press).

and D-linking status of the *wh*-phrase (e.g., *what, who, which movie, which colleague*). Here, it was found that bare inanimate *wh*-phrase *what* was less acceptable than the other three *wh*-phrase types, which did not differ from each other. This may suggest that inanimate nouns may be easier to extract out of an island, but this result is difficult to relate to the present study for two reasons. First, the animacy effect did not hold for the D-linked *wh*-phrases, suggesting that this may not be a robust effect. Second, *whether*-islands are different from *wh*-islands in (4) since the intervener (i.e., *whether*) itself does not relate to another (distant) thematic position. Goodall (2015) found clear evidence that D-linked *wh*-phrases ameliorate *wh*-islands that are more similar to those used in the present study. However, his D-linking manipulation compared bare *wh*-phrase against partitive *wh*-phrase (*What / Which of the cars do you wonder who might buy __ ?*). We note that, potentially, this partitive *wh*-phrase may have inflated the amelioration effect for a variety of reasons; for example, it contains a richer semantic content, which is known to facilitate retrieval processes in general (Hofmeister, 2011; Hofmeister and Vasishth, 2014). For this reason, our experiments will focus on D-linking manipulation that does not involve the partitive, in line with the D-linking manipulation that has been used more widely in the syntax literature.

Before presenting the experiments, it is important to clarify the scope of the present paper. The similarity-based interference accounts provide the motivation for the present study, as well as the critical predictions that semantic similarity should also play a role in acceptability variation in *wh*-islands. However, offline acceptability judgment data that we report here does not necessarily shed light on whether the observed acceptability variation in *wh*-islands actually reflects working memory constraints on encoding and retrieval processes during real-time sentence processing. As such, our aim is not to investigate how acceptability variation unfolds during real-time sentence processing, but rather to test whether the ultimate acceptability judgment data is compatible with the predictions of the similarity-based interference accounts.[2]

EXPERIMENT 1

This experiment investigates the acceptability of *wh*-island violations with D-linked identity and *wh*-island violations with an inclusion configuration, where only the extracted phrase is D-linked. We test this using a 2 × 2 design with movement from within a *wh*-island (non-island vs. island) and feature relation (non-identity vs. identity) as factors, as in **Table 2**. The extraction conditions contain extractions out of *wh*-islands. The non-extraction counterparts in do not contain *wh*-island violations and, hence, serve as baseline conditions.

Featural RM predicts that the D-linked identity condition should be severely degraded because the set of features on both D-linked *wh*-phrases (*which NP*, [+Q, +N]) are identical. On the other hand, the inclusion configuration should be less degraded than D-linked identity, because the features on the fronted phrase (*which NP*, [+Q, +N]) are a superset of the features on the intervener (*who*, [+Q]).

TABLE 2 | Sample item set from Experiment 1.

Non-identity	Non-island	Which student __ wondered who would invite the visitor?
	Island	Which visitor did you wonder who would invite __? *(Inclusion)*
Identity	Non-island	Which student __ wondered which teacher would invite the visitor?
	Island	Which visitor did you wonder which teacher would invite __? *(D-linked Identity)*

Method

Participants

Twenty-five self-reported native English speakers were recruited on the internet via Amazon Mechanical Turk, which has proven to be a useful venue in which participants provide reliable acceptability judgment data (Gibson et al., 2011; Sprouse, 2011). They were paid $0.30 for their participation. The data from 3 additional participants was excluded from the analysis, as they only used the extreme ends of the scale in the pre-test phase (see below). This and the following experiments were approved by the Johns Hopkins University Institutional Review Board, and all participants provided informed consent.

Materials

The stimuli for this experiment consisted of 16 sets of bi-clausal *wh*-questions (**Table 2**). These 16 items were counter-balanced across four lists, so that each participant saw only one version of each target item. Forty-eight filler items of comparable length and varying acceptability were randomly interspersed with these target items for a total of 64 items. Based on our informal judgments and acceptability judgment data in the literature, we manipulated the acceptability of filler items to create three groups of fillers: those that are expected to receive high acceptability rating (good fillers), those that are expected to receive low rating (bad fillers), and sentences whose acceptability was expected to fall in between (middle fillers). Fillers consisted of both declaratives and questions, which were included to ensure that the target items were not the only questions in the experiments. Having filler items with varying acceptability serves two purposes. First, this encourages the participants to use a large portion of the scale, which is critical for revealing subtle contrasts. Second, the data from fillers can serve as a baseline measure that can be used to estimate the magnitude of amelioration effects in target sentences. Stimuli from all four experiments, including the fillers, are provided in Supplementary Materials.

[2]While the present study does not directly tap the real-time generation of acceptability intuition, many studies have shown a correspondence between real-time comprehension difficulties and offline judgment data in processing of filler-gap dependencies (see, for example, Gibson and Thomas, 1999; Hofmeister and Sag, 2010; Vasishth et al., 2010; Hofmeister et al., 2013).

Procedure

All of the acceptability judgment experiments in this paper have the same basic procedure. Participants were instructed to rate sentences on a scale from 1 (bad) to 7 (good). Before beginning the experiment, participants were provided with detailed instructions and examples to illustrate that the task is not about stylistic considerations, prescriptive norms, or the plausibility of the event described. This was followed by additional examples with varying degrees of acceptability to illustrate what type of sentence corresponded to different parts of the scale. None of these example sentences used the same structure as the target sentences shown in (5).

Additionally, the first six experimental trials were identical for all participants and served as a pre-test phase. These six trials consisted of two highly acceptable sentences, two highly unacceptable sentences, and two marginal ones. These sentences were included to encourage participants to use the entire scale. The use of a large range of points on the scale was critical for the present study, because the target comparison involves two unacceptable sentence conditions. The acceptability contrast between such sentences may not be revealed if participants used, for example, only the two extreme ends of the scale and treated the task as a binary judgment task. If participants restricted their judgments to the extreme ends of the scale (i.e., 1 and 7) on these initial items, the data from these participants were excluded from further analyses, as it suggests that the participants are treating the scale as if it is a binary choice, which may skew the acceptability ratings in unexpected ways.[3]

Data Analysis

All experiments in this paper use the same data analysis procedure. First, the raw judgment ratings, including both targets and fillers, were converted to *z*-scores within participants (Schütze and Sprouse, 2013). The *z*-score transformation converts a participant's scores to units that represent the number of standard deviations a particular rating is from that participant's mean rating. This procedure corrects for the potential that individual participants treat the scale differently, e.g., using only a subset of the available ratings, because it standardizes all participants' results to the same scale. We also ran the reported analyses with the raw ratings and the results were unchanged in all experiments, although we will only report data and analyses based on *z*-scores.

Linear mixed-effect models were used to analyze the data; these models allow the simultaneous inclusion of random participant and random item variables (Baayen et al., 2008). Each model was fit using the maximal random effects structure that converged (Barr et al., 2013). These models were run in the R environment (R Core Development Team, 2015) using the lme4 package (Bates et al., 2015). *P*-value estimates for the fixed and random effects were calculated using the Sattherwaite approximation in the lmerTest package (Kuznetsova et al., 2015). When the results showed a significant interaction, planned pairwise comparisons were also performed to determine significance between individual conditions. These pairwise comparisons used separate linear mixed-effects models with maximal random effects structure; unlike other statistical analysis methods, mixed-effects models are robust to multiple comparisons.

[3]The overall pattern in our results did not change when the analysis included participants that would be removed according to this criterion. In this paper, we only present data that excluded those participants, as we think that this exclusion increases the chance of veridically representing the acceptability contrasts between conditions.

Results

Figure 1 presents the *z*-score transformed average ratings for each condition and for each filler type. Good filler sentences were rated as most acceptable (mean *z*-score = 0.80), while bad fillers were rated as least acceptable (mean *z*-score = −0.75). Middle fillers received ratings near participants' mean rating (i.e., near a *z*-score of 0, mean = −0.21). This pattern of acceptability for the fillers is common across all four experiments.

For the target items, we found that the island conditions were rated as less acceptable than the non-island conditions (island mean *z*-score = −0.71, non-island mean *z*-score = −0.05). Within the island conditions, the D-linked identity condition is rated as more acceptable than the inclusion condition (−0.58 vs. −0.84). In the non-island conditions, average *z*-scored ratings are around zero (means −0.04 and −0.07), suggesting that they were rated close to individual participants' mean ratings. This likely reflects the fact that sentences with two *wh*-phrases are generally uncommon and difficult to process out of context.

Table 3 presents the estimated coefficients and the standard error for the Linear Mixed Effect model with islandhood and feature relation as fixed effects and random intercepts and slopes for participants and items. Significant effects are marked by their beta estimates.

There is a main effect of islandhood such that *wh*-island violations are significantly less acceptable than non-island violating questions. There is no main effect of feature relation, but there is a significant interaction of islandhood and feature relation. The estimated coefficient of this interaction indicates that the feature combination had a significant effect in the island conditions, but not in the non-island conditions. This is supported by planned pairwise comparisons: the two non-island conditions are not significantly different from one another ($\beta = -0.02$, $SE = 0.12$, $p > 0.1$), while the D-linked identity condition is rated as significantly more acceptable than the inclusion condition ($\beta = 0.26$, $SE = 0.09$, $p < 0.01$).

Discussion

The results indicate that movement out of a *wh*-island generally results in severe degradation of acceptability. More importantly, this degradation is modulated by the feature relation between the two *wh*-phrases: the D-linked identity condition shows greater acceptability than the D-linked inclusion condition. These results replicate informal acceptability judgments in the literature that D-linking ameliorates *wh*-island effects, as well as judgment contrasts that D-linked identity leads to greater acceptability than inclusion (Comorovski, 1996; Shields, 2008). However, these results are not easily explained by the current formulation of

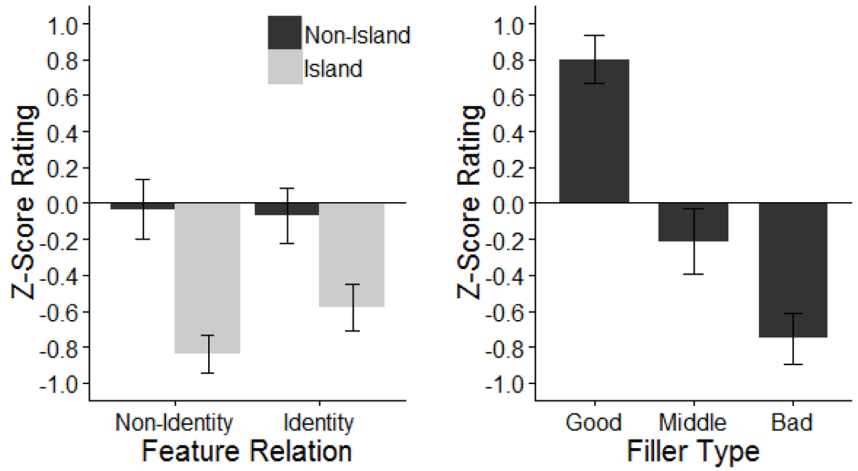


FIGURE 1 | Mean *z*-score acceptability rating of target questions by *wh*-phrase combination and islandhood, and mean *z*-score acceptability rating of filler sentences by filler type. Error bars indicate ± 1 SE.

Featural RM, which predicted that an identity configuration should be more degraded than an inclusion configuration. In fact, our results indicate that the D-linked identity configuration leads to a greater amelioration of the *wh*-island violation than an inclusion configuration.

We have so far focused only on the D-linked identity configuration. No items in this first experiment involve an identity configuration with bare *wh*-phrases, even though Rizzi's (2013) proposal critically relies on an acceptability difference between an identity configuration with bare *wh*-phrases and an inclusion configuration with a fronted, D-linked *wh*-phrase. In order to confirm the presence of *wh*-island amelioration in the inclusion configuration, as predicted by Featural RM, Experiment 2 compares the inclusion condition against a D-linked identity condition as well as a bare identity condition, where both the fronted *wh*-phrase and the intervener are bare *wh*-phrases.

EXPERIMENT 2

Method

Participants

Thirty-two self-reported native English speakers participated via Amazon Mechanical Turk. They were paid $0.50 for participating.

TABLE 3 | Fixed effects summary for Experiment 1 with maximal by-participant and by-item random effects.

	Estimate	*SE*
Intercept	−0.38***	0.08
Islandhood	−0.66***	0.11
Feature relation	−0.03	0.10
Islandhood × Feature relation	0.28*	0.13

*$p \leq 0.05$, **$p \leq 0.01$, ***$p \leq 0.001$.

Materials

The stimuli for this experiment consisted of 24 sets of biclausal sentences, which were constructed by using a 2 × 2 × 2 design with three factors: matrix *wh*-phrase (bare vs. D-linked), feature relation (non-identity vs. identity), and islandhood (non-island vs. island). The experimental conditions shown in **Table 4** include the same four conditions as Experiment 1 (those with a D-linked matrix *wh*-phrase) as well as four new conditions (those with a bare matrix *wh*-phrase) to test Featural RM's broader predictions for *wh*-island amelioration effects. First, the acceptability of the island conditions is predicted to be significantly lower than that of non-island conditions. Second, Featural RM predicts that the identity island conditions should be the most severely degraded compared to all other conditions, including their non-island counterparts. It also predicts that the magnitude of degradation should not differ between the two identity island conditions. Third, the inclusion configuration should yield an amelioration of *wh*-island violations. Thus, the inclusion condition should yield a degradation compared to its non-island counterpart due to a *wh*-island violation, but the resulting acceptability should still be higher than the island identity conditions. Finally, the reverse inclusion configuration and its non-island counterpart are included in the design to test all combinations of the three factors we used in this experiment. The feature set taxonomy of Featural RM (see **Table 1**) does not make explicit predictions for these conditions; however, given that Rizzi and colleagues generally attribute the amelioration effects to the superset-subset relation of feature set between the fronted *wh*-phrase and intervener, we can infer the predictions of Featural RM to be that the acceptability of the reverse inclusion configuration should be similar to that of the two island identity conditions, and lower than the acceptability of the inclusion condition.

These 24 items were counter-balanced across eight lists, so that each participant saw only one version of a target item. Forty-eight filler items of comparable length and varying acceptability were randomly interspersed with these target items.

TABLE 4 | Sample item set from Experiment 2.

Bare matrix *wh*-phrase	Non-identity	Non-island	Who __ wondered which teacher would invite the visitor?
		Island	Who did you wonder which teacher would invite __?
			(Reverse Inclusion)
	Identity	Non-island	Who __ wondered who would invite the visitor?
		Island	Who did you wonder who would invite __?
			(Bare Identity)
D-linked matrix *wh*-phrase	Non-identity	Non-island	Which student __ wondered who would invite the visitor?
		Island	Which visitor did you wonder who would invite __?
			(Inclusion)
	Identity	Non-island	Which student __ wondered which teacher would invite the visitor?
		Island	Which visitor did you wonder which teacher would invite __?
			(D-linked Identity)

Procedure and Data Analysis

This experiment used the same procedure and data analysis steps as Experiment 1. In the statistical analysis, we added planned pairwise comparisons for the island version of the bare identity, inclusion, and D-linked identity conditions, as the comparison of these three conditions is critical for establishing the amelioration of *wh*-island violations that are predicted by Featural RM.

Results

Similar to Experiment 1, all four island conditions were judged as less acceptable than their non-island counterparts (island mean z-score $= -0.54$, non-island mean z-score $= 0.10$), see **Figure 2**. Among the non-island conditions, the non-identity bare matrix *wh*-phrase condition received the highest rating (mean $= 0.25$), but we will leave this aside as it bears no relevance to our goal of testing the predictions of Featural RM. The other non-island conditions were judged similarly with mean z-score ratings around zero (means -0.03, 0.10, and 0.09). Among the island conditions, the D-linked identity condition was rated as the most acceptable (mean $= -0.38$). The remaining three extraction conditions received similar ratings (means -0.57, -0.58, and -0.62).

The Linear Mixed Effect model analysis confirmed that the overall pattern is consistent with Experiment 1. **Table 5** presents the estimated coefficients, the standard error, and the estimated p-value for the Linear Mixed Effect model with islandhood, feature relation, and matrix *wh*-phrase as fixed effects and random intercepts for participants and items.

TABLE 5 | Fixed effects summary for Experiment 2 with by-participant and by-item random intercepts for islandhood, feature relation, and matrix *wh*-phrase type.

	Estimate	*SE*
Intercept	−0.22***	0.05
Islandhood	−0.64***	0.12
Feature relation	−0.02	0.06
Matrix *wh*-phrase	−0.02	0.05
Islandhood × Feature relation	0.26**	0.10
Islandhood × Matrix *wh*-phrase	−0.09	0.10
Feature relation × Matrix *wh*-phrase	−0.26**	0.10
Islandhood × Feature relation × Matrix *wh*-phrase	0.01	0.19

The maximal random effects model did not converge; this model has random slopes for islandhood, feature relation, and their interaction.
*$p \leq 0.05$, **$p \leq 0.01$, ***$p \leq 0.001$.*

As in Experiment 1, there was a main effect of islandhood, but there was no main effect of either feature relation or matrix *wh*-phrase. Importantly, there was an interaction of islandhood and feature relation as well as feature relation and matrix *wh*-phrase, which suggests that the feature relation factor modulates the effects of islandhood or matrix *wh*-phrase type on the acceptability. Planned pairwise comparisons among island conditions revealed no significant difference between the bare identity condition and the inclusion condition ($\beta = 0.04$, $SE = 0.10$, $p > 0.1$). This suggests that the D-linking amelioration effect was not observed for the inclusion configuration. Additionally, there was no significant difference between the inclusion and reverse inclusion conditions ($\beta = 0.06$, $SE = 0.09$, $p > 0.1$). On the other hand, the D-linked identity condition is significantly more acceptable than the inclusion condition ($\beta = 0.23$, $SE = 0.11$, $p = 0.05$), and marginally more acceptable than the bare identity condition ($\beta = -0.19$, $SE = 0.11$, $p < 0.1$). This pattern suggests that the D-linked identity condition showed a reliable amelioration of *wh*-island violations. As reverse inclusion patterns with inclusion, there is no significant difference between reverse inclusion and bare identity ($\beta = -0.01$, $SE = 0.1$, $p > 0.1$), but D-linked identity is marginally more acceptable than reverse inclusion ($\beta = 0.18$, $SE = 0.1$, $p = 0.07$).

Discussion

Replicating the findings from Experiment 1, *wh*-island violations with D-linked identity received a reliably higher acceptability rating than bare identity or inclusion configurations. Furthermore, there was no clear evidence for amelioration of the *wh*-island violation in the inclusion condition. This selective *wh*-island amelioration effect is, again, not easily explained by Featural RM, which predicts that the inclusion configuration should be rated as more acceptable than bare or D-linked identity conditions. Finally, the finding that inclusion and reverse inclusion do not differ in acceptability also conflicts with the predictions of Featural RM.

The absence of an amelioration effect in the inclusion condition was surprising, given that amelioration effects in the inclusion configuration have been widely reported in the literature (Pesetsky, 1987; Cinque, 1990; Alexopoulou and Keller, 2013; Goodall, 2015). Experiment 3 explores whether the animacy of *wh*-phrases may play a role in amelioration of *wh*-island violations.

EXPERIMENT 3

Experiment 2 provided no evidence for *wh*-island amelioration in the inclusion configuration. One plausible source of this unexpected finding is the number of animate nouns in the stimuli. Examples for *wh*-island amelioration in the literature typically included a single animate *wh*-phrase (5a), whereas the stimuli used in Experiment 2 (5b) included two animate *wh*-phrases.

(5) a. Which book did you persuade which person to read __? (Pesetsky, 1987)
b. Which athlete did you wonder who would recruit __? (from **Table 3**)

It is plausible that having two animate *wh*-phrases makes them less distinct from one another, which may have increased confusability or processing demands in our stimuli. As discussed above, this is predicted by the similarity-based interference approach. In order to address this question, Experiment 3 replaces the animate *wh*-phrase [e.g., *which athlete* in (5b)] with an inanimate *wh*-phrase to more closely resemble the examples from the literature.

Method

Participants

Thirty-one self-reported native English speakers participated via Amazon Mechanical Turk. They were paid $0.50 for completing the task.

Materials

The stimuli for this experiment consisted of 24 sets of biclausal sentences, following the same 2 × 2 × 2 design used in Experiment 2, with three factors: islandhood, feature relation, and matrix *wh*-phrase (see **Table 6**). The non-island conditions were identical to those in Experiment 2, where the matrix *wh*-phrase was animate. In the new island conditions, on the other hand, the fronted *wh*-phrase was changed from an animate to an inanimate noun (e.g., *which event*). Because the animacy of the fronted NP has changed, *what* replaces *who* as the bare matrix *wh*-word in the bare identity and reverse inclusion conditions (i.e., *What did you wonder. . .?*).

The 24 items were counter-balanced across eight lists, such that each participant saw only one version of each. Forty-eight filler items of comparable length and varying acceptability were randomly interspersed with these target items for a total of 72 items.

Procedure and Data Analysis

The procedure and data analysis method were identical to those of Experiment 2.

Results

The acceptability judgment pattern in this experiment (**Figure 3**) resembles that of Experiment 2, as the D-linked identity condition received the highest rating among the extraction conditions (−0.06 vs. −0.62, −0.83, and −0.60).

These data were submitted to Linear Mixed Effect model analyses, which used islandhood, feature relation, and matrix *wh*-phrase as fixed effects and random intercepts for participants and items. The coefficient estimates, standard error, and estimated *p*-values are presented in **Table 7**.

The results revealed the same main effect of islandhood as in the previous experiments due to the decreased acceptability of the island violating conditions (island mean = −0.52, non-island mean = 0.11). Also, all three of the pairwise interactions are significant: islandhood and feature relation, islandhood and matrix *wh*-phrase, and feature relation and matrix *wh*-phrase. This suggests that all of these factors influence acceptability, even though the three-way interaction is not significant.

Next, following the data analysis procedure in Experiment 2, planned pairwise comparisons of the island conditions were conducted in order to examine the precise distribution of the amelioration effect. Replicating the results

TABLE 6 | Sample item set from Experiment 3.

Bare matrix *wh*-phrase	Non-identity	Non-island (Animate)	Who __ wondered which family should host the event?
		Island (Inanimate)	What did you wonder which family should host __? *(Reverse Inclusion)*
	Identity	Non-island (Animate)	Who __ wondered who should host the event?
		Island (Inanimate)	What did you wonder who should host __? *(Bare Identity)*
D-linked matrix *wh*-phrase	Non-identity	Non-island (Animate)	Which graduate __ wondered who should host the event?
		Island (Inanimate)	Which event did you wonder who should host __? *(Inclusion)*
	Identity	Non-island (Animate)	Which graduate __ wondered which family should host the event?
		Island (Inanimate)	Which event did you wonder which family should host __? *(D-linked Identity)*

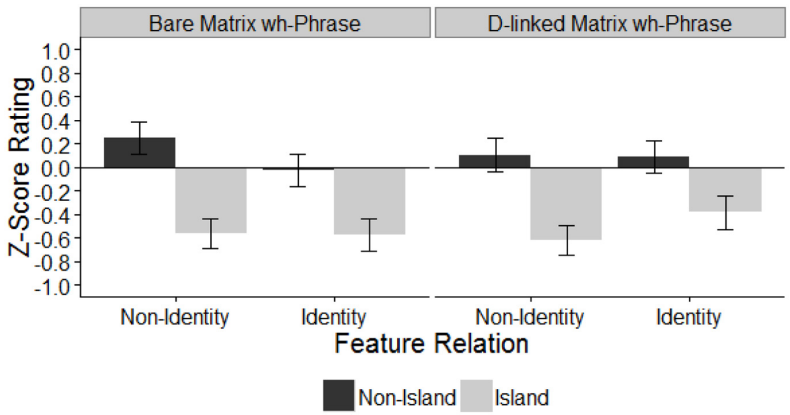


FIGURE 2 | Mean *z*-score acceptability rating in Experiment 2. Error bars indicate ± 1 SE.

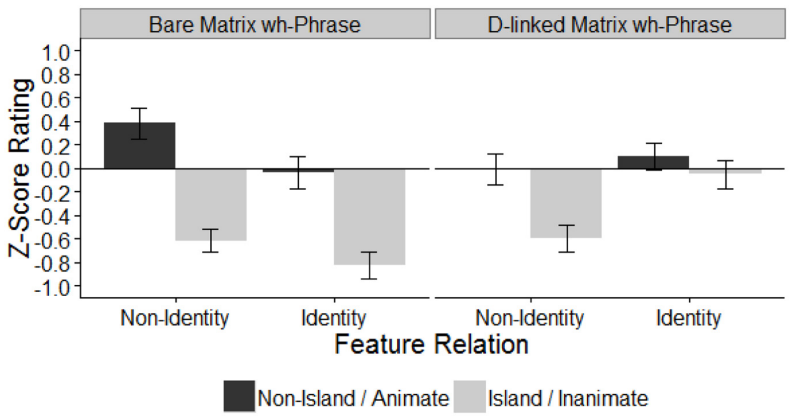


FIGURE 3 | Mean *z*-score acceptability rating in Experiment 3. Error bars indicate ± 1 SE.

of our previous experiments, the D-linked identity condition is significantly more acceptable than the inclusion condition ($\beta = 0.54$, $SE = 0.09$, $p < 0.001$) as well the bare identity condition ($\beta = 0.78$, $SE = 0.12$, $p < 0.001$). Also replicating Experiment 2, no difference was found between the inclusion and reverse inclusion conditions ($\beta = 0.02$, $SE = 0.09$, $p > 0.1$). Importantly, unlike Experiment 2, we found that the inclusion condition is significantly more acceptable than the bare identity condition ($\beta = -0.23$, $SE = 0.09$, $p < 0.05$). Again, reverse inclusion patterns with inclusion, so it is significantly more acceptable than bare identity ($\beta = -0.21$, $SE = 0.09$, $p < 0.05$) and marginally less acceptable than D-linked identity ($\beta = 0.13$, $SE = 0.07$, $p = 0.07$).

Discussion

Once again, this experiment found that the D-linked identity condition was more acceptable than the other island conditions. Also, the reverse inclusion conditions patterned with the inclusion conditions. Unlike Experiment 2, however, we found evidence for *wh*-island amelioration in the inclusion configuration, as the inclusion island condition was judged as more acceptable than the bare identity island condition. The fact that this effect was only found in Experiment 3 could be taken to suggest that the animacy manipulation plays a critical role in its emergence.

However, there are reasons to be cautious of this interpretation. In Experiment 3, island and animacy factors were confounded as the fronted *wh*-phrases were always inanimate in the island conditions. This design does not allow a direct comparison of *wh*-island violations with fronted animate *wh*-phrases to those with inanimate ones. Experiment 4 explores this issue by manipulating animacy within the island conditions.

EXPERIMENT 4

This experiment manipulates animacy and feature relation as in **Table 8**, in order to investigate whether *wh*-island amelioration in inclusion configurations is directly conditioned by the animacy of the fronted *wh*-phrase.

This allowed us to investigate the extent to which animacy contributed to wh-island amelioration effects. Given the results of Experiment 3, we predicted that the contrast between the

TABLE 7 | Fixed effects summary for Experiment 3 with by-participant and by-item random intercepts for extraction type, feature relation, and matrix *wh*-phrase type.

	Estimate	*SE*
Intercept	−0.21***	0.04
Islandhood	−0.63***	0.08
Feature relation	0.02	0.06
Matrix *wh*-phrase	0.14**	0.05
Islandhood × Feature relation	0.31**	0.11
Islandhood × Matrix *wh*-phrase	0.52***	0.09
Feature relation × Matrix *wh*-phrase	0.62***	0.09
Islandhood × Feature relation × Matrix *wh*-phrase	0.25	0.19

The maximal random effects model did not converge; this model has random slopes for islandhood, feature relation, and their interaction.
*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.*

TABLE 8 | Sample item set from Experiment 4.

Animate	Bare Identity	Who did you wonder who should host __?
	Inclusion	Which visitor did you wonder who should host __?
Inanimate	Bare Identity	What did you wonder who would host __?
	Inclusion	Which event did you wonder who should host __?

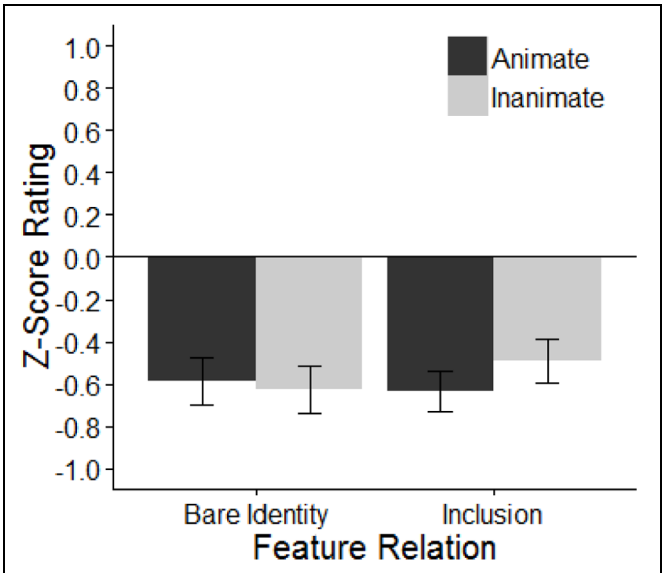


FIGURE 4 | Mean *z*-score acceptability rating in Experiment 4. Error bars indicate ± 1 SE.

inclusion and bare identity conditions should only appear in conditions with an inanimate *wh*-phrase.

Method

Participants

Twenty-nine self-reported native English speakers participated via Amazon Mechanical Turk. They were paid $0.50 for completing the experiment. Three additional participants were excluded for using a single value ($n = 1$) or only the extremes of the scale ($n = 2$) during the calibration items.

Materials

The stimuli for this experiment consisted of 24 sets of biclausal sentences with a 2 × 2 design (**Table 8**), using animacy of the matrix *wh*-phrase (animate vs. inanimate) and feature relation (bare identity vs. inclusion) as factors. These items were largely based on stimuli from the previous experiments. The 24 test items were counter-balanced across four lists, such that each participant only rated a single item from each set. The addition of 48 length-matched filler sentences resulted in a total of 72 items.

Procedure and Data Analysis

The procedure and data analysis method were identical to those of previous experiments. Regardless of the presence of a significant interaction, planned pairwise comparisons of feature relation within animacy were conducted to directly test whether the amelioration effect of inclusion was modulated by animacy of the fronted *wh*-phrase.

Results

Figure 4 presents the mean *z*-score ratings in each condition. Overall, inanimate *wh*-phrase conditions are rated as more acceptable than those with animate *wh*-phrases (inanimates = −0.55, animates = −0.61), but the bare identity and inclusion conditions show little difference in their acceptability ratings (bare identity = −0.59, inclusion = −0.57). Within the animate conditions, bare identity and inclusion show little difference in their acceptability ratings (−0.59 vs. −0.63). Within the inanimate conditions, however, inclusion was rated as more acceptable than bare identity (−0.51 vs. −0.60).

These data were analyzed using a Linear Mixed Effect model analysis with feature relation and animacy as fixed effects. The coefficient estimates, standard error and estimated *p*-values are given in **Table 9**.

The model revealed no main effect of animacy or feature relation, but there was a marginal interaction between the two factors. Planned pairwise comparisons revealed that inclusion was marginally more acceptable than bare identity when the extracted *wh*-phrase was inanimate (inanimate: $\beta = 0.13$, $SE = 0.07$, $p < 0.1$), but not when the extracted phrase was animate ($\beta = -0.04$, $SE = 0.07$, $p > 0.1$).

Discussion

This experiment investigated whether the animacy distinctness between two *wh*-phrases is a pre-requisite for *wh*-island amelioration in inclusion configurations. The results provide weak support for this hypothesis: when the fronted *wh*-phrase was animate, there was little difference between bare identity and inclusion conditions, but there was a marginal difference between these configurations when the fronted *wh*-phrase was inanimate. This finding has two implications. First, the results of Experiments 3 and 4 taken together suggest that the animacy of the extracted *wh*-phrases can modulate *wh*-island amelioration effects, but that the effect can be weak. Second, *wh*-island amelioration in inclusion configurations is generally not as robust

as it has been reported in the literature; a weak amelioration may emerge when the fronted *wh*-phrase and intervener are distinct in animacy, but its effect is clearly not as consistently present as the amelioration effect observed in D-linked identity configuration in Experiments 1 through 3.

GENERAL DISCUSSION

The main goal of this study was to investigate the distribution of *wh*-island amelioration effects, and the extent to which they are modulated by morpho-syntactic and semantic features of *wh*-phrases. Specifically, we tested the acceptability of a *wh*-island violation involving two D-linked *wh*-phrases (i.e., D-linked identity) against violations with an intervening bare *wh*-phrase (i.e., inclusion) or with no D-linked *wh*-phrases (i.e., bare identity).

There are two main findings from the experiments reported above. First, we found consistent evidence against the predictions of Featural RM about D-linked identity configurations: such configurations reliably led to a higher acceptability than inclusion configurations. Featural RM predicts the opposite. Moreover, a study that was conducted in parallel in French used a similar design to our Experiment 3 and found the same pattern (Villata et al., in press). Thus, the increased acceptability of the D-linked identity configuration is robust across experiments and across English and French.

Second, we found that the D-linking amelioration effect for *wh*-island violations can be modulated by animacy, although the animacy effects were not always robust. Experiment 2 used only animate *wh*-phrases and found no evidence for *wh*-island amelioration in the inclusion configuration. Experiment 3 used inanimate nouns for extracted *wh*-phrases, and revealed evidence for amelioration in the inclusion configuration. This contrast between the experiments suggests that animacy might play a role. However, this effect did not hold robustly in Experiment 4, which showed that the amelioration effect was somewhat stronger for inclusion configuration than bare identity condition, which in turn showed no sign of amelioration regardless of the animacy manipulation. While a complete understanding of the role of animacy or the status of the inclusion configuration awaits further research, it is safe to conclude at this point that the *wh*-island amelioration effects for the inclusion configuration are not as robust as it has been reported in the literature.

TABLE 9 | Fixed effects summary for Experiment 4 with by-participant and by-item random intercepts for feature relation and animacy of the matrix *wh*-phrase.

	Estimate	*SE*
Intercept	−0.59***	0.05
Feature relation	−0.05	0.06
Animacy	0.05	0.05
Feature relation × Animacy	−0.18†	0.11

†$p \leq 0.1$, *$p \leq 0.05$, *$p \leq 0.01$, ***$p \leq 0.001$.

These findings are summarized in (6), which depicts the ranking of acceptability variation among the *wh*-island violations that were examined in this paper. We will now discuss the theoretical implications of these findings.

(6) Bare identity ≤ (Reverse) inclusion with an animate *wh*-phrase extraction ≤ (Reverse) inclusion with an inanimate *wh*-phrase extraction < D-linked identity ≤ no extraction

Implications for Featural RM

Our data suggests that Featural RM does not fully account for the distribution of *wh*-island amelioration effects, especially the fact that the D-linked identity configuration led to a robust amelioration effect. We do not present this as an argument against Featural RM *per se*, but minimally something else must be said to account for the behavior of D-linked *wh*-items beyond the inclusion/identity featural distinction. One potential implication is that the set of morpho-syntactic features assumed in papers by Rizzi and colleagues may need to be enriched. We will explore below the addition of Topic or Animacy features, but demonstrate that neither of these features provides a satisfactory explanation.

Rizzi (personal communication) suggests that the extracted D-linked *wh*-phrase has a [+Topic] feature that the intervening D-linked *wh*-phrase does not, as this feature is only licensed by the left periphery of the matrix clause (for a similar suggestion that the extracted *wh*-phrase may have a presupposition feature, see Grohmann, 2000; Boeckx and Jeong, 2003). If this is the case, then the sentences with two D-linked phrases are cases of inclusion rather than identity (7).

(7) **Which athlete** did you wonder **which coach** would recruit __?
[+Q, +N, +Topic] [+Q, +N] [+Q, +N, +Topic]

This amendment allows Featural RM to account for the increased acceptability of the D-linked identity configuration. However, this featural augmentation does not explain why this configuration should be reliably more acceptable than the inclusion condition with a bare *wh*-phrase in the intervener position. Given the feature sets assumed in (7), both of these configurations are inclusion configurations, which are not predicted to show a contrast in acceptability. If we were to grade acceptability based on the degree of featural overlap, the prediction would again go the wrong direction: the bare inclusion condition should have less featural overlap, and therefore be more acceptable than the D-linked identity condition under the analysis in (7).

Another morpho-syntactic feature that may deserve to be added to the Featural RM framework is an animacy feature. It is typically assumed that animacy features do not actively participate in syntactic operations in English. However, animacy is known to play important roles in syntax of other languages (e.g., Slavic languages, see Rappaport, 2003). Our observations of superior *wh*-island amelioration effects for inanimate *wh*-phrases may be the first evidence that animacy plays an important role in English syntax as well. However, the addition of an animacy feature with the same status as e.g., [+Q] above is not fully

motivated by our data either. First, it offers no explanation for the observed acceptability contrast between the D-linked identity and inclusion configuration in Experiments 1 and 2. Second, using animacy features in Experiment 3 would change the D-linked identity feature relation to that of a reverse inclusion, as shown in (8). Under this configuration, Featural RM predicts the sentence to be equally as degraded as identity configurations, which is the opposite of what was found in Experiment 3. Rather, if Experiment 3 is taken at face value, (8) should be ameliorated simply because the two D-linked *wh*-phrases have a different value for animacy.

(8) **Which award** did you wonder **which actress** should receive __?
[+Q, +N] [+Q, +N, +animate] [+Q, +N]

Finally, incorporating an animacy feature would predict that animacy based amelioration effects hold robustly across all *wh*-island violations, but this prediction is inconsistent with the observation in Experiment 4 that the animacy manipulation showed a selective, weak modulation of the acceptability of the inclusion conditions but not the bare identity configuration. While an animacy distinction is clearly relevant, it cannot easily be captured in featural terms.

In summary, it is not obvious what featural adjustments could account for the amelioration patterns we have shown in this paper in a way that is entirely internal to the principles of Featural RM.[4] If this effect cannot be accounted for with featural manipulations, then (minimally) something external to the featural system must lead to the amelioration pattern.

Memory Constraints and Semantic Distinctness in Acceptability Variation

More generally, these results present a challenge to any account of *wh*-island effects that assumes that D-linked identity examples are acceptable or fully amelioriated: the variable amelioration effect for even this case suggests that some constraint like Relativized Minimality may well be active (in contrast to accounts of D-linking that simply assign it a different LF where the constraint leading to the violation is not at play; Pesetsky, 1987, 2000 on superiority). An explanation for the distribution of *wh*-island amelioration effects in our experiments must take into account the superior amelioration effects in D-linked identity configurations, as well as the fact that extraction of an inanimate *wh*-phrase sometimes leads to a further increase in acceptability. Before we present such explanations, we first argue for a new descriptive generalization: the degree of semantic distinctness of the extracted *wh*-phrase and the intervener (rather than the distinctness of morpho-syntactic features) predicts the distribution of *wh*-island amelioration effects.

We suggest that participants in these experiments were able, to varying degrees, to use *semantic distinctness*, rather than morphosyntactic distinctness, as a strategy for interpreting ill-formed *wh*-island examples. First, we will adopt a broadly Hamblin semantics of *wh*-questions, and assume that (i) questions denote a set of possible answers (Hamblin, 1973; see also Karttunen, 1977, and many others), and (ii) *wh*-phrases denote a set of potential referents (Hamblin, 1973; Kratzer and Shimoyama, 2002). Intuitively, the set of referents for the *wh*-item in a single-*wh* question corresponds to possible fragment NP answers to that question. Under this family of assumptions, bare *wh*-phrases like *who* denote the set of all human individuals, whereas a D-linked *wh*-phrase like *which award* would denote a presupposed set of entities satisfying the NP restrictor, in this case award, and require the answer to the *wh*-question to be constructed from some referent in this set only. With these assumptions, let us examine the distinctness of sets of individuals or objects denoted by *wh*-phrases in **Table 10**, which illustrates the main feature configurations that were investigated in our acceptability judgment experiments.

In the bare identity condition with *who* as an extracted *wh*-phrase, both the extracted *wh*-phrase and the intervener denote the set of all humans, and therefore their domains are identical and non-distinct. If the extracted *wh*-phrase is *what,* we assume that *what* denotes a set of everything in the world, which includes human individuals.[5] Here, the set denoted by *what* is a superset of the set denoted by *who*, and these sets are thus overlapping. As for the inclusion configuration with animate *wh*-phrases, *which visitor* denotes a presupposed set of visitors, while *who* denotes a set of all human individuals. Thus, the sets of individuals denoted by these two *wh*-phrases are also overlapping. On the other hand, for the inclusion configuration with inanimate and animate *wh*-phrases, the set denoted by *which event* and the set denoted by *who* are distinct. This explains the amelioration effect that was observed in the comparison of Experiments 2 and 3. Finally, in the D-linked identity conditions, the sets of individuals or objects denoted by the two *wh*-phrases (*which visitor* and *which family,* or *which event* and *which family*) are clearly distinct. Thus, these observations lead to the generalization that the *wh*-island violations that were amenable to amelioration effects were those in which the sets denoted by the extracted *wh*-phrase and the intervener are distinct. We take this as a necessary condition for *wh*-island amelioration.

[4]One reviewer suggested the inclusion of both a topic and an animacy feature. In example (8), this would result in a configuration known as *intersection*, where the fronted *wh*-phrase and the *wh*-intervenor have distinct sets of features that share a subset (in the terms of **Table 1**, the fronted phrase is [+A,+B] and the intervener is [+A,+C]) (Belletti et al., 2012). However, it is unclear whether this configuration should pattern with disjunction or with intersection in acceptability judgments. Additionally, this does not address the concern that example (7) becomes a case of intersection with the addition of these features. It is still unclear why sentences like (7) are consistenly more acceptable than the other cases of inclusion included in our experiments.

[5]There are three empirical reasons for assuming that *what* is underspecified for human or animacy features, and therefore is able to denote humans (see Grosu, 2003). First, *what* can be combined with either animate or inanimate nouns to form complex *wh*-phrases (e.g., *What doctor did you see? What textbook did you buy?*), whereas this type of composition is not possible for *wh*-phrases like *who* with clear human and animacy feature specification (**who doctor*). Second, the answer to *what* can be human or non-human, especially when there are multiple answers (e.g., *What can you see? John, Mary, and a tree.*). This is not possible for *wh*-phrases that are specified for human features (e.g., *Who did you see? *John, Mary, and a tree.*). Third, free relative clauses with *what* can take a human or a non-human referent (Grosu, 2003). In a sentence like *What I thought was a policeman was just a log*, the *wh*-phrase *what* is treated as human (a policeman) internally to the free relative clause, whereas it is treated as inanimate (a log) externally.

TABLE 10 | Distribution of amelioration effects and semantic distinctness.

Conditions	Sentence	Amelioration?	Semantic distinctness
Bare identity	Who/what did you wonder who would host__?	No	Non-distinct
Inclusion (animate)	Which visitor did you wonder who would host __?	No	Non-distinct
Inclusion (inanimate)	Which event did you wonder who would host __?	Maybe?	Distinct
D-linked identity	Which visitor did you wonder which family would host __?	Yes	Distinct
D-linked identity	Which event did you wonder which family would host __?	Yes	Distinct

The semantic distinctness of the *wh*-phrases provides the beginnings of an explanation of many of the patterns in our data, but clearly we do not have evidence for any sort of categorical amelioration; in fact, our results could be taken as evidence against it. One possible explanation for this state of affairs is that similarity-based interference during memory retrieval operations is sensitive to the semantic distinctness of two *wh*-phrases. As noted in the Introduction, it has been widely observed that the processing of filler-gap dependencies can be impeded when the dependencies contain two similar NPs. This similarity interference effect is considered to follow from limitations of the memory system in either encoding two similar NPs as distinct items, or in retrieving the target NPs with accurate syntactic and semantic features. It is plausible that the semantic distinctness of *wh*-phrases modulates the ease of encoding or retrieval processes, and when these processes are readily performed, participants may perceive the *wh*-island violations to be less severely degraded. In this sense, the semantic distinctness of *wh*-phrases may serve as a formal characterization of NPs that are particularly confusable for memory operations.

This psycholinguistic explanation for the role of semantic distinctness and memory constraints has implications for theories of islands and syntactic amelioration effects in general. We suggest two potential approaches for integrating syntactic and psycholinguistic constraints, both of which are equally compatible with our findings. The first approach is to reduce island constraints to cognitive constraints on memory operations, such that "island violations" merely reflect difficulties in establishing *wh*-dependencies during real-time parsing (Kluender and Kutas, 1993; Hofmeister and Sag, 2010; for related explanations for Superiority effects, see Hofmeister et al., 2013). With respect to *wh*-islands, according to this reductionist approach, what used to be considered violations of Featural RM constraints would be reanalyzed as severe instances of similarity-based interference effects, which are sensitive to both syntactic and semantic features of retrieval candidates. Simplifying the theory of grammar and postulating fewer constraints that are specific to linguistic representations is a welcome result (Chomsky, 1995; Phillips, 2013), and it highlights how syntactic theories can be refined by a further collaboration between linguistics and broader cognitive science research. The future agenda for this approach includes extension of experimental investigations to other syntactic phenomena that Featural RM provided explanations for (e.g., intervention effects in *combien* extraction in French; Obenauer, 1983, 1994), as well as addressing counter-arguments for cognitive explanations of island constraints (Sprouse et al., 2012; see also Phillips, 2006). We leave these questions for future research.

The second approach for integrating syntactic constraints on *wh*-dependency formation and memory constraints is to situate similarity interference effects in *repair processes* that the parser initiates in order to cope with a violation of formal, syntactic constraints; we term this approach the Amelioration-as-Repair hypothesis. This explanation of amelioration effects relies on the following three assumptions. First, we assume that acceptability judgment intuitions minimally reflect the well-formedness of syntactic derivations and semantic representations that the parser assigns to a given sentence. When this process fails due to linguistic or other cognitive constraints, we perceive degradation in sentence acceptability (Schütze, 1996), and the severity of degradation reflects the number of constraint violations at all levels of representations (Legendre et al., 1991; Keller, 2000; Smolensky and Legendre, 2006; Haegeman et al., 2014). Second, we also assume that syntactic constraints on *wh*-islands do play an important role in accounting for the general acceptability degradation due to extraction out of *wh*-islands, and this constraint could be the original Relativized Minimality constraint in Rizzi (1990, 2004) which did not distinguish bare identity *wh*-island from inclusion *wh*-island. Finally, we also assume that in the face of sentences that violate syntactic constraints, the parser attempts to repair the structure in order to assign an interpretation to the structurally unintegrated *wh*-phrase. Such interpretive repair processes are well documented in the psycholinguistics literature on severe garden-path sentences (e.g., Christianson et al., 2001; Ferreira and Patson, 2007). While this style of repair may not "cancel" the initial violation of syntactic constraints, it would at least provide a strategy for obtaining a legitimate semantic representation for the sentence that can be passed onto the interpretive process.

Given these assumptions, acceptability judgment data should reflect the degree to which this repair process is able to (a) identify a gap position inside an island, and (b) retrieve the relevant *wh*-phrase in order to complete the *wh*-dependency for the semantic representation. Under the Amelioration-as-Repair approach, it is during this repair/retrieval process that the similarity interference effects arise. It is well known that the parser typically respects island constraints during real-time sentence processing (e.g., Stowe, 1986; Traxler and Pickering, 1996); thus, initially the parser should generate an ungrammatical structure with no gap for the *wh*-phrase. This syntactic violation initiates the repair process, and the search for a gap inside an island. This search process identifies a verb with a missing complement, which indicates that the verb could be a host for the gap. This gap identification subsequently triggers a retrieval of a *wh*-phrase,

using the thematic role and morphological features as retrieval cues.[6] This retrieval process should be sensitive to the semantic distinctness of *wh*-phrases. If the repair process fails due to similarity interference effects (e.g., in the bare identity condition), the semantic representation would veridically reflect the syntactic violation of the *wh*-island constraint (i.e., no gap for the *wh*-phrase), and the sum of these two violations results in more severe degradation. On the other hand, if the parser identifies a gap inside an island due to the lack of similarity interference effects (e.g., in D-linked identity conditions with semantically distinct *wh*-phrases), the resulting semantic representation no longer contains any violation, even though it is derived from a structure that does, and therefore the only source of acceptability degradation is the initial violation of the *wh*-island constraint (see Huang, 1982 for arguments that the semantic representation of islands with argument gaps does not incur any violation).

One consequence of the Amelioration-as-Repair hypothesis is that it provides a new direction toward a mechanistic understanding of acceptability judgment in general. To this day, even though acceptability judgment data has served as the primary source of data for linguists, there is very little theory of how such intuitions arise (cf. Schütze, 1996), or how the process of judging sentence acceptability reflects psycholinguistic constraints. As such, regardless of whether island constraints or Featural RM should remain as a formal constraint on linguistic representations, integration of perspectives and insights from psycholinguistics could help advance the field of syntax.

Finally, we note that either approach raises new research questions that need to be addressed in future research. First, the current study does not provide time course measures that shed light on the memory encoding and retrieval mechanisms that are assumed under either explanation. Second, it remains to be answered why the animacy-based modulation of *wh*-island amelioration effects was not reliably observed across experiments. Following the psycholinguistic explanations above, we tentatively suggest that the real-time encoding and comparison of semantic distinctness information could be subject to a variety of conceptual or cognitive factors that will then impact the behavior of amelioration. For example, accessing the set of all individuals denoted by *who* may be inherently complex when it is presented out of context, as in the current experiments. This difficulty may sometimes mask the potential advantage of semantic distinctness in the inclusion configuration with an inanimate *wh*-phrase, suggesting also that it may not be generally safe to test amelioration effects out of context.

[6] It is also plausible that the animacy effect observed in our experiments reflects the fit of the verb semantic retrieval cues and the *wh*-phrases (e.g., *event* may be a better object for *host* than *visitor*). Testing this hypothesis requires a careful control of verb-noun co-occurrence frequency and plausibility. We leave this question open for future research.

CONCLUSION

The present study investigated the distribution of *wh*-island amelioration effects, with a special focus on how it is modulated by morpho-syntactic features and semantic features of *wh*-phrases. We found that morpho-syntactic features alone, such as those to which Featural RM in its current form appeals, failed to account for the distribution of *wh*-island amelioration effects. We suggested that a full explanation of our results requires the consideration of semantic representations, which may, in turn, be related to constraints on the sentence processing mechanisms that give rise to similarity interference effects. This observation calls for future work that re-examines amelioration effects in other syntactic environments in light of constraints on sentence processing mechanisms.

ACKNOWLEDGMENTS

This work was supported in part by NSF BCS-1423117 to AO, and NSF BCS-1344269 to KR and AO. Our thanks to Eleanor Chodroff and Bob Wiley for their contributions to Experiment 1.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpsyg.2015.02048

REFERENCES

Alexopoulou, T., and Keller, F. (2007). Locality, cyclicity, and resumption: at the interface between the grammar and the human sentence processor. *Language* 83, 110–160. doi: 10.1353/lan.2007.0001

Alexopoulou, T., and Keller, F. (2013). "What vs. who and which: kind-denoting fillers and the complexity of whether-islands," in *Experimental Syntax and Island Effects*, eds J. Sprouse and N. Hornstein (Baltimore, MD: Cambridge University Press), 310–340.

Baayen, R. H., Davidson, D. J., and Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *J. Mem. Lang.* 59, 390–412. doi: 10.1016/j.jml.2007.12.005

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2015). *lme4: Linear Mixed-Effects Models using Eigen and S4. R Package Version* 1.1-9. Available at: https://CRAN.R-project.org/package=lme4

Belletti, A., Friedmann, N., Brunato, D., and Rizzi, L. (2012). Does gender make a difference? Comparing the effect of gender on children's comprehension of relative clauses in Hebrew and Italian. *Lingua* 122, 1053–1069. doi: 10.1016/j.lingua.2012.02.007

Bentea, A., and Durrleman, S. (2014). "Children don't like restrictions: evidence from the acquisition of object A'-dependencies in French," in *Proceedings of the 39th Annual Boston University Conference on Language Development,* Boston, MA.

Boeckx, C., and Jeong, Y. (2003). "The fine structure of syntactic intervention," in *Proceedings of the Thirty-First Western Conference on Linguistics*, Vol. 14, eds B. Agbayani, P. Koshkinen, and V. Samiian (Fresno, CA: California State University, Department of Linguistics Publications), 33–41.

Chomsky, N. (1964). *Current issues in Linguistic Theory*. Hague: Mouton.

Chomsky, N. (1977). "On wh-movement," in *Formal Syntax*, eds P. Culicover, T. Wasow, and A. Akmajian (New York, NY: Academic Press), 71–132.

Chomsky, N. (1995). *The Minimalist Program*. Cambridge, MA: MIT Press.

Christianson, K., Hollingworth, A., Halliwell, J. F., and Ferreira, F. (2001). Thematic roles assigned along the garden path linger. *Cogn. Psychol.* 42, 368–407. doi: 10.1006/cogp.2001.0752

Cinque, G. (1990). *Types of A'-dependencies. Linguistic Inquiry Monographs* 17. Cambridge, MA: MIT Press.

Comorovski, I. (1996). *Interrogative Phrases and the Syntax-Semantics Interface*. New York, NY: Springer.

Fanselow, G., Schlesewsky, M., Vogel, R., and Weskott, T. (2011). Animacy effects on crossing wh-movement in German. *Language* 494, 657–683.

Featherston, S. (2005). That-trace in German. *Lingua* 115, 1277–1302. doi: 10.1016/j.lingua.2004.04.001

Ferreira, F., and Patson, N. D. (2007). The "good enough" approach to language comprehension. *Lang. Linguist. Compass* 1, 71–83. doi: 10.1111/j.1749-818X.2007.00007.x

Friedmann, N., Belletti, A., and Rizzi, L. (2009). Relativized relatives: types of intervention in the acquisition of A' dependencies. *Lingua* 119, 67–88. doi: 10.1016/j.lingua.2008.09.002

Gibson, E., Piantadosi, S., and Fedorenko, K. (2011). Using mechanical Turk to obtain and analyze English acceptability judgments. *Lang. Linguist. Compass* 5, 509–552. doi: 10.1111/j.1749-818X.2011.00295.x

Gibson, E., and Thomas, J. (1999). Memory limitations and structural forgetting: the perception of complex ungrammatical sentences as grammatical. *Lang. Cogn. Process.* 14, 225–248. doi: 10.1080/016909699386293

Goodall, G. (2015). The D-linking effect on extraction from islands and non-islands. *Front. Psychol.* 5:1493. doi: 10.3389/fpsyg.2014.01493

Goodluck, H. (2010). Object extraction is not subject to Child Relativized Minimality. *Lingua* 120, 1516–1521. doi: 10.1016/j.lingua.2009.10.005

Gordon, P. C., Hendrick, R., and Johnson, M. (2001). Memory interference during language processing. *J. Exp. Psychol.* 27, 1411–1423.

Gordon, P. C., Hendrick, R., and Johnson, M. (2004). Effects of noun phrase type on sentence complexity. *J. Mem. Lang.* 51, 97–114. doi: 10.1016/j.jml.2004.02.003

Gordon, P. C., Hendrick, R., Johnson, M., and Lee, Y. (2006). Similarity-based interference during language comprehension: evidence from eye tracking during reading. *J. Exp. Psychol.* 32, 1304–1321.

Gordon, P. C., Hendrick, R., and Levine, W. H. (2002). Memory-load interference in syntactic processing. *Psychol. Sci.* 13, 425–430. doi: 10.1111/1467-9280.00475

Grohmann, K. K. (2000). *Prolific Peripheries: A Radical View from the Left*. Doctoral dissertation, University of Maryland, College Park, MD.

Grosu, A. (2003). A unified theory of 'standard' and 'transparent' free relatives. *Nat. Lang. Linguist. Theory* 21, 247–331. doi: 10.1023/A:1023387128941

Haegeman, L., Jiménez-Fernández, A., and Radford, A. (2014). Deconstructing the subject condition in terms of cumulative constraint violation. *Linguist. Rev.* 31, 73–150. doi: 10.1515/tlr-2013-0022

Hamblin, C. L. (1973). Questions in montague English. *Found. Lang.* 10, 41–53.

Hofmeister, P. (2011). Representational complexity and memory retrieval in language comprehension. *Lang. Cogn. Process.* 26, 376–405. doi: 10.1080/01690965.2010.492642

Hofmeister, P., Jaeger, T. F., Arnon, I., Sag, I. A., and Snider, N. (2013). The source ambiguity problem: distinguishing the effects of grammar and processing on acceptability judgments. *Lang. Cogn. Process.* 28, 48–87. doi: 10.1080/01690965.2011.572401

Hofmeister, P., and Sag, I. A. (2010). Cognitive constraints and island effects. *Language* 86, 366–415. doi: 10.1353/lan.0.0223

Hofmeister, P., and Vasishth, S. (2014). Distinctiveness and encoding effects in online sentence comprehension. *Front. Psychol.* 5:1237. doi: 10.3389/fpsyg.2014.01237

Huang, C. J. (1982). *Logical Relations in Chinese and the Theory of Grammar*. Doctoral dissertation, Cambridge, MA: MIT Press.

Karttunen, L. (1977). Syntax and semantics of questions. *Linguist. Philos.* 1, 1–44. doi: 10.1007/BF00351935

Keller, F. (2000). *Gradience in Grammar: Experimental and Computational Aspects of Degrees of Grammaticality*. Doctoral dissertation, University of Edinburgh, Edinburgh.

Kluender, R., and Kutas, M. (1993). Subjacency as a processing phenomenon. *Lang. Cogn. Process.* 8, 573–633. doi: 10.1080/01690969308407588

Kratzer, A., and Shimoyama, J. (2002). "Indeterminate pronouns: the view from Japanese," in *Proceedings of the Third Tokyo Conference on Psycholinguistics*, ed. Y. Otsu (Tokyo: Hituzi Syobo), 1–25.

Kush, D., Johns, C. L., and Van Dyke, J. A. (2015). Identifying the role of phonology in sentence-level reading. *J. Mem. Lang.* 79, 18–29. doi: 10.1016/j.jml.2014.11.001

Kuznetsova, A., Brockhoff, B., and Christensen, H. B. (2015). *lmerTest: Tests in Linear Mixed Effects Models. R Package Version 2.0–29*. Available at: http://CRAN.R-project.org/package=lmerTest

Legendre, G., Miyata, Y., and Smolensky, P. (1991). "Unifying syntactic and semantic approaches to unaccusativity: A connectionist approach," in *Proceedings of the Seventeenth Annual Meeting of the Berkeley Linguistics Society: General Session and Parasession on The Grammar of Even Structure*, Vol. 17, eds C. Johnson, L. A. Sutton, and R. Shields (Berkeley, CA: University of California, Berkley Linguistics Society), 156–167.

Lewis, R. L., and Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cogn. Sci.* 29, 375–419. doi: 10.1207/s15516709cog0000_25

McDaniel, D., and Cowart, W. (1999). Experimental evidence for a minimalist account of English resumptive pronouns. *Cognition* 70, B15–B24. doi: 10.1016/S0010-0277(99)00006-2

Obenauer, H. (1983). On the Identification of empty categories. *Linguist. Rev.* 4, 153–202.

Obenauer, H. (1994). *Aspects de la Syntaxe A-barre*. Doctoral dissertation, Université de Paris VIII, Paris.

Pesetsky, D. (1987). "Wh-in-situ: movement and unselective binding," in *The Representation of in Definiteness*, eds E. J. Reuland and A. G. B. ter Meulen (Cambridge, MA: MIT Press), 98–129.

Pesetsky, D. M. (2000). *Phrasal Movement and its Kin*. Cambridge, MA: MIT press.

Phillips, C. (2006). The real-time status of island phenomena. *Language* 82, 795–823. doi: 10.1353/lan.2006.0217

Phillips, C. (2013). Some arguments and nonarguments for reductionist accounts of syntactic phenomena. *Lang. Cogn. Process.* 28, 156–187. doi: 10.1080/01690965.2010.530960

Rappaport, G. C. (2003). "The grammatical role of animacy in a formal model of Slavic morphology," in *American Contributions to the Thirteenth International Congress of Slavists (Ljubljana, 2003)*, Vol. 1, eds R. A. Maguire and A. Timberlake (Bloomington, IN: Slavica), 149–166.

R Core Development Team (2015). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rizzi, L. (1990). *Relativized Minimality*. Cambridge, MA: MIT Press.

Rizzi, L. (2004). "Locality and left periphery," in *Structures and Beyond: The Cartography of Syntactic Structures*, Vol. 3, ed. A. Belletti (Oxford: Oxford University Press), 223–251.

Rizzi, L. (2013). Locality. *Lingua.* 130, 169–186. doi: 10.1016/j.lingua.2012.12.002

Ross, J. R. (1967). *Constraints on Variables in Syntax*. Doctoral dissertation, Massachusetts Institute of Technology, Cambridge.

Schütze, C. T. (1996). *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology*. Chicago: University of Chicago Press.

Schütze, C. T., and Sprouse, J. (2013). "Judgment data," in *Research Methods in Linguistics*, eds R. J. Podesva and D. Sharma (New York, NY: Cambridge University Press), 27–50.

Shields, R. (2008). *What's So Special about D-Linking? Poster Session Presented at NELS 39*. New York, NY: Cornell University.

Smolensky, P., and Legendre, G. (2006). *The Harmonic Mind: From Neural Computation to Optimality-Theoretic Grammar. Vol. 1: Cognitive Architecture; vol. 2: Linguistic and Philosophical Implications*. Cambridge, MA: MIT Press.

Sprouse, J. (2011). A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behav. Res. Methods* 43, 155–167. doi: 10.3758/s13428-010-0039-7

Sprouse, J., and Hornstein, N. (2013). *Experimental Syntax and Island Effects*. Cambridge: Cambridge University Press.

Sprouse, J., Wagers, M., and Phillips, C. (2012). A test of the relation between working-memory capacity and syntactic island effects. *Language* 88, 82–123. doi: 10.1353/lan.2012.0029

Starke, M. (2001). *Move Reduces to Merge: A Theory of Locality*. Doctoral dissertation, University of Geneva, Geneva.

Stowe, L. E. (1986). Parsing WH-constructions: evidence for on-line gap location. *Lang. Cogn. Process.* 1, 227–245. doi: 10.1080/01690968608407062

Traxler, M. J., and Pickering, M. J. (1996). Plausibility and the processing of unbounded dependencies: an eye-tracking study. *J. Mem. Lang.* 35, 454–475. doi: 10.1162/jocn.1993.5.2.196

Van Dyke, J. A., and Johns, C. L. (2012). Memory interference as a determinant of language comprehension. *Lang. Linguist. Compass* 6, 193–211. doi: 10.1002/lnc3.330

Van Dyke, J. A., and McElree, B. (2006). Retrieval interference in sentence comprehension. *J. Mem. Lang.* 55, 157–166. doi: 10.1016/j.jml.2006.03.007

Vasishth, S., Suckow, K., Lewis, R. L., and Kern, S. (2010). Short-term forgetting in sentence comprehension: crosslinguistic evidence from verb-final structures. *Lang. Cogn. Process.* 25, 533–567. doi: 10.1080/01690960903310587

Villata, S., Rizzi, L., and Franck, J. (in press). Intervention effects in weak islands and Relativized Minimality: new experimental evidence from graded judgments Lingua.

Warren, T., and Gibson, E. (2002). The influence of referential processing on sentence complexity. *Cognition* 85, 79–112. doi: 10.1016/S0010-0277(02)00087-2

Warren, T., and Gibson, E. (2005). Effects of NP type in reading cleft sentences in English. *Lang. Cogn. Process.* 20, 751–767. doi: 10.1080/01690960500051055

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer, Dario Leander Jim Felix Paape, and handling editor declared their shared affiliation, and the handling editor states that the process nevertheless met the standards of a fair and objective review.

Copyright © 2016 Atkinson, Apple, Rawlins and Omaki. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 05 January 2015
doi: 10.3389/fpsyg.2014.01493

The D-linking effect on extraction from islands and non-islands

Grant Goodall*

Department of Linguistics, University of California, San Diego, La Jolla, CA, USA

Edited by:
Colin Phillips, University of Maryland, USA

Reviewed by:
Philip Hofmeister, University of Essex, UK
Jon Sprouse, University of Connecticut, USA

***Correspondence:**
Grant Goodall, Department of Linguistics, University of California, San Diego, 9500 Gilman Dr., MC 0108, La Jolla, CA 92093-0108, USA
e-mail: ggoodall@ucsd.edu

"D-linked" *wh*-phrases such as *which car* are known to increase the acceptability of sentences with island violations. One influential account of this attributes the effect to working memory: the D-linked filler is easier to retrieve at the site of the gap and this leads to the amelioration in acceptability. Such an account predicts that this effect should occur in general with non-trivial *wh*-dependencies, not just in island environments. An experiment is presented here to test this prediction. *Wh*-questions with both D-linked and bare *wh*-phrases and with both island and non-island embedded clauses are presented to participants, who rate their acceptability on a 7-point scale. Results show that D-linking significantly increases acceptability in both island and non-island environments, in accord with analyses that attribute the effect to working memory. In addition, the increase in acceptability is uniform in both types of environments, suggesting that the island effect itself may not be attributable to working memory.

Keywords: filler-gap dependencies, D-linking, island constraints, working memory, sentence acceptability

INTRODUCTION

The contrast between *wh*-phrases such as *which car* in (1a) and *what* in (1b) has been a major topic of research over the last few decades (e.g., Pesetsky, 1987; Cinque, 1990; Szabolcsi and Zwarts, 1993).

(1) a. **Which car** did you buy?
b. **What** did you buy?

Following terminology introduced in Pesetsky (1987), *wh*-phrases like *which car* are "discourse-linked or "D-linked," in that they naturally prompt an answer chosen from referents already existing in the discourse, whereas *wh*-phrases like *what* do not. (1a), for instance, is typically taken to be asking about a set of cars already known to the speaker and hearer, while (1b), under its most natural reading, is not (see also Katz and Postal, 1964 and Kuroda, 1968).

This distinction has been claimed to have two major consequences for the syntax of *wh*-dependencies. The first has to do with clauses containing two or more *wh*-phrases. English requires that one of these appear at the left edge of the clause, and generally, the syntactically more prominent *wh*-phrase (e.g., the subject vis-à-vis an object) is strongly preferred to play this role, as in (2a) and (2b), even though the less prominent *wh*-phrase is able to when there is no other, as in (2c).

(2) a. I wonder **who** bought **what**.
b. *I wonder **what who** bought.
c. I wonder **what** the man bought.

This is known as the Superiority effect (Chomsky, 1973). D-linking of the *wh*-phrases is claimed to weaken or erase this effect, such that any *wh*-phrase may appear at the left edge of the clause, as in (3) (Karttunen, 1977; Pesetsky, 1987; Comorovski, 1989).

(3) a. I wonder **which man** bought **which car**.
b. I wonder **which car which man** bought.

The second major consequence has to do with the gaps that are obligatorily associated with *wh*-phrase fillers. These gaps are not permitted in certain environments within the clause, a phenomenon known as an island effect (Ross, 1967). (4a) and (4b) show two such island environments, while (4c) shows a non-island environment, in which a gap is permitted.

(4) a. ***What** do you wonder [who bought __] ?
b. ***What** do you believe [the claim that the man bought __] ?
c. **What** do you think [that the man bought __] ?

As with Superiority, island effects are claimed to be weakened or erased when the *wh*-phrase is D-linked (Maling and Zaenen, 1982; Cinque, 1990; Rizzi, 1990; de Swart, 1992; Kiss, 1993; Chung, 1994):

(5) a. **Which car** do you wonder [who bought __] ?
b. **Which car** do you believe [the claim that the man bought __] ?

The above two consequences are surprising, at least initially, in that one might not expect *wh*-dependencies, which are often taken to be a quintessentially syntactic phenomenon, to be so sensitive to discourse-related factors. The effects of D-linking thus present an interesting puzzle, and a number of analyses have been proposed to explain them.

This paper explores this second consequence, the effect of D-linking on islands. We present evidence from a formal acceptability experiment showing that D-linking does indeed improve acceptability of sentences containing island violations, but that they are still significantly degraded compared to sentences without such violations. Moreover, D-linking results in a similar improvement in acceptability even in non-island environments, a finding that has important consequences for determining the sources of the D-linking and island effects.

We review the main classes of proposed explanations for the D-linking effect in islands in Section Three Accounts of D-linking and consider earlier acceptability experiments in this domain in Section Earlier Acceptability Studies. Section Experiment presents and discusses the experiment itself. Section Implications for Formal Acceptability Experiments discusses implications of the experiment for acceptability experiments in general, and general conclusions are presented in Section Conclusion.

THREE ACCOUNTS OF D-LINKING

One influential analysis (Szabolcsi and Zwarts, 1993, 1997; see also Honcoop, 1998) claims that the D-linking effect in islands is primarily due to semantic factors. Certain island domains, under this analysis, contain operators that require a Boolean operation (e.g., intersection), which in turn requires sets made up of discrete individuals. A D-linked *wh*-phrase facilitates an interpretation in which the set questioned consists of individuals, thus allowing for a coherent semantic interpretation of the sentence. With bare *wh*-words like *what*, on the other hand, an interpretation involving a set of individuals is unlikely (though possible under certain circumstances, as Szabolcsi and Zwarts discuss), so the sentence is perceived as ill-formed.

In another set of analyses, the source of the unacceptability of island violations such as (4a) is syntactic. In Rizzi (2001, 2004), for instance, the *wh*-dependency between *what* and its gap site in (4a) violates a putative fundamental property of syntax known as Relativized Minimality, which roughly speaking, disallows dependencies between a filler and a gap when there is an intervening filler [*who*, in the case of (4a)] that could also potentially enter into a dependency of the same type with this gap. Fronted topics are known to be immune to Relativized Minimality effects, so it is important to note in this analysis that D-linked *wh*-phrases bear certain crucial similarities to fronted topics: they contain lexical material beyond the *wh*-word itself, and they are dependent on previously mentioned elements in the discourse. To the extent that D-linked *wh*-phrases may be interpreted as topics, then, they should be able to circumvent the Relativized Minimality requirement and acceptability should increase.

In a third family of analyses, island violations such as (4a) result from limitations in working memory (Kluender and Kutas, 1993; Kluender, 1998; Hofmeister and Sag, 2010; Hofmeister, 2011). The filler *what* must be held in working memory until it can be reintegrated into the structure at the gap site in the embedded clause. Maintaining this filler in working memory while also processing a clause boundary and an intervening filler (*who*) overwhelms the limited capacity of the processor, so filler reintegration is less likely to succeed and the sentence is perceived as unacceptable. The situation changes when the filler is D-linked, because such a filler requires more initial processing, given its more referential nature and the presence of lexical material. The D-linked filler thus has a higher level of initial activation in working memory, and this enables it to survive more successfully until the point where it can be reintegrated at the gap site. There is considerable evidence that such a processing advantage for D-linked fillers exists (e.g., Kluender, 1998; Frazier and Clifton, 2002; Diaconescu and Goodluck, 2004; Hofmeister, 2007a,b, 2011; Hofmeister and Sag, 2010; Hofmeister et al., 2013), and it is reasonable to assume that it could result in higher acceptability [see Hofmeister et al., 2007, for an application of this type of analysis to the D-linking effect on Superiority, as in (2)-(3) above].

This working memory account of islands and D-linking differs from the other two in two important ways. First, it claims that the island and D-linking effects are essentially extragrammatical. That is, the grammar itself has nothing to say about island structures and D-linked fillers, other than that they are allowed, and the effects observed result from capacity constraints on working memory. In the other accounts, on the other hand, these same effects arise because the sentences in question would require an ill-formed semantic or syntactic structure, independently of how such a structure would be processed. Second, all three accounts attribute special properties to D-linked fillers, but only in the working memory account would these special properties be expected to increase acceptability even without an island structure. More concretely, D-linked fillers more readily allow for individuation in the semantic account and for a topic-like interpretation in the syntactic account. These properties permit the filler to avoid island effects, but there is no reason to expect them to affect acceptability in non-island environments. In the working memory account, however, the special property of D-linked fillers is that they have a higher level of activation, and this should facilitate retention in working memory and reintegration at the gap site regardless of the particular structure. Since easier reintegration is assumed to result in higher acceptability, this then predicts that making a filler D-linked will increase acceptability in both island and non-island environments.

We thus arrive at a clear distinction between the working memory analysis and the other two: The working memory account predicts that D-linking will increase acceptability in both islands and non-islands, while the grammatical (semantic and syntactic) accounts do not make this prediction. On the other hand, the three analyses are in agreement that without any auxiliary assumptions, whatever D-linking effect occurs in non-islands should be smaller than that in islands. In the grammatical accounts, this is straightforward: no prediction is made for non-islands, but a very clear effect is predicted for islands. In the working memory account, the predictions result from the way in which island phenomena themselves are accounted for. In these analyses, islands occur because of two main factors: the processing difficulty associated with a filler-gap dependency and that associated with a particularly complex embedded clause [such as the *wh*-clause in (4a) or the complex noun phrase in (4b)]. Crucially, there is an interaction between these two factors, in that the decline in acceptability when both occur together is greater than what would be expected given the decline associated with each

one on its own. Assuming that this interaction is straightforward (e.g., multiplicative), a weakening of one of the factors by amount *x* should result in an overall effect greater than *x*. More specifically, if D-linking lessens the processing difficulty found with filler-gap dependencies, the effect should be amplified when this difficulty is in interaction with the difficulty stemming from a complex embedded clause, and we thus expect D-linking to have a greater effect on acceptability in islands than it does in non-islands.

Two questions may now be posed: (i) Does D-linking increase acceptability in both islands and non-islands, and (ii) is the effect larger in islands than in non-islands? If the answer to the first question is positive, this would lend support to the working memory account of D-linking, and if it is negative, this would argue against it. As we have seen, the grammatical accounts do not make a specific prediction with regard to this question. As for the second question, a positive answer would confirm the predictions made by both the working memory and grammatical accounts. A negative answer would be consistent with the working memory account of D-linking, though inconsistent with the working memory account of islands, given straightforward assumptions about the nature of the interaction taken to underlie island effects. With regard to grammatical accounts, on the other hand, a negative answer would be inconsistent with the accounts of D-linking, though consistent with accounts of islands.

EARLIER ACCEPTABILITY STUDIES

The questions that we are now facing, whether D-linking of fillers increases acceptability even in non-island environments and whether the effect is greater in islands than in non-islands, are in principle able to be addressed experimentally, and some earlier studies have attempted to do so. Hofmeister (2007a) reports the results of a pilot study exploring the effect in non-islands, in which 16 subjects rate 9 sentences using a 7-point scale. The fillers are bare *wh*-words or phrases consisting of either *which* + noun or *which* + *of* + *the* + noun, as in the sample stimuli in (6).

(6) a. Justin proved **what** the engineers lied that they had invented __ without any help or instruction.
b. Justin proved **which devices** the engineers lied that they had invented __ without any help or instruction.
c. Justin proved **which of the devices** the engineers lied that they had invented __ without any help or instruction.

The differences in acceptability among the sentences are marginally significant, with type (6b) more acceptable than (6c), and (6c) more than (6a), but given the small-scale nature of the experiment and the lack of clear results, it is difficult to draw firm conclusions from this. Nonetheless, the study shows that designing an experiment that begins to address these questions is possible in principle.

Alexopoulou and Keller (2013) report on a study consisting of two sub-experiments. In one, the stimuli consist of *wh*-questions with gaps inside embedded *whether*-clauses, a known island environment. In the other, the gap is either in the main clause or in an embedded *that*-clause. In both sub-experiments, there are two factors: gap type (true gap vs. resumptive pronoun) and filler type (*what* vs. *what* + noun vs. *which* + noun vs. *which* + *of* + *the* + noun). Samples of the stimuli with a gap in a *whether*-clause are given in (7a), in the main clause in (7b), and in a *that*-clause in (7c).

(7) a. **What/What movie/Which movie/Which of the movies** does Jean wonder [whether they will watch __ at the cinema]?
b. **What/What movie/Which movie/Which of the movies** will they watch __ at the cinema?
c. **What/What movie/Which movie/Which of the movies** does Mary think [they will watch __ at the cinema]?

The stimuli are arranged in 8 lists using a Latin square design, and subjects respond to the stimuli using magnitude estimation (Bard et al., 1996).

Alexopoulou and Keller find some evidence of a D-linking effect in the *whether*-island case, with *which* + noun (though not *what* + noun or *which* + *of* + *the* + noun) resulting in significantly higher acceptability than bare *what* in cases like (7a). Crucially, however, this effect is not found in either of the non-island environments (see Sprouse et al., for a similar finding, though with D-linked vs. bare as a between-subjects factor). That is, when the gap is in the matrix clause, as in (7b), or in an embedded *that*-clause, as in (7c), there is no significant difference among the four filler types. As discussed above, a result such as this presents straightforward evidence against the working memory account of the D-linking effect, since this account predicts that D-linked fillers will be easier to reintegrate into the structure and that this will lead to increased acceptability, both in island and non-island contexts. The lack of an observed effect in the non-island contexts is entirely consistent with the grammatical accounts and thus provides an argument in their favor.

(7b-c) are standardly considered fully acceptable with any of the fillers, however, so in order to detect a D-linking effect in these cases, the experiment will need to be able to distinguish among sentences at the very high end of the acceptability scale. There is some indication in Alexopoulou and Keller's results that their experiment is not able to do this reliably. Sentences with short dependencies as in (7b), where the filler and the gap are within the same clause, have always been found in previous experimental work to be much more acceptable than those with long dependencies, such as in (7c), where the filler and the gap are in separate clauses, despite the fact that both are standardly treated as grammatical (e.g., Cowart, 1997; Alexopoulou and Keller, 2007). In Alexopoulou and Keller's results, though, the two sentence types are virtually identical, strongly suggesting the presence of a ceiling effect. If this is true for short vs. long filler-gap dependencies, for which the literature reports a very robust difference, then the fact that they find no difference among the four filler types is perhaps not as telling as it appears at first.

A similar lack of expected distinctions in the mid-range of the acceptability scale suggests that the experiment may not have attained a level of sensitivity sufficient to detect all potential contrasts of interest. The *whether*-islands tested are a canonical example of the type of island that is thought to exhibit D-linking effects (see, e.g., Szabolcsi, 2006), yet recall that this was only found with *which* + noun, not *what* + noun or *which* + *of* + *the* + noun, contrary to expectations. The absence of a D-linking effect with *that*-clauses in this experiment is thus perhaps not

surprising, given that this effect was also not detected in some of the cases where it would be most expected.

The possibility that the experiment was not sensitive enough to detect all potential D-linking effects gains further plausibility when one looks at the details of the experimental design, which show several features that could have contributed to a lowered level of sensitivity. In terms of the materials, each participant saw just one token of each condition, and there was a 1:1 filler/experimental ratio. In addition, there was only partial counterbalancing of the stimuli: There were 24 conditions overall, yet only 8 lexicalizations of each condition, and 8 lists of stimuli were created. These lists were distributed among 22 participants, so some lists (and stimuli) were seen by more participants than others. As for the participants themselves, they were self-reported native speakers of English recruited over the Internet. Given the nature of the English-speaking community, where bilingualism in many forms is very common and it is not always clear who counts as a "native speaker," it is possible that the participants' language histories were very heterogeneous, which in turn could have led to increased variability in their responses. In addition, participants took part in the experiment over the Internet. Although indications are that performing sentence acceptability experiments in this way gives adequate results (Gibson et al., 2011; Sprouse, 2011b), there is still the realistic possibility that it will result in increased noise, especially when the number of participants is small. Finally, the response method used with participants (magnitude estimation) may have also contributed to a decrease in sensitivity. This is still a matter of some controversy, but there are suggestions in the literature that magnitude estimation may not be as sensitive as initially thought and that it may even obscure fine-grained distinctions (Sprouse, 2011a; Weskott and Fanselow, 2011; Fukuda et al., 2012).

We of course cannot be sure that any of the above factors resulted in a decrease in the experiment's sensitivity, but given that the D-linking effect is likely very subtle, it would be prudent to avoid design features that might make detecting such an effect more difficult.

Given the existing literature, then, it is still an open question whether D-linking increases the acceptability of *wh*-dependencies in non-island environments and if it does, whether this effect is smaller in non-islands than in islands. In the Hofmeister (2007a) study, the results are not clear enough to draw firm conclusions, and in the Alexopoulou and Keller (2013) study, there are reasons to suspect that the results are compromised by a ceiling effect and a general lack of sensitivity. In the following section, we describe an experiment that is designed to address directly the questions of a possible D-linking effect in non-island environments and how this might compare to that in island environments.

EXPERIMENT

PARTICIPANTS

Fifty six people participated in this experiment. All were undergraduate students at the University of California, San Diego who were participating for course credit. The experiment was performed in a laboratory setting, with prior authorization from the university's Institutional Review Board. All participants gave their informed consent.

The results of two groups of participants were excluded. The first included those who on a language background questionnaire, gave a language other than English as their native language or their dominant language, or who indicated that they had been born outside of the U.S. This eliminated 6 participants. The second group included those who did not appear to be attending to the task, as evidenced by their responses on 9 key filler items that were unquestionably grammatical or unquestionably ungrammatical. Participants who made 2 or more "errors" on these fillers were excluded, where "errors" are defined as a response of 3 or below (on a 1–7 scale) to a grammatical filler or a response of 5 or above to an ungrammatical filler. 2 participants were eliminated in this way, leaving 48 in total (2 per experimental list).

MATERIALS AND METHOD

Experimental items were all *wh*-questions and were prepared using a 2 × 3 design, crossing filler type (bare vs. D-linked) and type of structure in which the gap is located (embedded complex noun phrase vs. *wh*-clause vs. *that*-clause). With regard to filler type, the bare filler was always *what* and the D-linked fillers all had the form *which of the* + plural noun. With regard to structure type, the complex noun phrases all contained a singular head noun (e.g., *claim*, *plan*, *idea*), followed by a clausal complement, and the *wh*-clauses all contained *who* as subject of that clause. The 6 conditions are exemplified in (8).

(8) a. **What / Which of the cars** do you believe the claim that he might buy?
b. **What / Which of the cars** do you wonder who might buy?
c. **What / Which of the cars** do you believe that he might buy?

(8a) and (8b) are classic violations of island constraints: the Complex Noun Phrase Constraint (CNPC) and the *Wh*-island Constraint, respectively (Ross, 1967). The gap in (8c) is within a *that*-clause, a classic non-island environment.

Twenty four sets of lexically matched stimuli were created and distributed into 6 counterbalanced lists using a Latin square design, such that each list contained 4 tokens of each condition. 81 filler items were added to each list, and the lists were then pseudo-randomized twice, resulting in 12 lists. An additional 12 lists were created by reversing the order of items, resulting in a total of 24 lists. 2 participants were randomly assigned to each list; each experimental item was thus seen by 8 participants. The full set of stimuli is presented in the Supplementary Material.

Participants saw the stimuli on a computer screen and were instructed to rate each sentence on a scale from 1 ("very bad") to 7 ("very good") based on how it sounded to them as a native speaker of the language. The scale was presented horizontally in evenly spaced increments with only the two extremes labeled and participants indicated their response by clicking on the appropriate number. They were told to rely on their first reaction, without trying to analyze the sentence, and that there were no "correct" answers. They were also told to rate each sentence on its own, regardless of how simple or complicated the sentence might seem.

RESULTS

The results were transformed to z-scores prior to analysis. The z-score mean and standard error for each of the six conditions is presented in **Figure 1**.

A linear mixed effects model was run with filler type and structure type as fixed factors, participant and item as random intercepts, and by-participant and by-item random slopes for filler type and a by-participant random slope for structure type, using the *lmer* function in the *lme4* package for R (Bates et al., 2014a,b; R Core Team, 2014). All *p*-values were calculated by Satterthwaite approximation, using the *lmerTest* package (Kuznetsova et al., 2014). This revealed a significant main effect for filler type (D-linked: −0.168 vs. bare: −0.444; $t = 3.446$; $p < 0.001$), and this effect remained significant when the model was restricted to each of the three structures individually: CNPC (D-linked: −0.441 vs. bare: −0.705; $t = 3.476$; $p < 0.01$), *wh*-island (D-linked: −0.545 vs. bare: −0.923; $t = 3.982$; $p < 0.001$), and *that*-clause (D-linked: 0.483 vs. bare: 0.295; $t = 2.416$; $p < 0.02$). To test for an interaction between filler type and structure type, a second model was constructed without an interaction between these two fixed factors and the results compared to the first by means of the *anova* function. This revealed no significant difference between the two models ($p = 0.155$) and thus no significant interaction between these two factors. The interaction between filler type and structure type was also not significant when the CNPC data were excluded and the model run as a 2 × 2 design, with *wh*-island and *that*-clause as the levels for structure type ($t = 1.866$; $p = 0.062$) and when the *wh*-island data were excluded and CNPC and *that*-clause used as the levels for structure type ($t = 0.771$; $p = 0.440$).

To a large extent, earlier observations in this domain are confirmed (e.g., general island effects and D-linking effects are readily apparent), but there are two novel findings here. First, the increase in acceptability associated with D-linked fillers occurs in all three structure types, not just in the islands. Second, this increase is uniform across all three types. That is, the amount of increase associated with D-linking does not appear to vary significantly between islands and non-islands.

As noted earlier, achieving sufficient sensitivity is a concern in this type of study, but the fact that all of the island effects and D-linking effects that the existing literature predicts did emerge suggests that the experiment was successful in this regard. It also appears that the experiment avoided a ceiling effect in the case of *wh*-questions with a gap in the *that*-clause. Although significantly more acceptable than the island violations, these sentences are still within the mid-range of the acceptability of the fillers. As seen in **Figure 2**, the acceptability of the fillers went as high as 1.61, much higher than the mean acceptability of the *that*-clause sentences with either a bare or D-linked filler (0.295 and 0.483, respectively).

DISCUSSION

The main purpose of this study is to determine whether D-linking of the filler improves the acceptability of *wh*-questions where the gap is in a non-island, and if so, whether this improvement is of the same size as that which occurs when the gap is within an island. We have now seen that the effect does occur in non-islands and that it is not different in size from that observed in islands. More specifically, D-linking leads to a significant increase in acceptability when the gap is in a non-island *that*-clause, and in addition, significant increases are also found in the two island cases examined. There is no significant interaction between filler type and structure type, suggesting that the amelioration due to D-linking is essentially uniform regardless of whether the gap is within an island or non-island.

These results confirm one crucial prediction of the working memory analysis of D-linking effects. If, as this analysis claims, D-linking effects arise because the nature of D-linking allows for easier reintegration of the filler at the gap site, and if this in turn results in higher acceptability, then we would expect to be able to

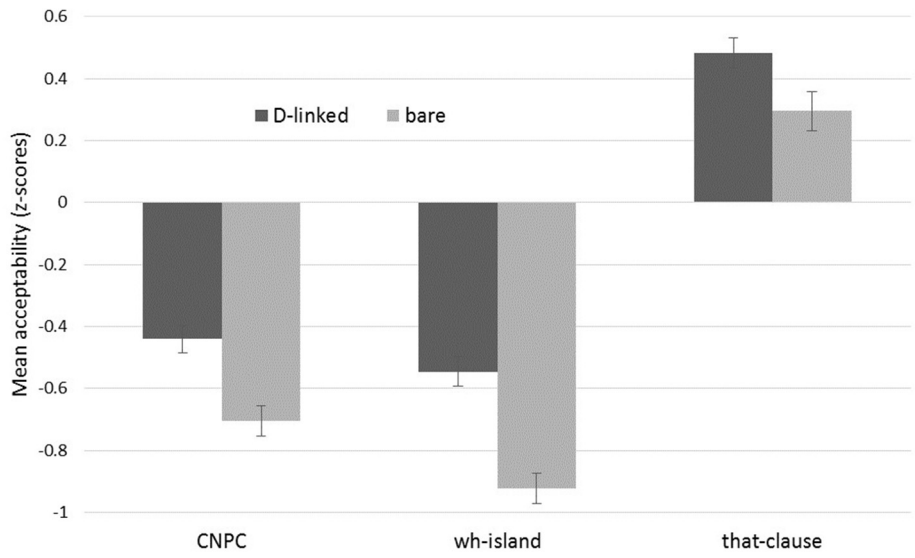


FIGURE 1 | Mean acceptability of experimental conditions (z-scores; error bars indicate SE).

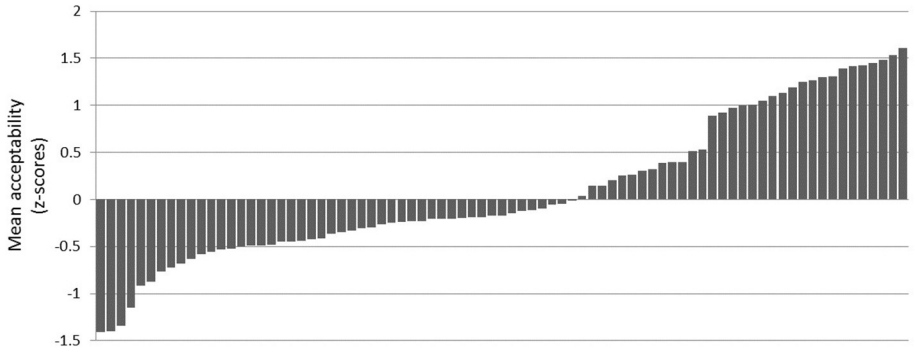


FIGURE 2 | Mean acceptability of fillers (z-scores).

detect this increase in acceptability no matter whether the gap is located in an island or a non-island. The results seen here suggest that this prediction is correct and thus provide new evidence in favor of the working memory analysis. This new evidence from acceptability complements and is in accord with the considerable evidence already existing that D-linking facilitates the processing of filler-gap dependencies.

The results of the experiment are at odds, however, with another prediction that is shared by both the working memory analysis and the grammatical analyses. Namely, the experiment finds an essentially uniform D-linking effect in both islands and non-islands, whereas both types of analyses, in their most straightforward forms, predict a larger effect in the case of islands. For the working memory analysis, this is because island phenomena are the result of an interaction between the difficulty of the dependency and the difficulty of the structure, so if we assume that this interaction is simple (e.g., multiplicative), facilitating the dependency in this case should lead to an increase in the acceptability of the island that is larger than what would be expected by facilitating the dependency alone, as in a non-island structure. For the grammatical analyses, it is because island phenomena are the result of limitations on the operation of the syntax and/or semantics, and D-linking has the effect of removing these limitations. In non-islands, these limitations do not exist, so no effect of D-linking is expected. Both the working memory and the grammatical analyses, then, predict a difference in behavior between islands and non-islands with regard to D-linking, but this difference is not found here.

On the one hand, then, the results of the experiment here provide important support for the idea that the D-linking effect is ultimately due to an effect of working memory. We have found that D-linking increases acceptability in both island and non-island environments, just as would be expected if D-linking facilitates reintegration of the filler at the gap site in filler-gap dependencies. On the other hand, though, the results suggest caution with the idea that the island phenomenon itself is ultimately due to working memory. As we have seen, we would expect a larger D-linking effect in islands than in non-islands if this were true, and this is not what we observe. The results here are most compatible, then, with the view that the D-linking effect is due to working memory and that the island effect is due to some independent mechanism. Crucially, this mechanism and the working memory effect should be such that they do not interact, as would be expected, for example, if the island effect (but not the D-linking effect) were the result of a grammatical constraint. Given the types of grammatical constraints that have been proposed for islands (e.g., Rizzi, 2004; Boeckx, 2008; Truswell, 2011), one would expect them to combine additively with working memory effects, without any interaction, and the results here thus provide some support for such an account of islands and D-linking. Clearly, though, any conclusion that islands themselves are independent of working memory effects must be approached with caution, given the evidence that has been put forward suggesting that the two are closely related (for recent discussion of the evidence for and against this idea, see Hofmeister et al., 2012a,b; Sprouse et al., 2012a,b; and Michel, 2014).

Further support for the idea that the D-linking effect itself is due to the effects of working memory comes from an experimental result not yet highlighted: both CNPC and *wh*-islands show a significant amelioration with D-linking. This finding is of interest because much of the literature on D-linking assumes that it affects only *weak* islands (i.e., those in which acceptability of an argument gap is much higher than that of an adjunct gap) and not *strong* islands (i.e., those in which argument gaps and adjunct gaps are equally unacceptable) (e.g., Cinque, 1990). *Wh*-islands are a standard example of a weak island and CNPC is typically taken to be a strong island (e.g., Szabolcsi, 2006), so the fact that both show a clear D-linking effect in the results here runs counter to common assumptions. It is exactly what the working memory analysis of D-linking predicts, however, so this finding represents additional support for it.

Another area where the experimental results here run counter to common assumptions in the literature concerns the relation between D-linking and islands. It is often stated that D-linking makes gaps within islands licit (e.g., Szabolcsi, 2006). The results here point to a more nuanced view, however. Although a D-linked filler does significantly increase the acceptability of a gap within an island, this increased acceptability is still relatively low: the mean z-scores are well below 0 (-0.441 for CNPC and -0.545 for *wh*-islands) and below most of the filler items (see also Goodall, 2004, 2010; and Sprouse et al., in press). The contribution of D-linking to acceptability seen here may thus be more modest

than what is sometimes suggested, but this fact is compatible with both processing and grammatical analyses of D-linking. In the processing analyses, the idea that D-linking leads to easier reintegration of the filler at the gap site does not mean that no difficulty remains, and this residual difficulty would reasonably be expected to lead to low acceptability. In the grammatical analyses, similarly, D-linking may make it easier to construe the filler as being individuated or as referring to material in the previous discourse, but it is very conceivable that such accommodation would come with a processing cost that would suppress acceptability. The fact that the increase in acceptability due to D-linking is relatively small is thus important to note, but it does not in itself necessarily differentiate among various analyses of the D-linking effect.

The experiment here was designed to test for D-linking effects across a range of syntactic environments. As is always the case, one must be cautious about generalizing the results beyond those structures tested. The experimental design included reasonable representative samples of a non-island structure (*that*-clause), for instance, and of island structures (CNPC and *wh*-islands), but these of course do not exhaust the possibilities (see Sprouse et al., in press, for an investigation of subject and adjunct islands, in addition to those explored here). Similarly, the type of D-linked filler used (*which of the* N) is a prototypical one, but there are other possibilities (*which* N or *what* N) that could also be tested. In addition, the stimuli in this experiment were presented without context (although by their very nature, D-linked fillers provide a kind of context that bare fillers do not), but D-linking is known to be sensitive to context, to such an extent that even bare *wh*-words can behave as D-linked *wh*-phrases if the context is strong enough (e.g., Cinque, 1990; Szabolcsi and Zwarts, 1993; Rizzi, 2004). There is no particular reason to expect that manipulating either the island/non-island structure or the properties of the filler would alter the results presented here, but prudence dictates caution in extending too far beyond what this study provides evidence for.

IMPLICATIONS FOR FORMAL ACCEPTABILITY EXPERIMENTS

The use of formal experiments to measure acceptability is relatively recent, primarily coming after the publication of Schutze (1996) and Cowart (1997), and has only become common in the last few years (see, e.g., Myers, 2009 and Sprouse and Hornstein, 2013 for overviews). As a consequence, there are still certain methodological concerns and questions for which there does not yet exist a full consensus, and some of these relate to aspects of the present study.

One of these concerns the proper way to interpret participant responses to stimuli on the numerical scale. In this study, as in many others, participants were asked to indicate their responses using a 7-point scale, where 1 was labeled "very bad" and 7 "very good." This method is known to yield results that are reasonably valid, reliable and sensitive (Myers, 2009; Weskott and Fanselow, 2011; Fukuda et al., 2012), but there remain concerns that participants may use different areas of the scale in different manners. In particular, Poulton (1979, 1989) demonstrates equalizing biases in rating tasks in which participants spread out responses over the full range of the scale and tend to use each response category equally often. In acceptability studies, this means that if there were a large number of low-acceptability stimuli and many fewer high-acceptability stimuli, for example, the differences among the lower ones could be exaggerated (i.e., participants would spread their responses out over a larger portion of the scale) while differences among the higher ones could be suppressed (i.e., participants would compress their responses into whatever portion of the scale was not being used for the lower stimuli). A possibility like this is a special concern in the present study for two reasons. First, the essential question being asked is whether a small difference in the lower end of the scale (i.e., the D-linking effect in island environments) is also found in the higher end of the scale (i.e., in non-island environments). Since this latter difference was indeed found, one could legitimately worry that this finding results simply from a tiny difference being exaggerated because of an equalizing bias. Second, there is some initial indication that the results are consistent with an equalizing bias, in that many of the response categories were used at similar rates, as seen in **Figure 3** (especially categories 2, 3, 5, and 7), and furthermore, the number of responses at the lower end (categories 1–3) and at the higher end (categories 5–7) of the scale were almost identical: 2183 and 2151, respectively.

There is thus a real concern that the results are influenced by an equalizing bias on the part of participants. However, closer inspection of participant responses reveals that despite the overall distribution in **Figure 3**, most individual participants used the seven response categories at very uneven rates, as seen in **Figure 4**, suggesting that there was no clear equalizing bias for most participants.

Moreover, Cowart (1997), notes that rating experiments can be designed so as to discourage the possibility of equalizing bias. For example, the stimuli (including filler items) can be created so that no particular area of the scale is likely to predominate, thus decreasing the possibility of distortion in one area of the scale. In addition, the response scale can be presented to subjects in such a way that clearly invites an interpretation of the numbers as representing equal intervals. Both of these measures were taken in the present study. The stimuli included many filler items that were unquestionably of very high acceptability, as in (9), and of very low acceptability, as in (10), as well as many of intermediate status, as in (11).

(9) What do you think was on the table yesterday? (raw mean = 6.67).

Are all of the children in the room? (raw mean = 6.88).

(10) What would the girl could the tiger suddenly do? (raw mean = 1.54).

Would the this store is successful? (raw mean = 1.54).

(11) What does everybody say that Marge saw the books? (raw mean = 2.69).

Who were sculptures of on exhibit in the gallery? (raw mean = 3.58).

About which bike will several ads be shown to the athletes? (raw mean = 4.61).

Second, the response categories were presented after each stimulus in left-to-right increasing order in evenly spaced increments, in the manner of a ruler, with each numeral underneath its

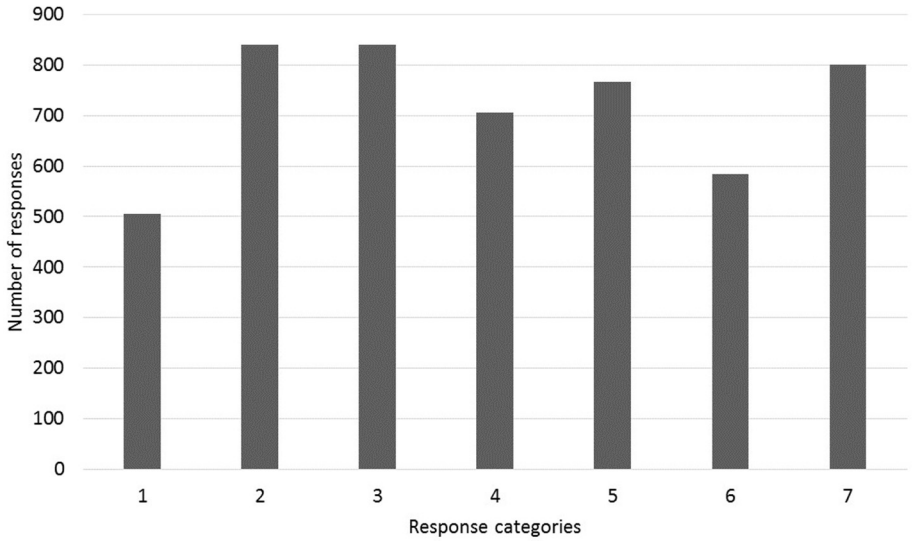


FIGURE 3 | Distribution of the 5040 participant responses per response category.

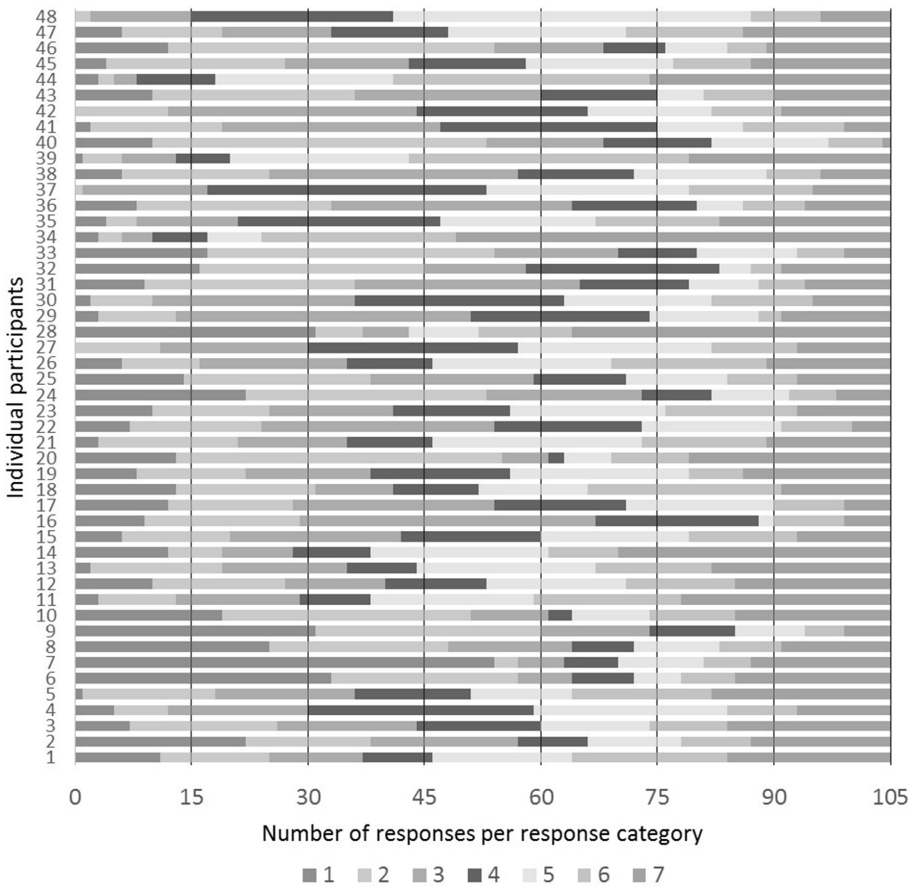


FIGURE 4 | Distribution of responses by individual participant. An equal distribution of responses would consist of 15 responses per category per participant.

corresponding response button. Neither of these steps can eliminate the possibility of response biases, but together, they make it more likely that the D-linking amelioration that we observed with non-islands at the higher end of the scale is in fact similar and comparable to the amelioration seen with islands at the lower end of the scale.

Another area of concern in the recent literature on formal acceptability experiments has been cases where the experimental results and those obtained through more traditional means (i.e., by asking a small number of speakers (perhaps including the investigator) for judgments on a representative set of sentences) seem to diverge (Sprouse and Almeida, 2012, 2013; Gibson and Fedorenko, 2013; Gibson et al., 2013; Sprouse et al., 2013). The present experiment is of interest in this regard, because some of the results align with the traditional literature and others do not. For instance, the D-linking effect that was observed here with *wh*-islands lines up well with what has been reported in more traditional studies, but the similar effect seen with *that*-clauses does not. This then leads to a clear question: If there really is a D-linking effect with gaps in *that*-clauses, why has this never been observed in studies using more traditional methodology? Two possible answers arise. First, it may be simply that no one found this effect because no one was looking for it. From the standpoint of a researcher exploring properties of the grammar, gaps within *that*-clauses are highly acceptable and thus presumably grammatical (i.e., allowed by the grammar). Finding that these gaps become even more acceptable when the filler is D-linked would not be informative, because in standard models, there is no way for the sentence to become even more grammatical. Put simply, standard grammatical models can capture gradations of ungrammaticality (e.g., by counting the number of violations or their severity), but not gradations of grammaticality. From this standpoint, then, there would be no particular reason to look for D-linking effects in otherwise grammatical sentences.

A second answer might be that formal acceptability experiments appear to be very sensitive to strains on working memory in a way that more traditional methods are not, especially for sentences in the higher range of acceptability. For example, filler-gap dependencies within a single clause and those spanning two clauses are, other things being equal, taken to be equally acceptable in traditional studies, but formal acceptability experiments typically find a sharp decline in acceptability for the latter (Kluender and Kutas, 1993; Cowart, 1997; Alexopoulou and Keller, 2007). It is not clear why this divergence between the two methods occurs, but given that it does, the fact that the present study found a distinction that traditional studies have not begins to make sense. If the D-linking effect truly is a working memory effect, then we might not expect traditional methods to be sensitive to it in the case of *that*-clauses, which are of relatively high acceptability.

There thus appear to be reasonable ways in which one might explain the discrepancy between traditional methods and the experiment presented here with regard to the effect of D-linking in non-island environments. In this case or more generally, it is not a question of which of these methods is right or wrong, but of which is appropriate given the resources available and the nature of the phenomenon being investigated. Since the focus of investigation here concerns the possibility of small differences in acceptability among sentences that are taken to be grammatical, where working memory effects might crucially be involved, a formal experiment seems appropriate.

Finally, the present experiment highlights the fact that there is as much need for careful design and attention to detail in sentence acceptability experiments as in any other experimental methodology. Many of the acceptability contrasts that interest researchers are very robust and are easily detectable across a wide range of methodologies: traditional fieldwork, traditional introspection, very simple experiments, etc. For more subtle contrasts, however, the method may need to be chosen more carefully. In this study, several steps were taken in order to ensure adequate sensitivity and to avoid a ceiling effect, a particular danger in this case since the crucial sentences of interest were of relatively high acceptability. For example, participants were screened for language background and attention to task, and they performed the experiment in a laboratory setting. The materials were also fully counterbalanced: experimental stimuli were distributed across lists following a standard Latin square design, and each experimental item was seen by exactly the same number of participants. Filler items represented a wide range of acceptability, including many of very high acceptability. In addition, there was a relatively large number (192) of observations per condition (4 tokens of each condition per participant; 48 participants), and the response method used by participants (7-point scale) is one that has been shown capable of capturing small differences in acceptability (Weskott and Fanselow, 2011; Fukuda et al., 2012). These various aspects of the experimental design were chosen deliberately in response to the particular needs presented by this study.

CONCLUSION

It has been known for many years that D-linking, where the filler in a *wh*-question prompts an answer chosen from referents already existing in the discourse, increases the acceptability of sentences where the gap is inside an island configuration. It has been claimed in a number of analyses that this phenomenon reflects the way that working memory operates in sentence processing, in that at the point of the gap site, D-linked fillers are easier to access and then integrate into the existing structure, and that this ease of processing results in higher acceptability. These analyses clearly predict that this D-linking effect should be found not just with islands, but with filler-gap dependencies in non-islands as well. The experiment presented here tested this prediction directly by probing for D-linking effects on acceptability in two island and one non-island environments. It was seen that the effect occurs in all three cases, confirming the prediction made by the analyses that attribute the effect to the operation of working memory.

In addition, the effect is essentially uniform across all three cases, contrary to what many analyses of the islands themselves would predict. The combined results are most compatible with a view in which the D-linking effect is due to working memory and the island effects are due to something independent of this, such as grammar. The results here suggest that these two effects may combine additively, but do not interact.

ACKNOWLEDGMENTS

Aspects of this work have been presented at the CUNY Conference 2013, the Linguistic Society of America Annual Meeting 2014, the Workshop on Understanding Acceptability Judgments at the University of Potsdam, and the University of Chicago. I am grateful to the audiences there for their valuable feedback and to Adrienne LeFevre and Michelle McCadden for their assistance in carrying out this research.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fpsyg.2014.01493/abstract

REFERENCES

Alexopoulou, T., and Keller, F. (2007). Locality, cyclicity, and resumption: at the interface between the grammar and the human sentence processor. *Language* 83, 110–160. doi: 10.1353/lan.2007.0001

Alexopoulou, T., and Keller, F. (2013). "What vs. who and which: kind-denoting fillers and the complexity of whether-islands," in *Experimental Syntax and Island Effects*, eds J. Sprouse and N. Hornstein (Baltimore, MD: Cambridge University Press), 310–340.

Bard, E. G., Robertson, D., and Sorace, A. (1996). Magnitude estimation of linguistic acceptability. *Language* 72, 32–68. doi: 10.2307/416793

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2014a). *_lme4: Linear Mixed-Effects Models using Eigen and S4_. R package version 1.1-7*. Available online at: http://CRAN.R-project.org/package=lme4

Bates, D., Maechler, M., Bolker, B. M., and Walker, S. (2014b). "lme4: Linear mixed-effects models using Eigen and S4," in *ArXiv e-print; Submitted to _Journal of Statistical Software_*. arXiv:1406.5823.

Boeckx, C. (2008). *Bare Syntax*. Oxford; New York: Oxford University Press.

Chomsky, N. (1973). "Conditions on transformations," in *Festschrift for Morris Halle*, eds S. Anderson and P. Kiparsky (New York, NY: Holt, Reinhart & Winston), 232–286.

Chung, S. (1994). Wh-agreement and "referentiality" in Chamorro. *Linguistic Inquiry* 25, 1–44.

Cinque, G. (1990). *Types of A'-Dependencies*. Cambridge, MA: MIT Press.

Comorovski, I. (1989). *Discourse and the Syntax of Multiple Constituent Questions*. Doctoral dissertation, Cornell University.

Cowart, W. (1997). *Experimental Syntax: Applying Objective Methods to Sentence Judgments*. Thousand Oaks, CA: Sage Publications.

de Swart, H. (1992). "Intervention effects, monotonicity and scope," in *SALT II: Proceedings of the Second Conference on Semantics and Linguistic Theory*, eds C. Baker and D. Dowty (Columbus: The Ohio State University), 387–406.

Diaconescu, R. C., and Goodluck, H. (2004). The pronoun attraction effect for D(iscourse)-linked phrases: evidence from speakers of a null subject language. *J. Psycholinguist. Res.* 33, 303–319. doi: 10.1023/B:JOPR.0000035103.19149.dc

Frazier, L., and Clifton, C. Jr. (2002). Processing 'd-linked' phrases. *J. Psycholinguist. Res.* 31, 633–660. doi: 10.1023/A:1021269122049

Fukuda, S., Goodall, G., Michel, D., and Beecher, H. (2012). "Is magnitude estimation worth the trouble?," in *Proceedings of the 29th West Coast Conference on Formal Linguistics* (Somerville, MA: Cascadilla Proceedings Project), 328–336.

Gibson, E., and Fedorenko, E. (2013). The need for quantitative methods in syntax and semantics research. *Lang. Cogn. Process.* 28, 88–124. doi: 10.1080/01690965.2010.515080

Gibson, E., Piantadosi, S., and Fedorenko, K. (2011). Using mechanical turk to obtain and analyze English acceptability judgments. *Lang. Linguist. Compass* 5, 509–524. doi: 10.1111/j.1749-818X.2011.00295.x

Gibson, E., Piantadosi, S. T., and Fedorenko, E. (2013). Quantitative methods in syntax/semantics research: a response to Sprouse and Almeida (2013). *Lang. Cogn. Process.* 28, 229–240. doi: 10.1080/01690965.2012.704385

Goodall, G. (2004). "On the syntax and processing of wh-questions in Spanish," in *Proceedings of the 23rd West Coast Conference on Formal Linguistics* (Somerville, MA: Cascadilla Press), 101–114.

Goodall, G. (2010). "Experimenting with wh-movement in Spanish," in *Romance Linguistics 2008: Interactions in Romance. Selected Papers from the 38th Linguistic Symposium on Romance Languages (LSRL), Urbana-Champaign, April 2008* (Amsterdam; Philadelphia: John Benjamins Publishing), 233–248.

Hofmeister, P. (2007a). *Representational Complexity and Memory Retrieval in Language Comprehension*. Stanford, CA: Stanford University doctoral dissertation.

Hofmeister, P. (2007b). "Retrievability and gradience in filler-gap dependencies," in *Proceedings of 43rd Regional Meeting of the Chicago Linguistics Society,* Vol. 43, eds M. Elliott, J. Kirby, O. Sawada, E. Staraki, and S. Yoon (Chicago, IL).

Hofmeister, P. (2011). Representational complexity and memory retrieval in language comprehension. *Lang. Cogn. Process.* 26, 376–405. doi: 10.1080/01690965.2010.492642

Hofmeister, P., Arnon, I., Jaeger, T. F., Sag, I. A., and Snider, N. (2013). The source ambiguity problem: distinguishing the effects of grammar and processing on acceptability judgments. *Lang. Cogn. Process.* 28, 48–87. doi: 10.1080/01690965.2011.572401

Hofmeister, P., Casasanto, L. S., and Sag, I. A. (2012a). How do individual cognitive differences relate to acceptability judgments?: A reply to Sprouse, Wagers, and Phillips. *Language* 88, 390–400. doi: 10.1353/lan.2012.0025

Hofmeister, P., Casasanto, L. S., and Sag, I. A. (2012b). Misapplying working-memory tests: a reductio ad absurdum. *Language* 88, 408–409. doi: 10.1353/lan.2012.0033

Hofmeister, P., Jaeger, T. F., Sag, I. A., Arnon, I., and Snider, N. (2007). "Locality and accessibility in wh-questions," in *Roots: Linguistics in Search of its Evidential Base*, eds S. Featherston and W. Sternefeld (Berlin: Mouton de Gruyter), 185–206.

Hofmeister, P., and Sag, I. A. (2010). Cognitive constraints and island effects. *Language* 86, 366–415. doi: 10.1353/lan.0.0223

Honcoop, M. (1998). *Dynamic Excursions on Weak Islands*. Doctoral dissertation, Holland Academic Graphics, Leiden University, The Hague.

Karttunen, L. (1977). Syntax and semantics of questions. *Linguist. Philos.* 1, 3–44. doi: 10.1007/BF00351935

Katz, J. J., and Postal, P. M. (1964). *An integrated Theory of Linguistic Descriptions*. Cambridge, MA: MIT Press.

Kiss, K. É. (1993). Wh-movement and specificity. *Nat. Lang. Linguist. Theory* 11, 85–120. doi: 10.1007/BF00993022

Kluender, R. (1998). On the distinction between strong and weak islands: a processing perspective. *Syntax Semant.* 29, 241–280.

Kluender, R., and Kutas, M. (1993). Subjacency as a processing phenomenon. *Lang. Cogn. Process.* 8, 573–633. doi: 10.1080/01690969308407588

Kuroda, S. Y. (1968). English relativization and certain related problems. *Language* 44, 244–266. doi: 10.2307/411621

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (2014). *lmerTest: Tests for Random and Fixed Effects for Linear Mixed Effect Models (lmer objects of lme4 package). R package version 2.0-11*. Available online at: http://CRAN.R-project.org/package=lmerTest

Maling, J., and Zaenen, A. (1982). "A phrase structure account of Scandinavian extraction phenomena," in *The Nature of Syntactic Representation*, eds P. Jacobson and G. Pullum (Dordrecht: Reidel), 229–282.

Michel, D. (2014). *Individual Cognitive Measures and Working Memory Accounts of Syntactic Island Phenomena*. Doctoral dissertation, University of California, San Diego, CA.

Myers, J. (2009). Syntactic judgment experiments. *Lang. Linguist. Compass* 3, 406–423. doi: 10.1111/j.1749-818X.2008.00113.x

Pesetsky, D. (1987). "Wh-*in-Situ*: movement and unselective binding," *The Representation of (in)Definiteness*, eds E. Reuland and A. ter Meulen (Cambridge, MA: MIT Press), 98–129.

Poulton, E. C. (1979). Models for biases in judging sensory magnitude. *Psychol. Bull.* 86:777. doi: 10.1037/0033-2909.86.4.777

Poulton, E. C. (1989). *Bias in Quantifying Judgements*. Hove; Hillsdale: Taylor & Francis.

R Core Team. (2014). *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available online at: http://www.R-project.org/

Rizzi, L. (1990). *Relativized Minimality*. Cambridge, MA: MIT Press.

Rizzi, L. (2001). "Relativized minimality effects," in *The Handbook of Contemporary Syntactic Theory*, eds M. Baltin and C. Collins (Oxford: Blackwell), 89–110.

Rizzi, L. (2004). "Locality and left periphery," in *Structures and Beyond. The Cartography of Syntactic Structures*, Vol. 3 (New York, NY: Oxford University Press), 223–251.

Ross, J. R. (1967). *Constraints on Variables in Syntax*. Doctoral dissertation, Massachusetts Institute of Technology. Norwood, NJ: Ablex. Published as Infinite syntax (1986).

Schutze, C. T. (1996). *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology*. Chicago, IL: University of Chicago Press.

Sprouse, J. (2011a). A test of the cognitive assumptions of magnitude estimation: commutativity does not hold for acceptability judgments. *Language* 87, 274–288. doi: 10.1353/lan.2011.0028

Sprouse, J. (2011b). A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behav. Res. Methods* 43, 155–167. doi: 10.3758/s13428-010-0039-7

Sprouse, J., and Almeida, D. (2012). Assessing the reliability of textbook data in syntax: Adger's Core Syntax. *J. Linguist.* 48, 609–652. doi: 10.1017/S0022226712000011

Sprouse, J., and Almeida, D. (2013). The empirical status of data in syntax: a reply to Gibson and Fedorenko. *Lang. Cogn. Process.* 28, 222–228. doi: 10.1080/01690965.2012.703782

Sprouse, J., Caponigro, I., Greco, C., and Cecchetto, C. (in press). *Experimental Syntax and the Variation of Island Effects in English and Italian*. Dordrecht; Boston: Natural Language and Linguistic Theory.

Sprouse, J., and Hornstein, N. (eds.). (2013). *Experimental Syntax and Island Effects*. Cambridge: Cambridge University Press.

Sprouse, J., Schütze, C. T., and Almeida, D. (2013). A comparison of informal and formal acceptability judgments using a random sample from Linguistic Inquiry 2001–2010. *Lingua* 134, 219–248. doi: 10.1016/j.lingua.2013.07.002

Sprouse, J., Wagers, M., and Phillips, C. (2012a). A test of the relation between working-memory capacity and syntactic island effects. *Language* 88, 82–123. doi: 10.1353/lan.2012.0004

Sprouse, J., Wagers, M., and Phillips, C. (2012b). Working-memory capacity and island effects: a reminder of the issues and the facts. *Language* 88, 401–407. doi: 10.1353/lan.2012.0029

Szabolcsi, A. (2006). "Strong and weak islands," in *The Blackwell Companion to Syntax*, Vol. 4, eds M. Everaert and H. van Riemsdijk (Malden, MA: Blackwell Publishing), 479–532.

Szabolcsi, A., and Zwarts, F. (1993). Weak islands and an algebraic semantics of scope taking. *Nat. Lang. Semant.* 1, 235–284. doi: 10.1007/BF00263545

Szabolcsi, A., and Zwarts, F. (1997). "Weak islands and an algebraic semantics of scope taking," in *Ways of Scope Taking*, ed A. Szabolcsi (Dordrecht: Kluwer), 217–262.

Truswell, R. (2011). *Events, Phrases, and Questions*. Oxford: Oxford University Press.

Weskott, T., and Fanselow, G. (2011). On the informativity of different measures of linguistic acceptability. *Language* 87, 249–273. doi: 10.1353/lan.2011.0041

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 28 August 2014; accepted: 04 December 2014; published online: 05 January 2015.

*Citation: Goodall G (2015) The D-linking effect on extraction from islands and non-islands. Front. Psychol. **5**:1493. doi: 10.3389/fpsyg.2014.01493*

This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology.

Copyright © 2015 Goodall. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 12 December 2014
doi: 10.3389/fpsyg.2014.01237

Distinctiveness and encoding effects in online sentence comprehension

Philip Hofmeister[1]* and Shravan Vasishth[2,3]

[1] Department of Language and Linguistics, University of Essex, Colchester, UK
[2] Department of Linguistics, University of Potsdam, Potsdam, Germany
[3] School of Mathematics and Statistics, University of Sheffield, Sheffield, UK

Edited by:
Claudia Felser, University of Potsdam, Germany

Reviewed by:
Judith Koehne, Saarland University, Germany
David Gallo, University of Chicago, USA

***Correspondence:**
Philip Hofmeister, Department of Language and Linguistics, University of Essex, Wivenhoe Park, Essex CO4 3SQ, Colchester, UK
e-mail: phofme@essex.ac.uk

In explicit memory recall and recognition tasks, elaboration and contextual isolation both facilitate memory performance. Here, we investigate these effects in the context of sentence processing: targets for retrieval during online sentence processing of English object relative clause constructions differ in the amount of elaboration associated with the target noun phrase, or the homogeneity of superficial features (text color). Experiment 1 shows that greater elaboration for targets during the encoding phase reduces reading times at retrieval sites, but elaboration of non-targets has considerably weaker effects. Experiment 2 illustrates that processing isolated superficial features of target noun phrases—here, a green word in a sentence with words colored white—does not lead to enhanced memory performance, despite triggering longer encoding times. These results are interpreted in the light of the memory models of Nairne, 1990, 2001, 2006, which state that encoding remnants contribute to the set of retrieval cues that provide the basis for similarity-based interference effects.

Keywords: encoding, retrieval, similarity, distinctiveness, sentence processing

1. INTRODUCTION

In everyday life and in laboratory experiments, people remember the unusual better than the usual. Von Restorff's classic findings illustrate this in terms of superior memory for isolated items, such as a bright green word in the context of a list of words colored black (von Restorff, 1933). More generally, a background of homogeneous stimuli favors the recall and recognition of contextually isolated stimuli. These so-called isolation effects share certain key characteristics with another set of memory effects tied to meaning-related processing. The latter include findings that people recall random trivia facts better if they subsequently hear causally-related information (Bradshaw and Anderson, 1982). Word recall and recognition benefits, too, from meaning-related processing (e.g., assessing the pleasantness of word meanings) compared with the processing of superficial features (e.g., identifying whether the word contains the letter "e"), at least under conditions where the memory retrieval phase taps word meaning (Craik and Lockhart, 1972; Hyde and Jenkins, 1973; Craik and Tulving, 1975; Stein et al., 1978).

Although clearly different in some respects (meaning-related processing is not typically taken to be 'unusual' or 'bizarre'), these two sets of effects can be thought of as being parallel in light of their relationship to both encoding and retrieval. In particular, elaboration and isolation each tend to give rise to longer encoding or study times. Elaboration, like isolation, also raises the probability of contextually unique features that serve to differentiate study items at retrieval, because elaboration typically yields highly diagnostic, meaning-related units of information. Thus, these two memory phenomena both potentially reflect a common set of core principles on the encoding-retrieval relationship and the dynamics of retrieval interference.

Correspondingly, mechanistic explanations for both kinds of effects have hinged on processes operative at either the encoding or the retrieval stage. From one view, the mnemonic benefits may arise from increased processing or attention during the encoding phase (Hirshman et al., 1989; Watkins et al., 2000; Shiffrin, 2003), leading to higher fidelity representations, more highly activated representations, or simply a richer set of self-generated features that form a partly redundant network with the core memory representation. This implies a type of investment-reward strategy; by paying for the cognitive costs of "enhanced" representational encoding, the costs of memory retrieval are lessened.

From a different but not mutually exclusive perspective, semantic processing increases the *distinctiveness* of the stimuli at the time of retrieval: "additional conceptual or semantic features help to differentiate the studied words from each other, making these memories less susceptible to interference and/or providing more features that can be cued on a typical recall or recognition memory test" (Gallo et al., 2008, p. 1096; see also Moscovitch and Craik, 1976; Fisher and Craik, 1977; Jacoby and Craik, 1979; Hunt and Worthen, 2006). In other words, semantic processing of words trumps superficial processing because processing a word's meaning generates more contextually unique features than focusing on its sound or orthographic features. For instance, many words in memory may have the sound [aʊ] or the letter sequence "ch." But relatively few items in memory may be associated with features like "sandy" and "next to the ocean." Consequently, such accounts predict more than a simple contrast

between meaning-related and non-meaning related processing. If semantic processing increases the chances of conceptual distinctiveness, then as semantic processing increases, the chances for successful retrieval from memory should improve up to some arbitrary limit. One implication for at least some such distinctiveness accounts is that a memory target will contrast more with other stimuli, and hence be remembered better, if those *competing* representations elicit more semantic processing. That is, differentiation of two study items may in principle be modulated by the presence/absence of unique semantic features of *either* item, as adding contextually unique features to a competitor cuts down on potential overlap between a competitor and memory target.

Much of this prior research deals with explicit memory for language stimuli, particularly word lists. How linguistic representations are recovered in their most natural setting—online sentence processing—as a function of either elaboration or isolation has not played a significant part in this line of research. This is no doubt due to the implicit nature of memory retrieval during comprehension. Yet comprehending sentences perpetually requires reaccessing some previously perceived information, such as when a pronoun must be interpreted or when the subject of a verb needs to be remembered, and this prior content may vary considerably in the requisite amount of syntactic and semantic processing. Another context in which retrieval from memory happens is in so-called long-distance dependencies (a.k.a. filler-gap dependencies), as in 1:

(1) I finally gave up reading the novel that James Joyce wrote ___ in the 1930s.

To understand this sentence, "the novel" must be retrieved at the embedded verb "wrote" to be properly interpreted as the thematic patient. Evidence that memory retrieval of the argument takes place at the verb comes from reading time data, cross-modal priming tasks, neurophysiological studies, and speed-accuracy tradeoff data (Tanenhaus et al., 1985; Nicol and Swinney, 1989; Kluender and Kutas, 1993; Osterhout and Swinney, 1993; McElree, 2000).

The purpose of the present investigation is to identify whether elaboration and isolation effects occur in online sentence processing and the extent to which such effects might be explained by relating encoding times to retrieval times. The working hypothesis, therefore, is that factors that predict the *success of explict recall* also contribute to the *efficiency of implicit retrieval*. While extant sentence processing models generally ignore variation in the encoding stage as a potential source of processing variation at retrieval sites, cue-based models of retrieval do predict that unique features in a memory target can facilitate retrieval (McElree, 2000; Van Dyke and Lewis, 2003; Lewis and Vasishth, 2005; Lewis et al., 2006; Van Dyke and McElree, 2006, 2011). However, such theories do not make across-the-board predictions that targets with more semantic features, or contextually unique features, ought to be easier to retrieve. This is due to the fact that only those features cued by the retrieval trigger bear on assessments of similarity. For instance, in 2 below, "was complaining" initiates a retrieval probe targeting the animate subject NP "the resident":

(2) a. The worker was surprised that the resident who was living near the dangerous warehouse *was complaining* about the investigation. [= LOW INTERFERENCE]
b. The worker was surprised that the resident who said that the neighbor was dangerous *was complaining* about the investigation. [= HIGH INTERFERENCE]

In the high interference condition, the head and dependent are separated by an NP ("the neighbor") which is a type of semantic object that "can complain" and is also subject marked, similar to the retrieval target. The intervening NP in the low interference condition, in contrast, is inanimate and the object of a preposition, thus mismatching the target semantically and syntactically. Using such materials, Van Dyke (2007) observed evidence of a processing disruption in the high interference condition beginning at the key verbal cluster, which she interpreted in terms of the mechanics of cue-based retrieval. On such an account, features not in the retrieval probe triggered by the verb should have little bearing on memory interference. Whether a target is the only word to begin with an "r" or appears in an unusual font should be immaterial to retrieval efficacy, for example, if verbs do not trigger retrieval probes containing such cues.

In the present experiments, key targets for implicit retrieval in long-distance dependencies differ in the amount of elaboration or "complexity" associated with them (Experiment 1), or with respect to the homogeneity of their text color with the surrounding text (Experiment 2). In both cases, the key features—prenominal modifiers and text color—are unlikely to be directly cued by the retrieval triggering verbs, i.e., verbs don't normally select arguments on the basis of color or the number of modifiers. If elaboration and isolation effects pattern in implicit memory retrieval tasks as they do in explicit memory tasks, then we should expect to see retrieval-related benefits in sentence processing given elaboration or isolation.

Recent reading time data provide some initial evidence that memory retrieval in sentence processing is sensitive to a memory target's representational complexity (Hofmeister, 2011). The term "complexity" is shorthand for the idea that discourse references can differ in semantic complexity via category hierarchy differences, e.g., "a thing" vs. "a stethoscope," as well as syntactic complexity. For instance, "the landmark on the bluff" encodes both syntactic and semantic features absent in "the landmark." In clefted constructions like those in 3, participants spent longer reading the head noun of the clefted element as the number of modifiers increased. At the words immediately following the subcategorizing verb (underlined below), however, reading times were faster given more features associated with the target. It is at this subcategorizing verb and the immediately following regions that we expect to observe signs of reactivation and retrieval of the representation in the cleft. Notably, the faster reading times for elaborated conditions do not appear until the subcategorizing verb or shortly thereafter:

(3) a. It was a communist that the members of the club banned from ever entering the premises.
b. It was an alleged communist that the members of the club banned from ever entering the premises.

c. It was an alleged Venezuelan communist that the members of the club banned from ever entering the premises.

Further experiments showed this same pattern even when holding the number of words and syntactic complexity constant, e.g., "which person" vs. "which soldier." At least in some contexts, therefore, syntactic and semantic processing of linguistic representations facilitates their retrieval from memory. It further suggests that recoverability increases gradiently with semantic processing—something that the list memory literature has so far not shown.

The present self-paced reading studies expand upon these findings in several ways. In Experiment 1, not only the target noun phrase, but also a preceding non-target noun phrase varies in syntactic and semantic complexity. In 4, for example, the matrix object noun phrase is the target for retrieval at "encouraged" and appears in either elaborated or non-elaborated form:

(4) The (senior foreign) diplomat contacted the (ruthless military) dictator who the activist from the United Kingdom encouraged to preserve natural habitats and resources

In addition, the preceding matrix subject noun phrase also varies between an elaborated and non-elaborated form. This manipulation of a competitor's complexity serves two purposes (note: "the activist" serves as a second potential competitor). First, it addresses the previously discussed question of whether elaborative processing linked to non-targets/competitors may facilitate differentiation at retrieval points. Such an idea is plausible from the perspective that providing more detail about any discourse referent or event lowers the chances that it will be confused with some other candidate for memory retrieval. Second, in 4 above, the key retrieval region ("banned") appears later in the complex sentences than in the simpler ones, opening the door to an explanation based on word position effects. Due to the manipulation of the complexity of multiple phrases in Experiment 1, it will be possible to directly assess whether the effects observable at retrieval sites are reducible to word position effects.

In Experiment 2, the essential components of von Restorff's design are carried over to the domain of sentence processing. Key words in the test sentences are systematically manipulated to make them superficially homogeneous or isolated with the expectation that this will give rise to longer encoding times. The question is whether superficial isolation or differentiation of words in sentences produces retrieval effects that are qualitatively similar to the effects of elaboration in online sentence processing. If they do, then we have evidence of a tight correspondence between implicit and explicit retrieval processes targeting linguistic stimuli.

As we shall see, both elaboration and isolation give rise to longer encoding times, but only the former yields strong evidence for faster reading times at sentence-internal retrieval sites. Moreover, while the elaboration associated with a non-target has striking downstream effects on encoding processes for other discourse referents, the evidence for an effect of non-target complexity on the retrieval of target representations is considerably weaker.

2. EXPERIMENT 1: TARGET AND NON-TARGET COMPLEXITY

2.1. PARTICIPANTS

Fifty-two University of Essex undergraduates participated in this study for course credit or payment. All participants identified themselves as native English speakers without significant exposure to a second language before the age of five. No participant data was removed on the basis of accuracy, as all participants scored above 67% correct.

2.2. METHODOLOGY AND MATERIALS

In this 2 × 2 self-paced, moving window experiment, 28 items varied in terms of the complexity of a target noun phrase and a non-target noun phrase in the same sentence. Specifically, all sentences contained a transitive matrix clause of the form [NP V NP], where the object noun phrase was modified by an object relative clause. The matrix subject (NP1) appeared with either 0 or 2 modifying words, as did the matrix object NP (NP2), as illustrated below:

(5) a. The congressman interrogated the general who a lawyer for the White House advised to not comment on the prisoners. (= SIMPLE SIMPLE)
b. The conservative U.S. congressman interrogated the general who a lawyer for the White House advised to not comment on the prisoners. (= COMPLEX SIMPLE)
c. The congressman interrogated the victorious four-star general who a lawyer for the White House advised to not comment on the prisoners. (= SIMPLE COMPLEX)
d. The conservative U.S. congressman interrogated the victorious four-star general who a lawyer for the White House advised to not comment on the prisoners. (= COMPLEX COMPLEX)

The subject of the object relative clause (NP3) was always of the form [DET NOUN]. At the critical embedded verb ("advised" in the example above), proper interpretation of the sentence requires retrieval of the representation referred to by NP2. It is also at such sentence internal retrieval sites that prior psycholinguistic evidence has repeatedly identified signs of similarity-based memory retrieval interference from competing representations (Gordon et al., 2001, 2002, 2006; Van Dyke and McElree, 2006).

Each participant saw only one condition of each item. All sentences were followed by a yes/no comprehension question, and participants received feedback if they answered incorrectly. The comprehension questions targeted information about one of the three referents introduced in the sentence, e.g., "Was the general advised not to comment on the prisoners?" with numerous questions asking about the relationship between two referents, e.g., "Did a photographer embarrass a celebrity?" In Experiment 1, mean comprehension accuracy across all trials, including fillers, was 84% (min = 70%, max = 97%). 70 fillers accompanied the

main experimental items for this experiment. Twenty eight of these were from an unrelated experiment.

Materials were presented and randomized with the reading time software LINGER v. 2.94, developed by Doug Rohde (available at http://tedlab.mit.edu/~dr/Linger/). The experimental items were randomized by the experimental software, and at least one filler separated each critical item. At the beginning of each trial, a fixation cross at the left of the screen appeared on the same line where the target sentence subsequently appeared. On pressing a key, the cross disappeared and the first word of the sentence was shown. Words not currently being read were not presented on screen and were not masked with dashes, i.e., the screen was blank except for the word currently being read. We opted for this method to prevent participants from using end-of-sentence information to modulate their reading rate, since the target sentences differed in overall length.

Prior to statistical analysis, raw reading times greater than 5000 ms or less than 100 ms were removed, affecting a total of 0.001% of the data. No additional outlier removal processes were performed. All data were analyzed regardless of comprehension accuracy in order to capture any reading time differences that may reflect memory retrieval failures. In other words, as we are investigating not only retrieval efficiency but also success, excluding trials that were incorrectly responded to would eliminate an important and relevant subset of the data on which retrieval of the target NP potentially failed. However, in the Supplementary Materials, we also present secondary analyses using only data from correctly answered trials.

Reading times were log-transformed to normalize the residuals and reduce the effect of extreme data points. Then, the log reading times for all stimuli (fillers included) were regressed against several predictors known to affect reading times in self-paced reading tasks: word length and log list position (Ferreira and Clifton, 1986; Hofmeister, 2011). Specifically, longer words predict longer reading times and later list positions predict faster reading times as participants progress through the experiment. The model estimating these effects included a random effects term for participants, i.e., by-participant random intercept adjustments. We used data from fillers in this process to produce maximally general estimates of word length and list position. The residuals of this model—RESIDUAL LOG READING TIMES—are the dependent variable analyzed here (**Figure 1** shows raw reading times to provide a more interpretable scale for the effects). All categorical predictors variables were sum coded to reduce effects of collinearity.

All analyses were conducted with Bayesian hierarchical models, fit with Stan and the `R` package `rstan`. We employed these models because they allow us to fit complex hierarchical models with maximal random effect structures that often do not converge using other popular linear regression packages such as `lme4`. Moreover, as noted in Husain et al. (2014), using Bayesian models allows us to assess and compare the weights of evidence for particular hypotheses. This means that we avoid categorizing effects as significant or non-significant, eschewing traditional statistical inference based on p-values. Instead, we make statistical inferences for particular hypotheses by computing the posterior probabilities for relevant parameters θ_i by sampling from their posterior distribution.

Each word region model used 4 chains, 5000 samples per chain, a warm-up of 2500 samples, and no thinning, resulting in 10,000 samples for each parameter estimate. All models contained fixed effect parameters for NP1 complexity, NP2 complexity, and their interaction. They also included by-participant random intercept adjustments and random slopes for all fixed effect terms (3 parameters), and by-item random intercept adjustments and random slopes for NP1 complexity, NP2 complexity, and their interaction (3 parameters). We utilized weak, uninformative priors for all key parameters, including participant and item adjustments. For each model, $P(\theta|data)$ indicates the probability that the parameter estimate is negative, i.e., speeding up occurs. For instance, an estimate that $P(\theta_{complex} < 0) = 0.99$ signifies that we can be 99% certain that complexity speeds up reading; in contrast, if $P(\theta_{complex} < 0) = 0.01$, we can infer with 99% certainty that complexity slows down reading. These probabilities

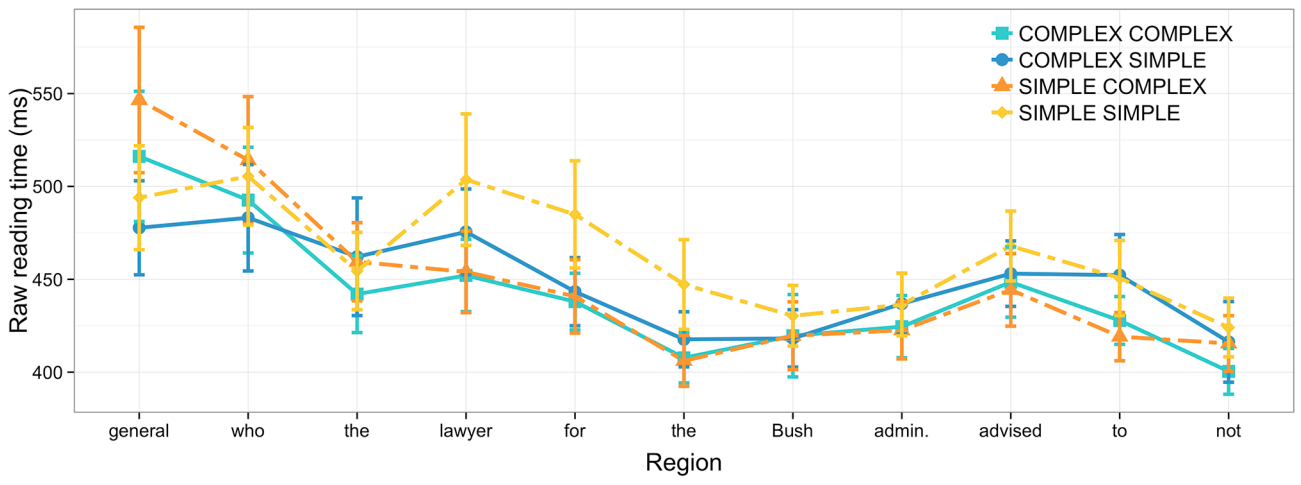


FIGURE 1 | Raw reading times in Experiment 1 by region; error bars show 95% confidence intervals.

were obtained by calculating the percentage of posterior samples above or below zero. To improve readability we will write $P(\theta < 0)$ for $P(\theta < 0|data)$.

Three regions are analyzed in Experiment 1: the head noun of NP2, the head noun of NP3, and the verb that subcategorizes for NP2. As reading time effects in self-paced reading experiments often spill over onto subsequent words, results for the word regions immediately after the relevant sites are also reported. No significant effects of the experimental manipulations on comprehension accuracy were found so they are not discussed here (see data in Supplementary Materials).

2.3. RESULTS

2.3.1. NP2 head noun

As shown in **Table 1**, greater syntactic and semantic complexity of NP2 leads to longer reading times at this region. Greater complexity of NP1, however, has a weaker effect in the opposite direction. That is, reading times at the NP2 head noun were somewhat faster when NP1 was complex, compared to when it was syntactically and semantically simple. There is no compelling evidence for an interaction at this word region.

2.3.2. NP3 head noun + spillover

Complexity of NP2 also has an effect on reading times at the head noun of NP3 (e.g., "lawyer"): reading times are faster when NP2 is relatively complex. At the word immediately following the head noun ("for" in 2.2), an interaction of NP1 & NP2 complexity arises, along with main effects of NP1 & NP2 complexity. This interaction stems from the fact that NP1 complexity leads to faster reading times only when NP2 is simple.

2.3.3. Relative clause verb + spillover

A main effect of NP2 complexity is evident at the critical relative clause verb: when NP2 is complex, reading times are faster than when NP2 is simple. Alongside this main effect, the results provide weak support of an interaction due to the fact that the complexity of NP1 affects reading times more when NP2 is simple. Put differently, there is no added processing facilitation due to the complexity of NP1 when NP2 is itself complex. The NP2 complexity effect also carries over onto the word immediately after the verb. In fact, the effect is even more pronounced at this region. Here, signs of an interaction are considerably weaker, as illustrated in **Figure 1**.

2.3.4. Correctly answered trials only

We conducted secondary, *post-hoc* analyses using only data from correctly answered trials to determine whether the observed complexity effects were tied to trials where participants answered incorrectly. As depicted in **Figure 2**, all main findings persist in this data subset with NP2 complexity effects at the NP3 and the relative clause verb slightly increasing in magnitude.

2.4. DISCUSSION

When readers encode additional syntactic and semantic features, they read faster at sentence-internal retrieval sites. This pattern holds, however, primarily for NP2—the downstream retrieval target. At the relative clause subject, reading times are faster when NP2 is syntactically and semantically complex, and this effect re-emerges at the retrieval triggering verb, continuing on into the spillover region.

Table 1 | Model summary for Experiment 1 for each region and fixed effect factor.

Region	Factor	Mean	CrI lower	CrI upper	$P(\beta < 0)$
Head noun	NP1 complexity	−0.012	−0.033	0.009	0.866
	NP2 complexity	0.038	0.011	0.065	0.002
	NP1 × NP2 complexity	−0.003	−0.024	0.017	0.621
RC subject head noun	NP1 complexity	−0.005	−0.025	0.014	0.702
	NP2 complexity	−0.022	−0.039	−0.005	0.996
	NP1 × NP2 complexity	0.002	−0.016	0.020	0.397
RC subject head noun + 1	NP1 complexity	−0.014	−0.032	0.002	0.955
	NP2 complexity	−0.012	−0.027	0.002	0.951
	NP1 × NP2 complexity	0.013	−0.002	0.027	0.043
RC verb	NP1 complexity	−0.005	−0.019	0.010	0.757
	NP2 complexity	−0.014	−0.027	−0.001	0.979
	NP1 × NP2 complexity	0.011	−0.005	0.025	0.079
RC verb + 1	NP1 complexity	−0.004	−0.009	0.017	0.281
	NP2 complexity	−0.019	−0.035	−0.004	0.991
	NP1 × NP2 complexity	0.006	−0.007	0.019	0.200

Summary includes the posterior 95% Credible Interval (CrI), i.e., the lower CrI refers to the 2.5% bound and the upper CrI refers to the 97.5% bound. $P(\beta < 0)$ indicates the probability that complexity slows reading times effects, i.e., values closer to 0 indicate slowing down and values closer to 1 indicate speeding up due to complexity.

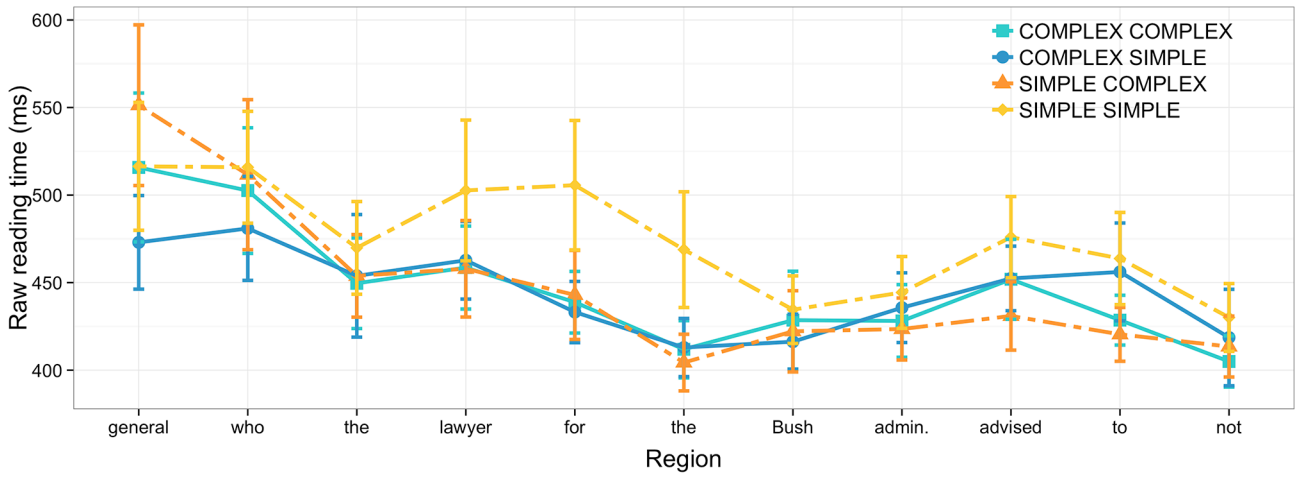


FIGURE 2 | Raw reading times in Experiment 1 for correctly answered trials; error bars show 95% confidence intervals.

Effects tied to NP1—the preceding non-target—are comparatively weaker and tied to the status of NP2. Whereas the effects of the complexity of NP2 show up at the head noun of NP3, the impact of NP1 complexity does not emerge until the head noun's spillover region. More tellingly, NP1 complexity affects reading rates selectively: only when NP2 is simple, and hence syntactically similar to NP3, does greater NP1 complexity reduce reading times. At the retrieval region, too, effects of the complexity of NP1 are weak compared to those of NP2. While there are hints at the relative clause verb that NP1 complexity has some facilitatory effects, such effects (1) do not have the duration of those tied to NP2, (2) are statistically weaker, and (3) only appear when NP2 is simple. In essence, differences in the feature-based complexity of a competitor do not weigh as significantly on retrieval in sentence comprehension as differences in target complexity. This suggests rather specific constraints on the dynamics of encoding and retrieval with respect to the computation of similarity-based interference in sentence processing that are dealt with in the General Discussion.

Two notable conclusions can be drawn from these results. First, word position alone cannot account for the reading time differences at the retrieval sites. Inside the relative clause, the COMPLEX-SIMPLE and SIMPLE-COMPLEX conditions match each other with respect to word position, yet display different profiles at the word following the subcategorizing verb. Moreover, if elaboration effects at the retrieval region owe their existence to a basic linkage between word position and reading rate, then we would expect the reading times for the conditions to be ordered according to word position. However, the COMPLEX-COMPLEX condition proved to be no faster than the SIMPLE-COMPLEX, despite the retrieval region appearing two words later in the sentence. Second, the lack of a main effect of NP1 complexity at the retrieval region argues against a general preference for maximal descriptiveness. Indeed, nowhere in the sentence does there seem to be a notable advantage for modifying both NPs in the matrix clause. As noted above, however, NP1 complexity does impact the processing of NP3 when NP2 is simple. We take this to mean that encoding interference arises at NP3 when all the NPs match in form, but altering the form of either of the preceding NPs mitigates these interference effects.

A valid concern with respect to these data concerns the relationship between the effects at NP3 and the verb. Are these separate effects, or do the effects at the verb simply reflect extended spillover effects that originate with processing NP3 in the above stimuli? This concern is amplified by signs of NP2 complexity effects at the region before the retrieval-triggering verb. Several arguments, however, speak against the interpretation that the differences at the verb and its spillover region reflect a continuation of previously initiated processes. First, a separate analysis revealed that the NP2 complexity effect at the verb remains intact even after including reading times from the word before the verb as a covariate ($\hat{\mu} = -0.022$; CrI Lower $= -0.038$; CrI Upper $= -0.006$: $P(\beta < 0) = 0.997$). Second, consideration of only correctly answered trials shows that the effects at the verb are magnified, while differences at the preceding region are minimized (see Supplementary Materials for model summaries). Some of the variation across conditions immediately prior to the verb thus comes from trials where encoding or retrieval processes may have been compromised. Further supporting this interpretation, it was found that several poorly-performing participants (who averaged 56% correct on the critical trials) were the primary source of reading times differences at the word region preceding the verb. In the case of these participants, it is indeed possible that encoding difficulties continued on into the retrieval region[1]. Taken together, these observations support the interpretation that the effects at the verb and subsequent word reflect cognitive processes that begin at the verb.

[1]This might be taken as justification to exclude these participants altogether; however, we see no reason to exclude participants because they encounter more encoding problems or read less accurately than their peers.

3. EXPERIMENT 2

If complexity effects arise during sentence processing because additional semantic or conceptual features distinguish representations from one another, this raises the question of whether all types of unique features distinguish comprehension-based representations. There may be nothing special, mnemonically speaking, about syntactic and semantic features in comprehension. Experiment 2 consequently looks at whether unique features in general stimulate faster processing at retrieval sites in comprehension. But this experiment also has a secondary purpose. In Experiment 1, longer encoding times match up with shorter reading times at or directly after the retrieval site. Thus, one take on the previous results is that additional semantic features stimulate more processing, which facilitates downstream retrieval. By manipulating the homogeneity of superficial features in Experiment 2, we address both issues due to the expectation that isolated word stimuli will not only generate contextually unique features (by definition), but will also lead to extended processing times during the encoding phase. The question is how this will bear, if at all, on the processing of words that trigger the retrieval of these encodings.

3.1. PARTICIPANTS

Forty-four UC-San Diego students participated in this study, in exchange for course credit. All subjects identified themselves as monolingual American English speakers without any known history of color blindness. The results from two participants were removed due to comprehension question accuracies below 67%.

3.2. METHODOLOGY AND MATERIALS

Thirty-two items were constructed with an object noun phrase in a transitive main clause modified by an object relative clause, as in 6 below. Textually, the conditions were identical to each other.

(6) The congressman interrogated the **general** who the lawyer for the Bush administration advised ___ to not comment on the detainees.

To manipulate processing during the encoding phase, the head noun of the object NP ("general" above) appeared either in the same color as the surrounding sentence text (white), or else in an incongruent color (bright green). Additionally, the color of the word that triggered retrieval ("advised") also varied between congruent and incongruent. This second manipulation provides a needed check to ensure that participants do not read later word regions faster because of anticipation for an incongruently colored word. Moreover, in the condition with the green head noun and green verb, we can assess whether reinstating features of the encoding phase aids in retrieval. Hence, each item had four conditions (WHITE-WHITE, WHITE-GREEN, GREEN-WHITE, GREEN-GREEN), but each subject saw only one condition of each item.

Participants received instructions that the color of the words in the sentences was immaterial to the task and that they did not need to respond to color changes. Yes/no comprehension questions followed each item, and participants received negative feedback if they answered a question incorrectly. Sixty fillers accompanied these critical items: 20 with 0 green words, 20 with 1 green word, and 20 with 2 green words. For filler items with 1 green word, the word was randomly selected from all words in the sentence. For fillers with 2 green words, one appeared randomly in the the first half of the sentence and the other in the second half. All fillers had a syntactic structure different from that used in the critical items.

The materials were presented in a self-paced, center presentation paradigm via a propriety software package. Only one version of each item appeared on each of four experimental lists, whose contents were pseudo-randomized such that at least one filler intervened between each critical item. A fixation cross in the center of the screen appeared before each trial, and a comprehension question followed every experimental trial, including fillers. Participants received feedback only on incorrectly answered trials.

The outlier removal process, computation of residual log reading times, and Bayesian analysis procedure all followed those used in Experiment 1. As in that experiment, there were no differences in comprehension accuracy (GREEN-GREEN = 76%, GREEN-WHITE = 77%, WHITE-WHITE = 76%, WHITE-GREEN = 76%). Here, we analyze residual log reading times at the head noun of the matrix object phrase and the relative clause verb that triggers its retrieval.

3.3. RESULTS

At the object head noun, incongruent, green words slow reading times, compared to the congruent, white words (see **Figure 3**). Similarly, looking at reading times at the retrieval region ["advised" in (3.2)], a perceptually incongruent, green verb slows reading speed compared to a congruent, white one.

In contrast to the pattern observed in Experiment 1, the increased encoding time at the object head noun due to superficial incongruence leads to relatively weak facilitation effects at the retrieval site, as shown in **Table 2**. In fact, the mean parameter value resides less than one standard deviation (=0.011) from zero, according to the model results. The mean value for the condition where both the noun and the verb are incongruently colored reflects slightly faster reading than for the condition where only the verb is incongruent (GREEN-GREEN: -0.015, $SE = 0.021$; WHITE-GREEN: 0.011, $SE = 0.024$). This difference of roughly one standard error is why the model acknowledges a relatively weak effect of noun color (and an interaction with verb color) on reading times at the verb. At regions after the verb, there is no evidence that processing an incongruently colored target noun facilitates processing.

3.4. DISCUSSION

Increased processing times triggered by incongruent stimuli at the encoding site had weak effects on processing at the retrieval site when compared to the complexity effects observed in Experiment 1. Only when the relevant perceptual features were reinstated at the retrieval site was there any numerical retrieval advantage for perceptually incongruous stimuli. Even in this case, the facilitating effects were quite mild and would be deemed insignificant on classical frequentist methods of analysis. These findings imply that contextually unique features do not necessarily lead to

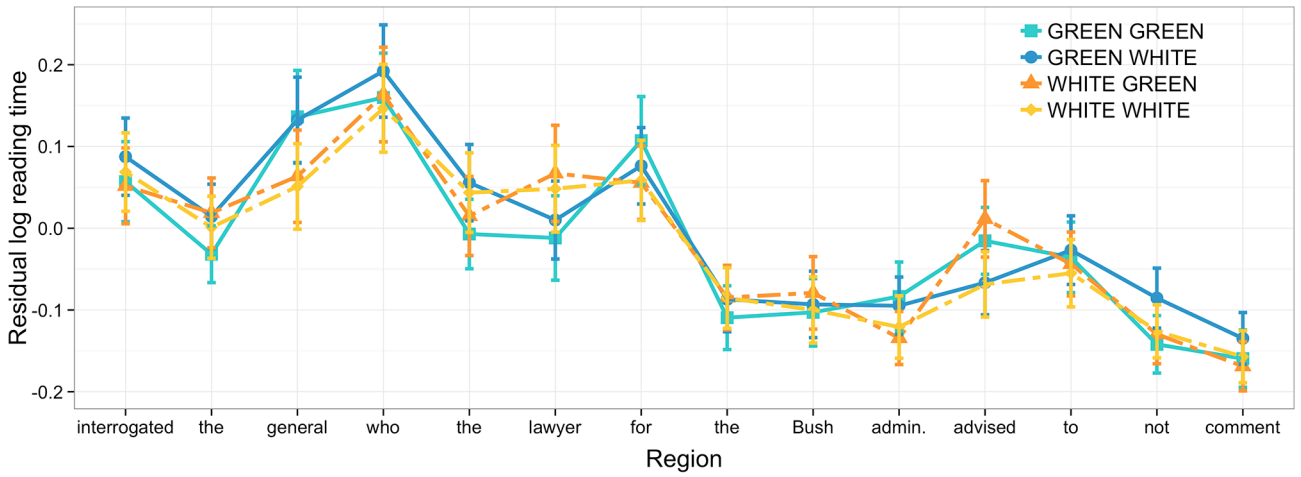


FIGURE 3 | Residual log reading times at verb in Experiment 2; error bars show 95% confidence intervals.

Table 2 | Model summary for Experiment 2.

Region	Factor	Mean	CrI lower	CrI upper	P($\beta < 0$)
Head noun	Noun color	0.039	0.007	0.071	0.008
	Verb color	0.005	−0.025	0.034	0.362
	Noun color × Verb color	−0.001	−0.031	0.028	0.532
RC verb	Noun color	−0.007	−0.029	0.016	0.731
	Verb color	0.032	0.008	0.056	0.001
	Noun color × Verb color	−0.008	−0.034	0.017	0.734

Summary includes the posterior 95% Credible Interval (CrI), i.e., the lower CrI refers to the 2.5% bound and the upper CrI refers to the 97.5% bound. P($\beta < 0$) indicates the probability that incongruence slows reading times effects, i.e., values closer to 0 indicate slowing down and values closer to 1 indicate speeding up due to incongruence.

improved memory performance, nor does increased processing time.

These findings may initially seem to contrast with memory results for recognition/recall of items presented in lists. For instance, von Restorff (1933) observed better recognition for words that appeared in superficially incongruent states. Similar findings of improved memory performance for superficially incongruent linguistic items (within mixed lists, but not unmixed lists) appear in Bruce et al., 1976, Hunt and Elliot, 1980, Hunt, 1995, Dunlosky et al., 2000, *inter alia*.

However, the current evidence reinforces the idea that the memory retrieval context is of utmost importance—a point frequently reiterated by memory researchers such as Tulving, Nairne, and others. In the present case, color or other superficial orthographical features rarely matter in written, sentence comprehension. Particularly if subjects are requested to ignore such information, there is little reason for subjects to recruit such potentially distinctive features in memory retrieval, whether or not they elicit more processing. In contrast, standard list recall or recognition tasks are novel encoding and retrieval contexts for participants—we are not standardly shown a list of words and then asked to retrieve them later, so we have few if any entrained habits. Consequently, in such novel circumstances, participants reasonably utilize all manner of perceptual features in recovering representations from memory.

In short, this experiment establishes that the uniqueness effects in language comprehension depend heavily on the retrieval context. What counts as unique critically depends on the nature and demands imposed at the retrieval site. Ultimately, if some set of representational features are unimportant for memory retrieval, then their congruence with other local feature appears to also have little import for memory retrieval.

4. GENERAL DISCUSSION

Increased processing during the encoding phase leads to more efficient retrieval processing in sentence comprehension, but only under certain conditions. Experiment 1 illustrated that increased processing associated with the downstream target benefits retrieval-related processing, whereas processing related to non-targets had relatively weak, short-lived effects that only arose when the target itself was not elaborated. Experiment 2 expanded on this by showing that not just any sort of extra processing facilitates memory (even for targets)—indeed, the results suggest that it is not about processing *per se* so much as the role of

the features themselves in the retrieval process. In many respects, these results parallel the findings of studies assessing the effects of elaboration on long-term memory performance for linguistic stimuli (Stein et al., 1978; Eysenck, 1979; Jacoby and Craik, 1979; Reder, 1980; Bradshaw and Anderson, 1982; Reder et al., 1986; McDaniel et al., 1988). At the same time, they add to these studies by showing that memory performance improves as meaning-related processing increases for linguistic stimuli in the context of sentence comprehension. Secondly, they demonstrate that these effects occur even in covert retrieval settings, where the time constraints of real-time comprehension limit the options for retrieval strategies. Third, the results from the final experiment demonstrate that unique representational target features and increased processing do not always lead to improved memory retrieval.

Both sets of findings—the advantage of additional processing for targets compared to non-targets, and the fact that increased processing time does not necessarily benefit memory retrieval—can be understood through the lens of the short-term, feature-based retrieval model of Nairne (1990, 2001, 2006), with some minor new assumptions (several other memory models make similar predictions, e.g., Oberauer and Kliegl, 2006 and Shiffrin, 2003, although the details differ). In Nairne's model, memory items are represented as a vector of features, e.g., [C X 1 2 3]. Retrieval cues consist of lingering, typically blurry, records of the immediate past, e.g., [C X ? 2 3], as well as cues from the local retrieval context. In turn, these two sets of cues form a memory probe that is compared against a set of candidate memory items. The ultimate objective is to "redintegrate" the retrieval cues with a memory item, as the cues by themselves cannot be directly interpreted (Ericsson and Kintsch, 1995). The probability of retrieving an event E_1, given a retrieval probe X_1 depends upon the similarity or feature-overlap of X_1 and E_1, as well as the similarity of X_1 to other memory candidates:

$$P_r(E_1|X_1) = \frac{s(X_1, E_1)}{\sum s(X_1, E_n)} \quad (1)$$

The similarity between a memory item and a retrieval probe is determined by the number of mismatching features divided by the total number of compared features (d):

$$s(X_1, E_1) = e^{-d(X_1, E_1)} \quad (2)$$

Because retrieval probes consist of remnants of the original encoding process that need to be interpreted by comparing them against candidate memory items, any contextually unique features in a target will improve the chances for successful retrieval. In short, a target's recoverability increases if it possesses a feature that no other competitor shares.

Nairne (2006) employs this model to explain isolation or distinctiveness effects, since odd/bizarre items possess features that mismatch with the features of some homogeneous background set. For instance, imagine a context where the original encoding is perfectly intact and acts as the sole source of retrieval cues, e.g., $X_1 = E_1$. Any contextually unique features will increase the dissimilarity or mismatch between the retrieval cues and competitors, even though contextual uniqueness does not directly affect the similarity value between the target and retrieval probe.

An implied consequence of such a theory is that simply adding features to a target is predicted to increase the odds of sampling from memory, so long as these features are unique. **Table 3** shows how the probability of sampling a target increases as the number of mismatching features between the target and non-targets increases, even though the number of shared features remains constant (see Hofmeister et al., 2013 for an application of this model to the processing and acceptability of multiple wh-questions in English). The added features Q, R, & N in the undegraded probe lack any correlates in the competitors, meaning that the mismatch between them and the probe increases, effectively upping the chances for sampling the target.

As **Figure 4** illustrates (left panel), the effect of adding mismatching or contextually unique features faces some restrictions: increasing the number of mismatches yields diminishing returns, ultimately asymptoting at a level that depends upon the number of features involved and the number of feature matches. In less formal terms, adding a little unique, diagnostic information can be quite helpful for memory retrieval, but adding lots of unique information is not likely to contribute much more. This model also predicts that the number of competitors affects retrieval probability much more dramatically than the number of overlapping features. On the right side, **Figure 4** shows that going from one competitor to three competitors which each share two features with the probe nearly halves the chances of retrieval. In contrast, the difference between two competitors with 2 vs. 10 matching features never exceeds 10% (see left side of **Figure 4**).

A key component of this type of model is that a fragile copy of the original encoding process stored in primary memory provides a source of retrieval cues. This makes explicit the idea that syntactic and semantic features not directly invoked by the local sentence context can influence retrieval processes, in contrast to assumptions that only the similarity of features "grammatically derived from the current word and context" enter into considerations of similarity-based interference (Lewis et al., 2006, p. 448)[2]. Sentence processing models built upon the latter kind of assumption face difficulty explaining some classic retrieval interference effects in the sentence processing literature (Logačev and Vasishth, 2012). For instance, Gordon et al. (2001) show that processing in object-cleft sentences like 7 is easier at the subcategorizing verb ("saw") when the two NPs are of different types

[2]Current sentence processing models are not without means to explain effects of complexity on memory retrieval. For instance, on the ACT-R-based theory of Lewis and Vasishth (2005), processing syntactic material that modifies some previously constructed representation requires the restoration of the stored memory item. This retrieval process, in turn, raises the overall activation level of the item, making it easier to retrieve subsequently. Thus, complexity-based effects on retrieval emerge most straightforwardly as the byproduct of encoding processes. Moreover, additional study time potentially allows for more accurate encoding, providing greater chances that target features will be cued at the retrieval site (see also Shiffrin, 2003). However, as retrieval cues are limited to those provided by local grammatical context, there is no guarantee that unique semantic or syntactic features will factor into estimates of similarity and thus retrieval difficulty.

(proper name vs. definite description), but that such effects are absent in subject relativization constructions:

(7) It was John/the barber that the lawyer/Bill saw in the parking lot.

These effects are commonly understood in terms of similarity-based interference: if the target noun phrase overlaps in form with another local noun phrase that appears before the verb, memory retrieval difficulty ensues, ostensibly because the retrieval cues match multiple memory representations. As the second NP occurs after the verb in subject relatives, no possibility for interference exists. Notably, the verb triggering retrieval ("saw") does not itself supply cues as to the nominal type of the clefted element; indeed, no language appears to explicitly code whether a verb requires a lexical, pronominal, or some other type of nominal argument. So, if the similarity effects arise because retrieval cues match multiple representations, then those cues must come from a source besides the verb. The original encoding of the target provides the most obvious source of such cues. Not only does this open up a way to explain similarity-based effects due to overlapping referential form, it can also accommodate phonological similarity effects such as the observed reading time contrast at the embedded verb in sentences like "The baker that the banker sought found the house" vs. "The runner that the banker sought found the house" (Acheson and MacDonald, 2011)[3].

Table 3 | Similarity values and predicted sampling probabilities for two retrieval contexts.

Cue	Traces	Similarity	Samp. prob.
[C C 2 3 1]	[C C 1 2 3]	0.55	0.26
	[C C 2 3 1]	1.0	0.48
	[C C 3 1 2]	0.55	0.26
[C C 2 3 1 Q R N]	[C C 1 2 3]	0.47	0.24
	[C C 2 3 1 Q R N]	1.0	0.52
	[C C 3 1 2]	0.47	0.24

The current findings add a further data point to our developing picture of similarity-based interference in sentence processing: non-target distinctiveness has a weaker role to play in retrieval interference than target distinctiveness. These effects can be straightforwardly accommodated with some specifications about how similarity is calculated. Following Nairne (2006), let's assume that similarity at retrieval sites is calculated by establishing mismatches with the lingering features of a target's encoding remnant and any other features in the retrieval probe. A memory probe such as [C X 1 2 3] will mismatch equally with a competitor representation like [C X 4 5 6] as [C X 4 5 6 L M], e.g., 3 out of 5 probe features will mismatch with competitor features. In other words, it is the number of features in the probe that determine how many mismatches there can be, and not the number of features in a memory retrieval candidate. Adding unique features to some non-target, therefore, will not directly affect the probability of sampling the target because it does not contribute to the set of retrieval cues.

The data hint nonetheless at some retrieval effects linked to the elaboration of non-targets, specifically when the retrieval target itself was syntactically and semantically simple. This would seem to initially contradict the above view that the uniqueness of non-targets does not directly bear on retrieval efficiency. There is no contradiction, however, if these non-targets effects are byproducts of encoding interference. That is, we presume that the uniqueness of features in non-target nominals affects how other

[3]Acheson and MacDonald (2011) illustrate similar effects in subject relatives, as well, suggesting that phonological similarity gives rise to encoding interference and not simply retrieval interference.

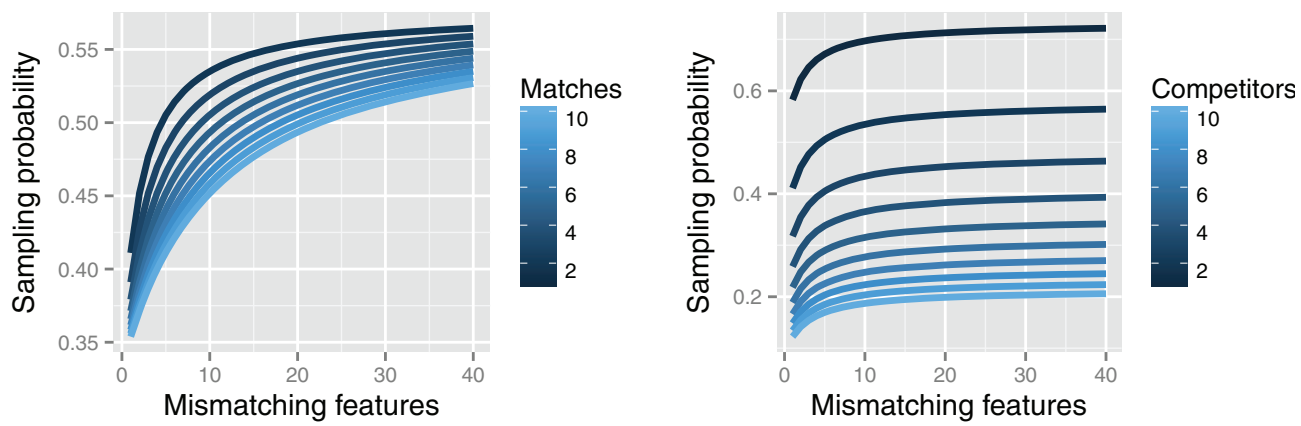


FIGURE 4 | Left: Relationship between number of unique target features (mismatching with non-targets) and average sampling probability of target with two competitors. In descending order, the lines show the varying sampling probability curves for 2 to 10 probe features matching with each competitor. **Right:** Relationship between number of unique target features and average sampling probability of target as a function of the number of competitors (from 1 to 10 in descending order), assuming two matching features between the probe and each competitor. The retrieval sampling curves illustrate the diminishing effects of mismatching features and the relatively greater effect of the number of competitors compared to the number of matching features.

local nominals, including downstream targets, are encoded, and indirectly influence retrieval operations *via such encoding effects*. Even more generally, encoding interference feeds into retrieval interference.

Already, evidence exists that similarity between linguistic representations in memory and those being encoded can lead to processing disruptions, during both encoding and retrieval stages (Gordon et al., 2002; Acheson and MacDonald, 2011). For example, Gordon et al. (2002) provide evidence of reading slowdowns when words on a sentence-external memory list are similar to key words inside the sentence, e.g., proper names vs. definite descriptions, both at the encoding site for the sentence-internal words and later at retrieval sites for those same words. We would add to this by hypothesizing that encoding interference may contribute to the degradation of memory representations, following research that suggests that forgetting in short-term memory for linguistic representations can stem from feature overwriting (Oberauer and Lange, 2008; Oberauer, 2009). Because these features that are susceptible to overwriting also contribute to retrieval cues on the account sketched above, feature loss could compromise any cue-based retrieval process.

Applying these hypotheses to the results of Experiment 1, encoding interference emerges as an indirect (and accordingly, weaker) contributor to retrieval differences, beyond what is predicted by the model of memory retrieval inspired by Nairne. Specifically, similarity between the referring expressions determines encoding interference, which can affect the integrity of the trace for the target nominal. So, when NP1 is complex and NP3 is simple or vice versa, this translates to a reduced danger of feature overwriting, compared to when they are both simple. In turn, the potential for retrieval interference is mitigated when the two initial NPs mismatch in complexity, because the trace for NP2 is more likely to be intact. Things are somewhat more complicated when NP1 and NP2 are both complex: while overlapping in structural form, the NPs carry more unique semantic features than their simpler counterparts. In this case, we tentatively take the results to mean that encoding interference is relatively low, compared to the case where both NPs are simple, but not any lower than when just one such NP is complex. These ideas require further tests to be substantiated, as the current experiments were not designed to test them. Nonetheless, we maintain that the relatively weak effects of non-targets can best be explained by appealing to the effect of encoding interference on memory retrieval.

Notably, redintegration-based models of memory do not require that every perceivable feature matters for memory retrieval. Listeners or readers may preferentially not encode some features in typical language settings, such as modality-specific features or exclude such features from the retrieval probe based on prior experience of the efficacy of such features. The advantage of increased processing thus depends upon the discourse context and the extent to which processing engenders unique features that come into play during the retrieval stage. From this perspective, encoding manipulations cannot have a predictable effect on memory in the absence of information about the encoding and retrieval contexts—what other memory candidates are available and what the retrieval cues are.

The results of Experiment 2 align with this perspective, in light of the absence of isolation or superficial processing effects. Modality-dependent features, such as orthography, font style, text color, etc., often play a large role in various laboratory tests of memory and in effects such as the auditory recency effect, but they appear to have a lesser role in guiding retrieval in sentence processing contexts. Such contrasts, though, are explicable in terms of task demands and prior experience. Word recall and recognition tasks lie outside the typical range of personal pastimes, whereas sentence comprehension is an everyday occurrence. This arguably leads participants to utilize a wider range of possible retrieval cues in word recall tasks, whereas prior experience with sentence processing would bias against the use of modality-specific features to distinguish memory representations. Instead, modality-independent features—properties that largely remain constant across presentations or modalities such as syntactic category and meaning—provide the basis for restoring linguistic representations during sentence processing because of their diagnostic potential. Thus, it is due to the fact that discrimination between language representations in sentence comprehension depends on syntactic and semantic features that the uniqueness of these features bears on determinations of retrieval ease and success. Correspondingly, the primary source of retrieval difficulty in language comprehension—overlapping semantic and syntactic representations and the resulting interference—is what gives additional linguistic processing mnemonic value, and why other types of processing such as superficial processing have little mnemonic value.

5. CONCLUSION

These tests of implicit memory establish that elaboration effects occur in online sentence processing tasks, as they do in explicit tests of memory. In Experiment 1, we found that increased processing of syntactic and semantic features connected to the target benefits memory retrieval in sentence processing; however, additional processing directed toward non-targets had substantially weaker effects on processing at retrieval sites. In Experiment 2, it was established that the processing of superficial features or features connected to non-targets yielded insubstantial processing advantages at retrieval sites, despite leading to longer encoding times. As sentence processing demands differ from those of explicit memory tasks, it is unsurprising that the effects of encoding manipulations can differ drastically across tasks with inherently different retrieval contexts. This apparent dynamic interaction between encoding and retrieval led Tulving (1983, p. 239) to argue against any statements of the form that "encoding operations of class X are more effective than encoding operations of class Y" (see also Neath and Surprenant, 2005 for a recent review). In short, encoding manipulations are unpredictable without additional information about the nature of the retrieval task and the background of competing representations.

The comparison of memory findings in the broader psychology and psycholinguistics literature also led to a unified theoretical account of distinctiveness effects, applicable across tasks. Capturing the interplay between representational uniqueness and retrieval probability, Nairne's feature-based model provides a means for introducing retrieval cues that are unlikely

to be cued by local grammatical memory triggers via the use of a fragile copy of the original encoding. This fills a critical gap in cue-based models of retrieval in sentence processing by pointing to alternative sources of retrieval cues beyond the local context, thus accounting for a variety of otherwise unexplained similarity-based effects in sentence processing.

ACKNOWLEDGMENTS

Valuable feedback on this article was received from Marta Kutas, Rick Lewis, and Michael Shvartsman. This research was supported by NIH Training Grant T32-DC000041 via the Center for Research in Language at UC-San Diego.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fpsyg.2014.01237/abstract

REFERENCES

Acheson, D. J., and MacDonald, M. C. (2011). The rhymes that the reader perused confused the meaning: phonological effects during on-line sentence comprehension. *J. Mem. Lang.* 65, 193–207. doi: 10.1016/j.jml.2011.04.006

Bradshaw, G., and Anderson, J. R. (1982). Elaborative encoding as an explanation of levels of processing. *J. Verb. Learn. Verb. Behav.* 21, 165–174. doi: 10.1016/S0022-5371(82)90531-X

Bruce, D., Gaines, I., and Marion, T. (1976). Tests of an organizational hypothesis of isolation effects in free recall. *J. Verb. Learn. Verb. Behav.* 15, 59–72. doi: 10.1016/S0022-5371(76)90007-4

Craik, F., and Lockhart, R. (1972). Levels of processing: a framework for memory research. *J. Verb. Learn. Verb. Behav.* 11, 671–684. doi: 10.1016/S0022-5371(72)80001-X

Craik, F., and Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *J. Exp. Psychol. Gen.* 104, 268–294. doi: 10.1037/0096-3445.104.3.268

Dunlosky, J., Hunt, R., and Clark, E. (2000). Is perceptual salience needed in explanations of the isolation effect? *J. Exp. Psychol. Learn. Mem. Cogn.* 26, 649–657. doi: 10.1037/0278-7393.26.3.649

Ericsson, K., and Kintsch, W. (1995). Long-term working memory. *Psychol. Rev.* 102, 211–244. doi: 10.1037/0033-295X.102.2.211

Eysenck, M. (1979). "Depth, elaboration, and distinctiveness," in *Levels of Processing in Human Memory*, eds L. Cermak and F. Craik (Hillsdale, NJ: Lawrence Erlbaum Associates), 89–118.

Ferreira, F., and Clifton, C. (1986). The independence of syntactic processing. *J. Mem. Lang.* 25, 348–368. doi: 10.1016/0749-596X(86)90006-9

Fisher, R., and Craik, F. (1977). Interaction between encoding and retrieval operations in cued recall. *J. Exp. Psychol. Hum. Learn. Mem.* 3, 701–711. doi: 10.1037/0278-7393.3.6.701

Gallo, D., Meadow, N., Johnson, E., and Foster, K. (2008). Deep levels of processing elicit a distinctiveness heuristic: evidence from the criterial recollection task. *J. Mem. Lang.* 58, 1095–1111. doi: 10.1016/j.jml.2007.12.001

Gordon, P., Hendrick, R., and Johnson, M. (2001). Memory interference during language processing. *J. Exp. Psychol. Learn. Mem. Cogn.* 27, 1411–1423. doi: 10.1037/0278-7393.27.6.1411

Gordon, P., Hendrick, R., Johnson, M., and Lee, Y. (2006). Similarity-based interference during language comprehension: evidence from eye tracking during reading. *J. Exp. Psychol. Learn. Mem. Cogn.* 32, 1304–1321. doi: 10.1037/0278-7393.32.6.1304

Gordon, P., Hendrick, R., and Levine, W. (2002). Memory-load interference in syntactic processing. *Psychol. Sci.* 13, 425–430. doi: 10.1111/1467-9280.00475

Hirshman, E., Whelley, M., and Palij, M. (1989). An investigation of paradoxical memory effects. *J. Mem. Lang.* 28, 594–609. doi: 10.1016/0749-596X(89)90015-6

Hofmeister, P. (2011). Representational complexity and memory retrieval in language comprehension. *Lang. Cogn. Process.* 26, 376–405. doi: 10.1080/01690965.2010.492642

Hofmeister, P., Jaeger, T., Sag, I., Arnon, I., and Snider, N. (2013). The source ambiguity problem: distinguishing effects of grammar and processing on acceptability judgments. *Lang. Cogn. Process.* 28, 48–87. doi: 10.1080/01690965.2011.572401

Hunt, R. (1995). The subtlety of distinctiveness: what von Restorff really did. *Psychon. Bull. Rev.* 2, 105–112. doi: 10.3758/BF03214414

Hunt, R., and Elliot, J. (1980). The role of nonsemantic information in memory: orthographic distinctiveness effects on retention. *J. Exp. Psychol. Gen.* 109, 49–74. doi: 10.1037/0096-3445.109.1.49

Hunt, R., and Worthen, J. (2006). *Distinctiveness and Memory*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780195169669.001.0001

Husain, S., Vasishth, S., and Srinivasan, N. (2014). Strong expectations cancel locality effects: evidence from Hindi. *PLoS ONE* 9:e100986. doi: 10.1371/journal.pone.0100986

Hyde, T., and Jenkins, J. (1973). Recall for words as a function of semantic, graphic, and syntactic orienting tasks. *J. Verb. Learn. Verb. Behav.* 12, 471–480. doi: 10.1016/S0022-5371(73)80027-1

Jacoby, L., and Craik, F. (1979). "Effects of elaboration of processing at encoding and retrieval: trace distinctiveness and recovery of initial context," in *Levels of Processing in Human Memory*, eds L. Cermak and F. Craik (Hillsdale, NJ: Lawrence Erlbaum Associates), 1–21.

Kluender, R., and Kutas, M. (1993). Bridging the gap: Evidence from ERPs on the processing of unbounded dependencies. *J. Cogn. Neurosci.* 5, 196–214. doi: 10.1162/jocn.1993.5.2.196

Lewis, R., and Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cogn. Sci.* 29, 1–45. doi: 10.1207/s15516709cog0000_25

Lewis, R., Vasishth, S., and Van Dyke, J. (2006). Computational principles of working memory in sentence comprehension. *Trends Cogn. Sci.* 10, 447–454. doi: 10.1016/j.tics.2006.08.007

Logačev, P., and Vasishth, S. (2012). "Case matching and conflicting bindings interference," in *Case, Word Order and Prominence*, eds M. Lamers and P. de Swart (Dordrecht: Springer), 187–216. doi: 10.1007/978-94-007-1463-2_9

McDaniel, M., Dunay, P., Lyman, B., and Kerwin, M. L. (1988). Effects of elaboration and relational distinctiveness on sentence memory. *Am. J. Psychol.* 101, 357–369. doi: 10.2307/1423084

McElree, B. (2000). Sentence comprehension is mediated by content-addressable memory structures. *J. Psychol. Res.* 29, 111–123. doi: 10.1023/A:1005184709695

Moscovitch, M., and Craik, F. (1976). Depth of processing, retrieval cues, and uniqueness of encoding as factors in recall. *J. Verb. Learn. Verb. Behav.* 15, 447–458. doi: 10.1016/S0022-5371(76)90040-2

Nairne, J. (1990). A feature model of immediate memory. *Mem. Cogn.* 18, 251–269. doi: 10.3758/BF03213879

Nairne, J. (2001). "A functional analysis of primary memory," in *The Nature of Remembering: Essays in Honor of Robert G. Crowder*, eds H. Roediger, J. S. Nairne, and A. Suprenant (Washington, DC: American Psychological Association), 282–296.

Nairne, J. (2006). "Modeling distinctiveness: implications for general memory theory," in *Distinctiveness and Memory*, eds R. Hunt and J. Worthen (New York, NY: Oxford University Press), 27–46.

Neath, I., and Surprenant, A. (2005). "Mechanisms of memory," in *Handbook of Cognition*, eds K. Lamberts and R. L. Goldstone (London: Sage), 221–238. doi: 10.4135/9781848608177.n9

Nicol, J., and Swinney, D. (1989). The role of structure in coreference assignment during sentence comprehension. *J. Psychol. Res.* 18, 5–19. doi: 10.1007/BF01069043

Oberauer, K. (2009). Interference between storage and processing in working memory: feature overwriting, not similarity-based competition. *Mem. Cogn.* 37, 346–357. doi: 10.3758/MC.37.3.346

Oberauer, K., and Kliegl, R. (2006). A formal model of capacity limits in working memory. *J. Mem. Lang.* 55, 601–626. doi: 10.1016/j.jml.2006.08.009

Oberauer, K., and Lange, E. (2008). Interference in verbal working memory: distinguishing similarity-based confusion, feature overwriting, and feature migration. *J. Mem. Lang.* 58, 730–745. doi: 10.1016/j.jml.2007.09.006

Osterhout, L., and Swinney, D. (1993). On the temporal course of gap-filling during comprehension of verbal passives. *J. Psychol. Res.* 22, 273–286.

Reder, L. (1980). The role of elaboration in the comprehension and retention of prose: a critical review. *Rev. Educ. Res.* 50, 5–53. doi: 10.3102/00346543050001005

Reder, L., Charney, D., and Morgan, K. (1986). The role of elaborations in learning a skill from an instructional text. *Mem. Cogn.* 14, 64–78. doi: 10.3758/BF03209230

Shiffrin, R. (2003). Modeling memory and perception. *Cogn. Sci.* 27, 341–378. doi: 10.1207/s15516709cog2703_2

Stein, B., Morris, D., and Bransford, J. (1978). Constraints on effective elaboration. *J. Verb. Learn. Verb. Behav.* 17, 707–714. doi: 10.1016/S0022-5371(78)90423-1

Tanenhaus, M., Carlson, G., and Seidenberg, M. (1985). "Do listeners compute linguistic representations?," in *Natural Language Parsing: Psycholinguistic, Theoretical, and Computational Perspectives*, eds A. Zwicky, L. Kartunnen, and D. Dowty (London; New York: Cambridge University Press), 359–408.

Tulving, E. (1983). *Elements of Episodic Memory*. Oxford: Clarendon Press.

Van Dyke, J., and Lewis, R. (2003). Distinguishing effects of structure and decay on attachment and repair: a cue-based parsing account of recovery from misanalyzed ambiguities. *J. Mem. Lang.* 49, 285–316. doi: 10.1016/S0749-596X(03)00081-0

Van Dyke, J., and McElree, B. (2006). Retrieval interference in sentence comprehension. *J. Mem. Lang.* 55, 157–166. doi: 10.1016/j.jml.2006.03.007

Van Dyke, J. A. (2007). Interference effects from grammatically unavailable constituents during sentence processing. *J. Exp. Psychol. Learn. Mem. Cogn.* 33:407. doi: 10.1037/0278-7393.33.2.407

Van Dyke, J. A., and McElree, B. (2011). Cue-dependent interference in comprehension. *J. Mem. Lang.* 65, 247–263. doi: 10.1016/j.jml.2011.05.002

von Restorff, H. (1933). Über die Wirkung von Bereichsbildungen im Spurenfeld. *Psychol. Res.* 18, 299–342. doi: 10.1007/BF02409636

Watkins, M., LeCompte, D., and Kim, K. (2000). Role of study strategy in recall of mixed lists of common and rare words. *J. Exp. Psychol. Learn. Mem. Cogn.* 26, 239–245. doi: 10.1037/0278-7393.26.1.239

Conflict of Interest Statement: The Associate Editor Claudia Felser declares that, despite being affiliated to the same institution as the author Shravan Vasishth, the review process was handled objectively and no conflict of interest exists. The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 10 May 2014; accepted: 11 October 2014; published online: 12 December 2014.

*Citation: Hofmeister P and Vasishth S (2014) Distinctiveness and encoding effects in online sentence comprehension. Front. Psychol. **5**:1237. doi: 10.3389/fpsyg.2014.01237 This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology.*

Copyright © 2014 Hofmeister and Vasishth. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 15 March 2016
doi: 10.3389/fpsyg.2016.00374

Elaboration over a Discourse Facilitates Retrieval in Sentence Processing

Melissa Troyer[1], Philip Hofmeister[2] and Marta Kutas[1,3]*

[1] Department of Cognitive Science, University of California at San Diego, La Jolla, CA, USA, [2] Department of Cognitive, Linguistic, and Psychological Sciences, Brown University, Providence, RI, USA, [3] Department of Neurosciences, University of California at San Diego, La Jolla, CA, USA

Language comprehension requires access to stored knowledge and the ability to combine knowledge in new, meaningful ways. Previous work has shown that processing linguistically more complex expressions ('*Texas cattle rancher*' vs. '*rancher*') leads to slow-downs in reading during initial processing, possibly reflecting effort in combining information. Conversely, when this information must subsequently be retrieved (as in filler-gap constructions), processing is *facilitated* for more complex expressions, possibly because more semantic cues are available during retrieval. To follow up on this hypothesis, we tested whether information distributed across a short discourse can similarly provide effective cues for retrieval. Participants read texts introducing two referents (e.g., two senators), one of whom was described in greater detail than the other (e.g., '*The Democrat had voted for one of the senators, and the Republican had voted for the other, a man from Ohio who was running for president*'). The final sentence (e.g., '*The senator who the* {*Republican*/*Democrat*} *had voted for*…') contained a relative clause picking out either the Many-Cue referent (with '*Republican*') or the One-Cue referent (with '*Democrat*'). We predicted facilitated retrieval (faster reading times) for the Many-Cue condition at the verb region ('*had voted for*'), where readers could understand that '*The senator*' is the object of the verb. As predicted, this pattern was observed at the retrieval region and continued throughout the rest of the sentence. Participants also completed the Author/Magazine Recognition Tests (ART/MRT; Stanovich and West, 1989), providing a proxy for world knowledge. Since higher ART/MRT scores may index (a) greater experience accessing relevant knowledge and/or (b) richer/more highly structured representations in semantic memory, we predicted it would be positively associated with effects of elaboration on retrieval. We did not observe the predicted interaction between ART/MRT scores and Cue condition at the retrieval region, though ART/MRT interacted with Cue condition in other locations in the sentence. In sum, we found that providing more elaborative information over the course of a text can facilitate retrieval for referents, consistent with a framework in which referential elaboration over a discourse and not just local linguistic information directly impacts information retrieval during sentence processing.

Keywords: sentence processing, retrieval, elaboration, representational complexity, semantic memory, self-paced reading

OPEN ACCESS

Edited by:
Matthew Wagers,
University of California, Santa Cruz, USA

Reviewed by:
Randi Martin,
Rice University, USA
Jeffrey Witzel,
University of Texas at Arlington, USA

****Correspondence:***
Melissa Troyer
mtroyer@ucsd.edu

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: *28 August 2015*
Accepted: *01 March 2016*
Published: *15 March 2016*

Citation:
Troyer M, Hofmeister P and Kutas M (2016) Elaboration over a Discourse Facilitates Retrieval in Sentence Processing. Front. Psychol. 7:374. doi: 10.3389/fpsyg.2016.00374

INTRODUCTION

Real-world knowledge is activated rapidly and richly in language comprehension (e.g., Kutas and Federmeier, 2000; DeLong et al., 2005; Metusalem et al., 2012). Knowledge about events, actions, and entities in the world can rapidly affect people's expectations about upcoming linguistic information (e.g., Kamide et al., 2003; DeLong et al., 2005; Borovsky et al., 2012). What's more, real-world knowledge use during language comprehension is dynamic, and new information can update, amend, or contradict prior information.

The ability to access this continually updated information depends on a number of factors, including the linguistic context. For instance, Bransford and Johnson (1972) provided participants with labeled and unlabeled versions of prose passages. One passage described an activity in which people typically arrange things into groups, go to the appropriate facilities, and perform a routine where a mistake may be rather expensive. Participants who initially received a label (e.g., *washing clothes*) had better memory for the passages. Similar effects have been observed when people are asked to remember information that has been causally linked [e.g., (1) someone needing change because (2) they need to do their laundry] compared to unrelated information (Smith et al., 1978; see also Bradshaw and Anderson, 1982). These findings, among others, demonstrate how language comprehension is fundamentally linked to the supporting knowledge structures, or schema, that are available to the comprehender (Radvansky and Zacks, 1991).

In addition to affecting offline processes like explicit memory, the availability of related linguistic information in a sentence (e.g., the number of adjectives modifying a noun) appears to affect online sentence processing (Hofmeister, 2011; Hofmeister and Vasishth, 2014). Modifying a referent's description with a likely attribute description (e.g., *a ruthless dictator*) leads to faster reading times at words that trigger retrieval of this discourse referent, compared to a referring expression with no modifiers. However, modification with attributes that are unlikely based on real-world knowledge (e.g., *a lovable dictator*) does not lead to the same facilitation, compared to the baseline condition (Hofmeister, 2011). In short, re-accessing previously encoded content appears to be influenced by the ability to access and use prior world knowledge in both online and offline language tasks.

Here, we test whether providing more (vs. less) information about referents across a discourse similarly can increase the ease of language comprehension when these referents are subsequently referred to. In previous work on the role of elaboration in sentence processing (Hofmeister, 2011; Hofmeister and Vasishth, 2014), the syntactic constructions used to investigate elaboration and retrieval were limited to pre-nominal modification and filler-gap dependencies that linked elements within a sentence. A natural question is whether the effects observed in such environments are specific to that particular combination of encoding and retrieval conditions, or whether elaboration can facilitate online language comprehension more generally. This work therefore examines the generality of conceptual elaboration effects in language processing.

Given variability in knowledge due to individual experience, it is likely that individuals also differ from one another in their ability to access and use any particular knowledge structure. If the performance profiles described above depend significantly on the availability of existing knowledge structures, then individual profiles ought to vary as a function of their experience accessing relevant knowledge or the availability of richer or highly structured representations in memory. Before outlining the current experiment, we briefly describe work underscoring the importance of world knowledge for guiding online language comprehension.

When understanding sentences, people seem to anticipate upcoming information based on the relationship between current linguistic information and prior world knowledge (e.g., Tanenhaus et al., 1995; Kamide et al., 2003; Borovsky et al., 2012). For instance, if a listener hears '*The pirate chases the. . .*,' it is reasonable for her to expect that the sentence will continue with something that a pirate (the agent) might chase (the action verb), such as a ship. Visual world eye-tracking paradigms, in which participants listen to spoken language while looking at images of items on a computer screen, have shown that both children and adults are sensitive to this type of information and use it to anticipate upcoming linguistic content (e.g., Kamide et al., 2003; Borovsky et al., 2012, 2013; Troyer and Borovsky, 2015).

In addition to eye-tracking paradigms, event-related brain potential (ERP) experiments support the role of real-world knowledge in guiding language comprehension. For instance, the N400 ERP component, whose amplitude is modulated by the semantic fit of meaningful input with prior context (Kutas and Hillyard, 1980, 1984; Kutas and Federmeier, 2000; see Kutas and Federmeier, 2011, for a recent review), is sensitive not only to fit of (or expectations about) semantic information but also to incoming information as it relates to individuals' real-world knowledge (Hagoort et al., 2004; Nieuwland and Van Berkum, 2006; Hald et al., 2007; Filik and Leuthold, 2013). For instance, Hagoort et al. (2004) presented participants with sentences drawing upon world knowledge, such as the fact that the color of Dutch trains is yellow. They found reduced N400 amplitude to words like '*yellow*' in the sentence '*Dutch trains are yellow and very crowded*' compared to sentences like '*Dutch trains are sour and very crowded*' (where '*sour*' is semantically inconsistent) and '*Dutch trains are white and very crowded*' (where '*white*' is semantically consistent but inconsistent with world knowledge about Dutch trains). These findings support the notion that experienced-based world knowledge (Dutch trains are yellow) affects language comprehension with the same time course as (and possibly via similar mechanisms to) semantic information (trains cannot be sour).

Furthermore, Metusalem et al. (2012) showed that rich information about events in the world is available during language comprehension. In their study, people read short scenarios about events—for example, a football game: '*Jeremy is a great athlete despite being prone to injury. During his last high school football game, he was knocked unconscious twice. That still didn't keep him from scoring the winning {TOUCHDOWN/HELMET/LICENSE} with only seconds remaining.*' Unsurprisingly, N400 amplitude was reduced to

predictable words fitting both with event-related information and with the semantics of the sentence (like '*touchdown*') compared to anomalous words (like '*license*'). Critically, N400 amplitude was intermediate to words which were not plausible continuations of the sentence but which were event-related (e.g., '*helmet*,' which is situationally related to football). These findings suggest that a rich landscape of knowledge can be rapidly activated during language comprehension, likely contributing to the flexibility of language comprehension.

Participants in the Metusalem et al. (2012) study also completed two tasks called the Author and Magazine Recognition Tests (ART and MRT, respectively), which require participants to select the authors and magazines that they recognize from lists containing both real and false examples (Stanovich and West, 1989). These tests provide an estimate of print experience, and the authors suggested that, by proxy, higher performance on the ART/MRT could reflect richer world knowledge. Indeed, performance on the ART/MRT predicts measures of declarative knowledge, including tests of cultural literacy recognition ($rs = 0.53 - 0.72$; West et al., 1993; Stanovich et al., 1995); tests about history and literature knowledge ($rs = 0.59 - 0.62$; Stanovich and Cunningham, 1992); a range of tests about cultural and practical knowledge ($rs = 0.53 - 0.85$, Stanovich and Cunningham, 1993); and, in children, the General Information subtest of the Peabody Individual Achievement Test (using a modified Title Recognition Test for Children; $r = 0.43$; Cunningham and Stanovich, 1991). If prior world knowledge influences access to event-related information, then N400 amplitude might vary with performance on the ART/MRT. The authors found that scoring higher on the ART and MRT was associated with a greater numerical reduction in N400 amplitude for implausible, yet event-related, continuations (e.g., '*helmet*,' in the example above), compared to participants who scored lower on the ART/MRT. However, the authors were unable to draw strong conclusions about the relationship between the N400 and scores on the ART/MRT, partly due to the number of participants ($N = 30$), which is relatively low for examining individual differences.

In combination with prior world knowledge, new information—for example, information encountered in the current discourse—can be exploited rapidly to aid future language processing. For example, Nieuwland and Van Berkum (2006) presented participants with short texts in which they ascribed human-like properties (e.g., the ability to fall in love) to typically inanimate objects (e.g., peanuts). In their experiments, the N400 was sensitive to these newly learned features, suggesting that people easily updated their mental models of the discourse to include these properties.

The current work investigates how variability in the amount of recently encountered information, providing elaboration of a referent, affects subsequent access. This work extends recent findings from self-paced reading studies that suggest that longer or more semantically complex linguistic representations of referents can facilitate subsequent access to those referents (Hofmeister, 2011; Hofmeister and Vasishth, 2014). For instance, Hofmeister (2011) asked participants to read (word-by-word) sentences in which a critical noun was described by zero, one, or two adjectives (low, mid, and high complexity conditions, respectively). Participants might read, '*It was a [famous (deaf)]* ***sculptor*** *that the aristocrats at the gallery ridiculed during the exclusive art show*.' At a subsequent critical verb (e.g., '*ridiculed*'), the critical noun had to be understood as the grammatical object of the verb. In order to access this information, participants must somehow retrieve information about the initial noun (e.g., '*sculptor*'). Hofmeister (2011) reported decreased reading times during (or in some cases, immediately following) the critical verb for items in the highest-complexity condition (i.e., where critical nouns were preceded by two adjectives) compared to the other conditions. In similar experiments, such findings also were observed for nouns which were semantically richer/more specific (e.g., '*soldier*') compared to less rich/less specific (e.g., '*person*'). Hofmeister (2011) interpreted these results as showing that additional semantic (and possibly syntactic) features of a linguistic representation led to facilitated retrieval of the information later in the sentence.

Studies like those of Hofmeister (2011) and Hofmeister and Vasishth (2014) have primarily focused on pre-nominal descriptors ('*Texas cattle rancher*') or differences in the semantic specificity/richness of a single word ('*soldier*' vs. '*person*') but have not explored the roles of other types of descriptions across a discourse. Pre-nominal adjectives are likely to change the processing of an upcoming noun for multiple reasons. First, in an information-theoretic sense, pre-nominal modification can lower the entropy of (or uncertainty about) the upcoming noun. Second, modifiers might be predictive of the noun for other reasons such as semantic relatedness (consider the relationship between the three words '*Texas*,' '*cattle*,' and '*rancher*,' for example). And finally, pre-nominal modification entails a specific type of syntactic relationship between modifiers and the noun, with the entire bundle of linguistic information [modifier(s) + noun] constituting a phrasal unit.

In the current study, we investigate how complex descriptions impact the subsequent retrieval of information about referents in language comprehension across sentence boundaries. We vary the additional linguistic information not in adjectival modifiers directly preceding the noun, but using post-nominal modification across multiple sentences in a short discourse. We predicted that providing higher-complexity descriptions about referents would make it easier for participants to process subsequent language referring to those referents compared to referents with linguistically simpler descriptions. Such a finding would indicate that conceptual complexity, above and beyond the phrasal unit, can influence retrieval in real-time language comprehension.

We also asked participants to complete a simple test designed to assess print exposure, which has been used as a proxy for real-world knowledge (e.g., Metusalem et al., 2012). We predicted that participants with greater world knowledge would be able to more effectively make use of additional information—possibly due to richer networks of conceptual representations and/or more effective access to relevant conceptual information. We therefore predicted these participants would be more likely to show effects of linguistic complexity at subsequent retrieval sites.

MATERIALS AND METHODS

Participants

A total of 101 participants, ages 18–29 ($M = 20.7$, 77 women) took part in the experiment. Participants were excluded from analysis if their overall accuracy on comprehension questions was less than 70%. This resulted in the exclusion of nine participants, for a total of 92 participants in the final dataset. Participants were students at UCSD who reported that they were native English speakers. They received partial class credit for participation. All participants provided informed consent for the study, which was approved by the University of California, San Diego Institutional Review Board.

Design and Materials

The materials for the study were 24 experimental items and 36 filler items of similar length and syntactic complexity. The majority of our materials were created by modifying materials from Fedorenko et al. (2012). A full listing of the experimental and filler items can be found in the Appendix in the Supplementary Data Sheet. Each item consisted of a short text of three sentences. All items began with two sentences, which were presented and read (self-paced) as whole sentences. The third sentence was presented word-by-word, using a moving-window self-paced reading paradigm (Just et al., 1982). Filler items were constructed to be similar to experimental items in length and content.

For experimental items, the first sentence always introduced four individuals, two of whom were referred to using the same noun (e.g., '*senator*,' in the example below). The second sentence always described relationships between the first two individuals (e.g., the two senators) and the second two (e.g., the Democrat and the Republican), with one of the first two individuals being described in more detail more than the other. In the third and final sentence, the second noun was varied to unambiguously pick out a referent for its object. In the example below, for instance, '*The senator who the Republican had voted for*' would refer to the senator from Ohio who was running for president (the Many-Cue condition), while '*The senator who the Democrat had voted for*' would refer to the other senator (the One-Cue condition).

(1) Sentence 1: Two senators were arguing with a Democrat and a Republican after a big debate.
Sentence 2: The Democrat had voted for one of the senators, and the Republican had voted for the other, a man from Ohio who was running for president.
Sentence 3: The senator who the {Republican/Democrat} had voted for was picking a fight about health care reform.

As described above, Cue condition refers to the presence or absence of additional descriptive information in the second sentence. To mitigate any effect of recency of information on reading times, we also created a second version of the materials in which the Many-Cue item came earlier than the One-Cue item. For example, in the second version of the example shown in (1), the second sentence would read, '*The Democrat had voted for one of the senators, a man from Ohio who was running for president, and the Republican had voted for the other*.' The factor Mention Order refers to whether the critical item (i.e., the object of the relative clause in Sentence 3) was mentioned relatively early or relatively late in the second sentence. In the example above (1), the information is Early for the One-Cue condition (i.e., '*The Democrat had voted for one of the senators*') but Late for the Many-Cue condition (i.e., '*The Republican had voted for one of the senators*'). The design was therefore a 2 × 2: Cue condition (Many-Cue, One-Cue) and Mention Order (Early, Late). This resulted in four lists, randomized across participants according to a Latin-square design such that no participant saw the same exact order of experimental and filler items.

Finally, each text was followed by a comprehension question, which participants answered with *yes* or *no* by key press. Across the experiment, comprehension questions queried each of the three sentences in a text so that a third focused on Sentence 1, a third on Sentence 2, and a third on Sentence 3. Half of the sentences were answered correctly with *no* and half with *yes*. For the example above in (1), the comprehension question asked about the first sentence and was correctly answered with *yes*: *Were the senators arguing before a big debate?* Similarly, filler questions asked about either the first, second, or third sentence, in equal proportions. Half of each set were correctly answered with yes, and half with no.

Author and Magazine Recognition Tests

Prior to testing, participants also completed an updated version of the ART and the MRT (Stanovich and West, 1989). These tasks were designed to provide a simple yet powerful way to estimate print experience and, by proxy, world knowledge. Previous work has found correlations in the range of r = 0.5 – 0.8 between ART/MRT and many measures of declarative/cultural knowledge (Cunningham and Stanovich, 1991; Stanovich and Cunningham, 1992, 1993; West et al., 1993; Stanovich et al., 1995); in addition, both tests correlate (rs = 0.3 – 0.4) with measures of reading comprehension, and the ART also correlates with measures of orthographic and phonological processing (Stanovich and West, 1989). Participants were given a printed list of 80 potential author names (ART) and 80 potential magazine titles (MRT; presented separately) and were asked to put a check mark next to the ones they knew to be true authors/magazines. In actuality, only half were real authors/magazines. Participants were asked to avoid guessing because some of the names on the lists were not actual authors/magazines. Scores for these tasks were calculated by summing the number of hits (correct items checked) minus the number of false alarms (checked items which were incorrect). The scores for both tasks were computed separately but combined (summed) for analyses.

Procedure

We used Linger (version 2.88) by Doug Rohde to collect self-paced reading data. For this part of the experiment, participants were instructed that they would be reading short texts made up of three sentences and that they should read the sentences for content, as there would be comprehension questions following each text. They were provided with examples and familiarized

with the task before they began, including practice on two items very similar to those used in the study, preceded by a few simpler examples of word-by-word self-paced reading.

Accuracy was computed on the fly and in aggregate in subsequent analyses. If participants responded incorrectly, a warning flashed on the screen to encourage them to try harder to answer correctly on subsequent questions. Participants were given a break halfway through the experiment and instructed to take short breaks as needed in between items.

Following testing, participants completed an exit questionnaire including questions about the ease of the experiment. The experiment was typically completed in under an hour.

Analysis

Although the final sentence of each text was presented word by word, five regions were created, the last four of which were analyzed (an example is demarcated below). Region 1 always consisted of a noun phrase (two words); Region 2 was the start of the relative clause (three words); Region 3 was the verb phrase of the relative clause (1–3 words); Region 4 was the matrix verb phrase region (2–5 words); and Region 5 was a final region including direct objects, adverbials, or prepositional phrases (2–7 words).

(2) The senator/who the Republican/had voted for/was picking a fight/about health care reform.

For the primary analyses, we first identified any trial containing single-word responses that were less than 100 ms or greater than 5000 ms and removed these trials from subsequent analysis, affecting less than 1% of the data. Next, for each trial, RTs for words within a region were averaged. These averaged RTs were then log-transformed, and data points falling more or less than 2.5 SDs from the mean (by region and condition) were eliminated, affecting ∼2.5% of the data.

Statistical analyses used linear mixed-effects models (Baayen, 2008) incorporating random effects for both items and subjects as well as fixed effects of Cue condition, Mention Order, and Spillover (log RT of the preceding region) as fixed effects, unless otherwise indicated. In addition, we included by-subjects and by-items random slopes for Cue condition, as this was our primary independent variable of interest. All analyses were performed in the statistical programming environment R.

RESULTS

Self-Paced Reading

Mean log reading times by region are shown in **Figure 1**, and full model estimates and statistics are provided in **Table 1**.

At the second region (which is the point at which the noun phrase '*The senator*' begins to be disambiguated), we observed no main effect of Cue condition or Mention Order, but there was a significant interaction of the two ($\beta = -0.011$, $SE = 0.005$, $t = -2.055$, $p < 0.05$). Visual inspection revealed this interaction appeared to be driven by slower reading times for conditions from Version 1 (Many-Late, One-Early) compared to Version 2 (Many-Early, One-Late; see above for an example of Version 1 vs. Version 2 of the materials). A follow-up analysis with Version (V1, V2) as fixed effects and Subject and Item as random effects indicated this was the case, with a significant difference between the two ($\beta = -0.011$, $SE = 0.005$, $t = -2.04$, $p < 0.05$).

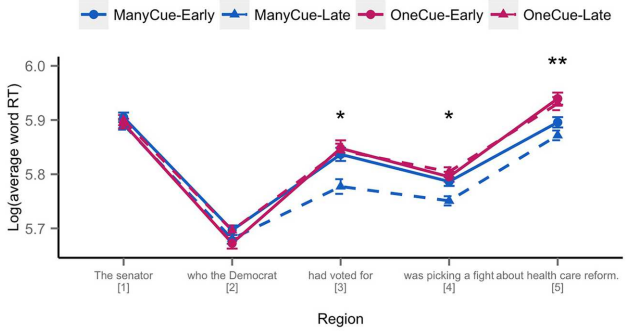


FIGURE 1 | Log average word reading times by region for sentence 3. Errors bars represent by-subject standard errors of the mean. There was a main effect of Cue condition at Regions 3–5 (Many-Cue > One-Cue; $^*p < 0.05$; $^{**}p < 0.001$). See **Table 1** for full model statistics.

Region 3 was the retrieval region where we predicted a main effect of Cue condition. Here, we observed the predicted main effect of Cue condition, with faster reading times in the Many-Cue compared to the One-Cue condition ($\beta = 0.019$, $SE = 0.008$, $t = 2.394$, $p < 0.05$). In addition, we also observed a marginal effect of Mention Order, with relatively Late information leading to faster reading times compared to Early information ($p = 0.07$) as well as a marginal interaction of Cue and Mention Order ($p = 0.09$).

The effect of Cue condition persisted into both Regions 4 ($\beta = 0.016$, $SE = 0.006$, $t = 2.632$, $p < 0.05$) and 5 ($\beta = 0.026$, $SE = 0.006$, $t = 4.074$, $p < 0.001$). No significant main effects or interactions with Mention Order were observed in either region, though there was a marginal interaction between Cue and Order in Region 4 ($p = 0.05$).

ART/MRT Scores

Scores on the ART and MRT were calculated separately and then summed to create a single composite score. For the ART, scores ranged from −5 (one participant checked more incorrect items than correct items, leading to the negative score) to 25, with a mean of 7.28 ($SD = 3.87$). Scores for the MRT ranged from 1 to 20, with a mean of 7.97 ($SD = 3.83$). The two tasks were positively correlated ($r = 0.415$, $p < 0.0001$). When combined by summation, the mean composite score was 15.25 ($SD = 6.47$).

Comprehension Question Accuracies

Comprehension questions were included primarily to encourage participants to read the texts carefully. Comprehension question accuracy was 88.32% ($SD = 6.14\%$) for filler materials. Analyses using mixed-effects logistic regression (with Cue condition and Mention Order as fixed effects and Subject and Item as random

TABLE 1 | Full model estimates and statistics for reading times from the final sentence.

Region	Effect	Estimate	Standard Error	*t*-value	*p*-value
Region 2	**(Intercept)**	**5.693**	**0.023**	**247.15**	**0.000**
	Cue condition	0.000	0.007	−0.059	0.953
	Mention Order	−0.003	0.005	−0.656	0.512
	Cue × Order	**−0.011**	**0.005**	**−2.055**	**0.040**
Region 3	**(Intercept)**	**5.834**	**0.026**	**222.01**	**0.000**
	Cue condition	**0.019**	**0.008**	**2.394**	**0.025**
	Mention Order	0.014	0.008	1.794	0.073
	Cue × Order	−0.013	0.008	−1.680	0.093
Region 4	**(Intercept)**	**5.786**	**0.022**	**264.604**	**0.000**
	Cue condition	**0.016**	**0.006**	**2.632**	**0.015**
	Mention Order	0.005	0.005	0.899	0.369
	Cue × Order	−0.011	0.005	−1.953	0.051
Region 5	**(Intercept)**	**5.916**	**0.025**	**238.968**	**0.000**
	Cue condition	**0.026**	**0.006**	**4.074**	**0.000**
	Mention Order	0.008	0.005	1.639	0.101
	Cue × Order	−0.003	0.005	−0.544	0.586

Statistically significant predictors ($p < 0.05$) are in bold.

effects) revealed that accuracy did not differ as a function of Cue condition or Mention Order, with a mean of 79.35% ($SD = 14.80\%$) for the Many-Cue condition and a mean of 77.26% ($SD = 13.82\%$) for the One-Cue condition. We therefore observed that our manipulation of interest, Cue condition, had no measurable effect on offline comprehension accuracies.

Accuracies were also analyzed by the type of question, that is, whether the question asked about the first, second, or third sentence. Mixed-effects logistic regression with question type (first, second, third sentence) as a fixed effect and Subjects and Items as random effects revealed that questions about the second sentence ($M = 70.92\%$, $SD = 20.89\%$) were answered less accurately than questions about the final sentence ($M = 84.51\%$, $SD = 13.54\%$; $\beta = -0.46$, $SE = 0.17$, $z = -2.75$, $p < 0.01$), though the difference between questions about the first sentence ($M = 79.48\%$, $SD = 14.30\%$) and second sentence did not reach significance ($p = 0.14$). This pattern likely reflects the fact that the second sentence was the most complex/longest of the three sentences.

Relationship between Reading Times and ART/MRT

We predicted that individuals scoring higher on the ART/MRT, and who are therefore likely to have greater world knowledge, would show the greatest effects of Cue condition during the retrieval region. However, adding the continuous ART/MRT composite scores as a predictor did not indicate any effect of ART/MRT on reading times during Region 3 nor was there any interaction with Cue or Mention Order (all $ps > 0.16$).

However, ART/MRT scores interacted with Cue condition at an un-predicted location, in Region 2 ($\beta = -0.002$, $SE = 0.001$, $t = -2.247$, $p < 0.05$). To follow up on this interaction, we used both group comparisons based on a median split as well as a correlational analyses. Numerically, individuals scoring higher on the ART/MRT had faster reading times for the One- ($M = 5.66$ log ms, $SD = 0.31$) compared to the Many-Cue condition ($M = 5.69$ log ms, $SD = 0.33$), but individuals scoring lower on the ART/MRT had the opposite numeric pattern (One-Cue, $M = 5.72$ log ms, $SD = 0.31$; Many-Cue, $M = 5.70$, $SD = 0.31$). Mixed-effects models performed separately over each group with Cue as a fixed effect and subject and item as random effects indicated that these were only trends ($ps = 0.09$, 0.11, respectively). However, a correlational analysis of ART/MRT scores and differences between One-Cue minus Many-Cue RTs was significant, $r = -0.216$, $p < 0.05$. We had no specific predictions for any effect of Cue at this region nor any interactions with ART/MRT (but see Discussion).

In addition, ART/MRT scores interacted with Cue condition in Region 4 ($\beta = -0.002$, $SE = 0.001$, $t = -2.172$, $p < 0.05$). We again inspected both group differences and correlations between ART/MRT and reading time differences. For the higher-scoring group, there was little difference based on Cue condition (One-Cue, $M = 5.80$ log ms, $SD = 0.33$; Many-Cue, $M = 5.79$ log ms, $SD = 0.34$; difference n.s.). However, a mixed-effects model (see above) revealed a difference between the One-Cue ($M = 5.81$ log ms, $SD = 0.32$) and Many-Cue ($M = 5.76$, log ms, $SD = 0.29$) conditions for the group scoring lower on the ART/MRT ($\beta = 0.027$, $SE = 0.008$, $t = 3.537$, $p < 0.001$). The correlation between ART/MRT scores and differences between One-Cue minus Many-Cue RTs was significant ($r = -0.283$, $p < 0.01$), indicating that lower scores were associated with larger differences between conditions. Although this pattern occurred at Region 4, a region subsequent to the critical retrieval region in our experiment (Region 3), it is possible the interaction between ART/MRT and Cue condition at this region relates to continued retrieval processes. We further discuss this possibility in the discussion.

There were no other interactions with ART/MRT at any other region in this analysis.

DISCUSSION

Summary of Findings

This study had two primary aims. The first was to test whether a greater amount of linguistic elaboration about a referent over a short discourse could facilitate subsequent access to that information during online language processing. If so, the second was to test whether this facilitation was greater for those with more world knowledge (determined using scores from the ART and MRT as a proxy) would lead to increased facilitation based on elaboration.

Supporting our hypothesis that elaborative information would provide more cues to retrieval, we found reduced reading times at a critical retrieval site when the referent had previously been described in more detail, albeit not more so for those with greater world knowledge. This work provides a novel contribution by suggesting that elaboration can affect retrieval-related processes in cross-sentential dependencies. These findings demonstrate the generality of elaboration effects in sentence processing (Hofmeister, 2011; Hofmeister and Vasishth, 2014).

It is particularly noteworthy that various formal syntactic theories treat anaphoric dependencies as fundamentally different than filler-gap dependencies. For instance, in transformational theories of syntax, filler-gap dependencies are licensed via cyclic movement of the filler, leaving behind a trace, whereas no such process applies to anaphoric dependencies (co-indexing provides the necessary connection; e.g., Chomsky, 1995, among many others). More importantly, the retrieval conditions in filler-gap dependencies are quite different from those in the current study. In filler-gap dependencies, the retrieval target is necessarily within the same sentence, which may limit the retrieval search space, relative to that for anaphoric dependencies. Further, the onset of a filler-gap dependency signals that the target information must be restored in the near future. That is, once a filler is encountered, a process is initiated that necessarily ends with retrieval; hence, it is predictable that the filler information will be needed again. Up to that point, the parser is actively engaged in searching for the first available integration point (Clifton and Frazier, 1989; Frazier and Clifton, 1989; Frazier and d'Arcais, 1989). This contrasts with anaphoric dependencies where there is no guarantee that a referent will ever be mentioned again—as was the case for the elaborative information presented in our short texts. In sum, anaphoric dependencies do not come with the same set of expectations or retrieval cues that accompany filler-gap dependencies. Thus, demonstrating that elaboration effects nevertheless arise in cross-sentential dependencies suggests that they are not contingent upon any of the idiosyncrasies of filler-gap dependencies.

We did not observe the predicted interaction between ART/MRT and Cue condition at Region 3. However, two unpredicted related results were the interactions between ART/MRT scores and Cue condition on reading times at Regions 2 and 4. In Region 2 ('The senator/who the Democrat/. . .'), participants may begin to anticipate the upcoming object of the relative clause, though there is still ambiguity with respect to which referent will be mentioned. We tentatively speculate that differences in language experience/world knowledge (as indexed by ART/MRT scores) may affect the individual's sensitivity to this ambiguity (or ability to predict an upcoming referent), possibly resulting in the observed interaction.

We initially hypothesized that having greater world knowledge (and higher scores on the ART/MRT, by proxy), would associate with greater ease of access for meaningful cues to retrieval. We therefore predicted greater facilitation in retrieval (at Region 3) for the Many-Cue condition, or possibly in a subsequent region, for those with greater world knowledge. However, the interaction between Cue and ART/MRT scores which we observed at Region 4 did not support our hypothesis; rather, individuals with lower ART/MRT scores drove effects of Cue condition in this region, with lower reading times associated with the Many-Cue compared to the One-Cue condition. One possibility is that for our materials, having more information benefited those with less language experience/less knowledge more, meaning that the group scoring lower on ART/MRT was able to benefit from the additional information in the Many-Cue condition while the higher-scoring group showed less of a difference between conditions. Future work using more tightly controlled stimuli (e.g., with identical numbers of words in each region, with identical syntax, etc.) might shed more light on the nature of these individual differences.

Overall, we interpret our findings as evidence that having more information about a referent is beneficial during retrieval and perhaps during subsequent comprehension, as the sentence progresses and information accumulates.

The Role of Elaboration in Online Sentence Processing

Work by Hofmeister (2011) and Hofmeister and Vasishth (2014) has shown that under many circumstances, elaborative information, typically in the form of adjectives preceding a noun, increases processing times at the point of encoding (at the noun) but facilitates processing times at a subsequent dependency. This finding holds for words which are more elaborated in the sense that they are semantically richer (e.g., '*soldier*' is richer than '*person*'), but it does not hold when adjectives preceding a noun are atypical descriptors (e.g., '*ruthless military dictator*' is typical but '*lovable military dictator*' is not). Here, we add to this literature by showing that elaborative information presented across multiple sentences, and not just locally (at the point of modifying a noun, for example), can facilitate subsequent access to or retrieval of that information.

What may account for the benefit of retrieving representations that have relatively many features associated with them, even across discourse boundaries? On one hand, such effects are surprising since it would seem to imply that more content must be retrieved. On the other, these effects align naturally with several non-mutually exclusive hypotheses about the nature of memory retrieval in language processing. For instance, in the cue-based retrieval model of Lewis and Vasishth (2005), the efficacy of retrieval for some item in memory is driven partly by its retrieval history, i.e., how many times an item has been restored and how recently. Modifying a word or phrase that has been encoded in the past reactivates that item, leading to an increase in its

activation. This reactivation process can even arguably offset any effects of time-based decay, giving rise to so-called anti-locality effects (Vasishth and Lewis, 2006). From this point of view, the increased ease of retrieval observed in Regions 3–5 is ascribable to a boosted level of activation of the target either prior to retrieval, or possibly during retrieval, as relevant cues spread activation to other cues (see Hofmeister, 2011). A separate, though not mutually exclusive, view suggests that adding semantic features to a discourse referent typically gives rise to a conceptually unique representation in the current discourse context. The advantage of this elaboration is manifested at the retrieval region, as the broader memory literature demonstrates a robust memory advantage for targets with contextually unique features (Moscovitch and Craik, 1976; Fisher and Craik, 1977; Jacoby and Craik, 1979; Hunt and Worthen, 2006; Gallo et al., 2008). In essence, adding details about a person or event increases the likelihood that this entity bears conceptual features that no other memory item (or very few others) shares, reducing the chance for similarity-based interference at retrieval. Both of these views capture the observed effects in our experiment without adjudicating between them.

CONCLUSION

The present findings are novel in showing that when (potentially) relevant semantic information is associated with a concept, it may directly impact its retrieval, even when the elaborative information is distributed across a discourse, and not just or at all in the local (within-sentence) linguistic context (as in Hofmeister, 2011; Hofmeister and Vasishth, 2014). Relatedly, one recent study found that when participants read longer descriptions (e.g., '*The actor who was frustrated and visibly upset*' vs. '*The actress*'), they were more likely to refer back to them with a pronoun, a finding the authors attributed to enhanced prominence of the referent due to the elaboration (Karimi et al., 2014). When concepts are more elaborated, subsequent processing advantages may occur because (a) there are more semantic features available and/or (b) those features lead to increased activation levels of the concept. Our findings suggest that variability in the elaboration of referents may have relatively long-term consequences for their processing across the subsequent discourse.

AUTHOR CONTRIBUTIONS

All authors listed, have made substantial, direct and intellectual contribution to the work, and approved it for publication.

ACKNOWLEDGMENTS

This work was supported by NICHD grant 22614 to MK. MT was supported by an NSF Graduate Research Fellowship and the Kroner Fellowship at UCSD. We would like to thank Joshua Davis, Katherine DeLong, Jeff Elman, Robert Kluender, Kevin Smith, and Tom Urbach for comments on previous versions of this work.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpsyg.2016.00374

REFERENCES

Baayen, R. H. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge: Cambridge University Press.

Borovsky, A., Burns, E., Elman, J. L., and Evans, J. L. (2013). Lexical activation during sentence comprehension in adolescents with a history of Specific Language Impairment. *J. Commun. Disord.* 46, 413–427. doi: 10.1016/j.jecp.2012.01.005

Borovsky, A., Elman, J. L., and Fernald, A. (2012). Knowing a lot for one's age: vocabulary skill and not age is associated with anticipatory incremental sentence interpretation in children and adults. *J. Exp. Child Psychol.* 112, 417–436. doi: 10.1016/j.jecp.2012.01.005

Bradshaw, G. L., and Anderson, J. R. (1982). Elaborative encoding as an explanation of levels of processing. *J. Verbal Learn. Verbal Behav.* 21, 165–174. doi: 10.1016/S0022-5371(82)90531-X

Bransford, J. D., and Johnson, M. K. (1972). Contextual prerequisites for understanding: some investigations of comprehension and recall. *J. Verbal Learn. Verbal Behav.* 11, 717–726. doi: 10.1016/S0022-5371(72)80006-9

Chomsky, N. (1995). *The Minimalist Program*. Cambridge, MA: MIT Press.

Clifton, C. Jr., and Frazier, L. (1989). "Comprehending sentences with long-distance dependencies," in *Linguistic Structure in Language Processing*, eds G. N. Carlson and M. K. Tanenhaus (Washington, DC: Linguistic Society of America), 273–317.

Cunningham, A. E., and Stanovich, K. E. (1991). Tracking the unique effects of print exposure in children: associations with vocabulary, general knowledge, and spelling. *J. Educ. Psychol.* 83, 264–274. doi: 10.1037/0022-0663.83.2.264

DeLong, K. A., Urbach, T. P., and Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nat. Neurosci.* 8, 1117–1121. doi: 10.1038/nn1504

Fedorenko, E., Piantadosi, S., and Gibson, E. (2012). Processing relative clauses in supportive contexts. *Cogn. Sci.* 36, 471–497. doi: 10.1111/j.1551-6709.2011.01217.x

Filik, R., and Leuthold, H. (2013). The role of character-based knowledge in online narrative comprehension: evidence from eye movements and ERPs. *Brain Res.* 1506, 94–104. doi: 10.1016/j.brainres.2013.02.017

Fisher, R., and Craik, F. (1977). Interaction between encoding and retrieval operations in cued recall. *Hum. Learn. Mem.* 3, 701–711. doi: 10.1037/0278-7393.3.6.701

Frazier, L., and Clifton, C. Jr. (1989). Successive cyclicity in the grammar and the parser. *Lang. Cogn. Process.* 4, 93–126. doi: 10.1080/01690968908406359

Frazier, L., and d'Arcais, G. B. F. (1989). Filler driven parsing: a study of gap filling in Dutch. *J. Mem. Lang.* 28, 331–344. doi: 10.1016/0749-596x(89)90037-5

Gallo, D. A., Meadow, N. G., Johnson, E. L., and Foster, K. T. (2008). Deep levels of processing elicit a distinctiveness heuristic: evidence from the criterial recollection task. *J. Mem. Lang.* 58, 1095–1111. doi: 10.1016/j.jml.2007.12.001

Hagoort, P., Hald, L., Bastiaansen, M., and Petersson, K. M. (2004). Integration of word meaning and world knowledge in language comprehension. *Science* 304, 438–441. doi: 10.1126/science.1095455

Hald, L. A., Steenback-Planting, E. G., and Hagoort, P. (2007). The interaction of discourse context and world knowledge in online sentence comprehension. Evidence from the N400. *Brain Res.* 1146, 210–218. doi: 10.1016/j.brainres.2007.02.054

Hofmeister, P. (2011). Representational complexity and memory retrieval in language comprehension. *Lang. Cogn. Process.* 26, 376–405. doi: 10.1080/01690965.2010.492642

Hofmeister, P., and Vasishth, S. (2014). Distinctiveness and encoding effects in online sentence comprehension. *Front. Psychol.* 5:1237. doi: 10.3389/fpsyg.2014.01237

Hunt, R., and Worthen, J. (2006). *Distinctiveness and Memory*. Oxford: Oxford University Press.

Jacoby, L., and Craik, F. (1979). "Effects of elaboration of processing at encoding and retrieval: trace distinctiveness and recovery of initial context," in *Levels of Processing in Human Memory*, eds L. Cermak and F. Craiks (Hillsdale, NJ: Lawrence Erlbaum Associates), 1–21.

Just, M. A., Carpenter, P. A., and Woolley, J. D. (1982). Paradigms and processes in reading comprehension. *J. Exp. Psychol.* 111, 228–238. doi: 10.1037/0096-3445.111.2.228

Kamide, Y., Altmann, G. T. M., and Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: evidence from anticipatory eye movements. *J. Mem. Lang.* 49, 133–156. doi: 10.1016/S0749-596X(03)00023-8

Karimi, H., Fukumura, K., Ferreira, F., and Pickering, M. J. (2014). The effect of noun phrase length on the form of referring expressions. *Mem. Cogn.* 42, 993–999. doi: 10.3758/s13421-014-0400-7

Kutas, M., and Federmeier, K. D. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends Cogn. Sci.* 4, 463–470. doi: 10.1016/S1364-6613(00)01560-6

Kutas, M., and Federmeier, K. D. (2011). Thirty years and counting: finding meaning in the N400 component of the event-related brain potential. *Annu. Rev. Psychol.* 62, 621–647. doi: 10.1146/annurev.psych.093008.131123

Kutas, M., and Hillyard, S. A. (1980). Reading senseless sentences: brain potentials reflect semantic incongruity. *Science* 207, 203–205. doi: 10.1126/science.7350657

Kutas, M., and Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature* 307, 161–163. doi: 10.1038/307161a0

Lewis, R. L., and Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cogn. Sci.* 29, 375–419. doi: 10.1207/s15516709cog0000_25

Metusalem, R., Kutas, M., Urbach, T. P., Hare, M., McRae, K., and Elman, J. L. (2012). Generalized event knowledge activation during online sentence comprehension. *J. Mem. Lang.* 66, 545–567. doi: 10.1016/j.jml.2012.01.001

Moscovitch, M., and Craik, F. (1976). Depth of processing, retrieval cues, and uniqueness of encoding as factors in recall. *J. Verbal Learn. Verbal Behav.* 15, 447–458. doi: 10.1016/S0022-5371(76)90040-2

Nieuwland, M. S., and Van Berkum, J. J. A. (2006). When peanuts fall in love: N400 evidence for the power of discourse. *J. Cogn. Neurosci.* 18, 1098–1111. doi: 10.1162/jocn.2006.18.7.1098

Radvansky, G. A., and Zacks, R. T. (1991). Mental models and the fan effect. *J. Exp. Psychol.* 17, 940–953. doi: 10.1037//0278-7393.17.5.940

Smith, E. E., Adams, N., and Schorr, D. (1978). Fact retrieval and the paradox of interference. *Cogn. Psychol.* 10, 438–464. doi: 10.1016/0010-0285(78)90007-5

Stanovich, K. E., and Cunningham, A. E. (1992). Studying the consequences of literacy within a literate society: the cognitive correlates of print exposure. *Mem. Cogn.* 20, 51–68. doi: 10.3758/bf03208254

Stanovich, K. E., and Cunningham, A. E. (1993). Where does knowledge come from? Specific associations between print exposure and information acquisition. *J. Educ. Psychol.* 85, 211–229. doi: 10.1037/0022-0663.85.2.211

Stanovich, K. E., and West, R. F. (1989). Exposure to print and orthographic processing. *Read. Res. Q.* 24, 402–433. doi: 10.1037/0022-0663.85.2.211

Stanovich, K. E., West, R. F., and Harrison, M. R. (1995). Knowledge growth and maintenance across the life span: The role of print exposure. *Dev. Psychol.* 31, 811–826. doi: 10.1037/0012-1649.31.5.811

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., and Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science* 268, 1632–1634. doi: 10.1126/science.7777863

Troyer, M., and Borovsky, A. (2015). "Childhood SES affects anticipatory language comprehension in college-age adults," in *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*, eds D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, et al. (Austin, TX: Cognitive Science Society).

Vasishth, S., and Lewis, R. L. (2006). Argument-head distance and processing complexity: explaining both locality and anti-locality effects. *Language* 82, 767–794. doi: 10.1353/lan.2006.0236

West, R. F., Stanovich, K. E., and Mitchell, H. R. (1993). Reading in the real world and its correlates. *Read. Res. Q.* 28, 34–50. doi: 10.2307/747815yy

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2016 Troyer, Hofmeister and Kutas. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 09 April 2015
doi: 10.3389/fpsyg.2015.00347

Representing number in the real-time processing of agreement: self-paced reading evidence from Arabic

Matthew A. Tucker[1]*, *Ali Idrissi*[2] *and Diogo Almeida*[1]

[1] *Language, Mind, and Brain Laboratory, Science Division, Department of Psychology, New York University Abu Dhabi, Abu Dhabi, UAE,* [2] *Department of English Literature and Linguistics, Qatar University, Doha, Qatar*

In the processing of subject-verb agreement, non-subject plural nouns following a singular subject sometimes "attract" the agreement with the verb, despite not being grammatically licensed to do so. This phenomenon generates agreement errors in production and an increased tendency to fail to notice such errors in comprehension, thereby providing a window into the representation of grammatical number in working memory during sentence processing. Research in this topic, however, is primarily done in related languages with similar agreement systems. In order to increase the cross-linguistic coverage of the processing of agreement, we conducted a self-paced reading study in Modern Standard Arabic. We report robust agreement attraction errors in relative clauses, a configuration not particularly conducive to the generation of such errors for all possible lexicalizations. In particular, we examined the speed with which readers retrieve a subject controller for both grammatical and ungrammatical agreeing verbs in sentences where verbs are preceded by two NPs, one of which is a local non-subject NP that can act as a distractor for the successful resolution of subject-verb agreement. Our results suggest that the frequency of errors is modulated by the kind of plural formation strategy used on the attractor noun: nouns which form plurals by suffixation condition high rates of attraction, whereas nouns which form their plurals by internal vowel change (ablaut) generate lower rates of errors and reading-time attraction effects of smaller magnitudes. Furthermore, we show some evidence that these agreement attraction effects are mostly contained in the right tail of reaction time distributions. We also present modeling data in the ACT-R framework which supports a view of these ablauting patterns wherein they are differentially specified for number and evaluate the consequences of possible representations for theories of grammar and parsing.

Keywords: working memory, agreement, plurals, abstract morphology, self-paced reading, Arabic, sentence processing

OPEN ACCESS

Edited by:
Colin Phillips,
University of Maryland, USA

Reviewed by:
Darren Tanner,
University of Illinois at Urbana-Champaign, USA
Kepa Erdocia,
University of the Basque Country, Spain

***Correspondence:**
Matthew A. Tucker, Language, Mind, and Brain Laboratory, Science Division, Department of Psychology, New York University Abu Dhabi, (A2-166B), P.O. Box 129188, Abu Dhabi, UAE
matt.tucker@nyu.edu

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: *03 November 2014*
Accepted: *11 March 2015*
Published: *09 April 2015*

Citation:
Tucker MA, Idrissi A and Almeida D (2015) Representing number in the real-time processing of agreement: self-paced reading evidence from Arabic. Front. Psychol. 6:347. doi: 10.3389/fpsyg.2015.00347

1. Introduction

A fundamental feature of language comprehension in real time is the online integration of grammatical information in the form of structural cues expressed morphologically on individual lexical items. For instance, many languages display grammatical agreement—a process whereby verbs co-vary in form with features of their arguments. Integrating agreement cues to resolve verb-argument agreement dependencies provides the parser with valuable information concerning

structural relations in the input and therefore provides important clues to the correct parse. Moreover, humans are quite good at completing this resolution: it is conducted relatively quickly, and failures to resolve agreement dependencies result in failures of parsing in many instances.

Despite this relative aptitude in comprehending agreement, speakers do make mistakes in both the comprehension and production of agreement dependencies. Since the initial study of Bock and Miller (1991), a large amount of theorizing concerning the nature of this integration has been based upon failures of agreement called *agreement attraction* errors. In an agreement attraction error, an agreeing element does not correctly match its controller in all features but instead matches a local *distractor* or *attractor* in a subset of the mismatching features. A key property of these errors is that this distractor NP is typically thought to be grammatically inaccessible insofar as it is not normally capable of controlling agreement because of its structural position. For instance, in English subject-verb agreement dependencies, attraction errors have been noted for several configurations, including prepositional phrase modifiers/complements, relative clauses, and the like (1)[1]:

(1) a. *The sheer weight* of **all these figures *make*** them hard to understand.
(based upon Ronald Reagan 13 October 1982; quoted in Francis, 1986 and Wagers et al., 2009)

b. *The boy* that liked **the snakes *sleep*** throughout the afternoon.
(based upon Bock and Miller, 1991)

c. *The request* to begin **the projects *were*** overwhelming because of the cost.
Tucker and Wagers (2010)

d. *Studying* **micro-climates** like this ***have*** helped me to understand if it's gonna be a minor event, or a catastrophic one.
(Jim Wood, professional blog post 13 August 2012[2])

Errors such as these are often discussed by both grammarians and syntactic theorists alike (see Jespersen, 1924; Zandvoort, 1961; Kimball and Aissen, 1971; Quirk et al., 1985; Francis, 1986; Kayne, 2000; Den Dikken, 2001; *inter alia*) and—despite their *prima facie* ungrammaticality—are common in both everyday speech and formal writing. Both production and comprehension studies have shown that the probability of agreement attraction errors is influenced by a large number of factors, including linear order, relative structural embedding, and the amount of featural overlap between distractor and verb (Bock and Cutting, 1992; Bock and Eberhard, 1993; Vigliocco and Nicol, 1998; Pearlmutter et al., 1999; Hartsuiker et al., 2001; Franck et al., 2002; Thornton and MacDonald, 2003; Haskell and MacDonald, 2005; Wagers et al., 2009; *inter alia*).

The factors which have been shown to influence the possibility of these errors include both processing and grammatical constraints. For instance, several researchers beginning with Bock and Miller (1991) have noted that agreement attraction errors are asymmetric in both their occurrence and salience in English. Specifically, whereas errors leading to erroneously plural verbs (2a) are commonly produced and more difficult to notice, erroneously singular verbs (2b) are rarely produced and seem much more salient to speakers:

(2) a. *The key* to **the cabinets *have*** become rusty from years of disuse. (**SG → PL**)

b. *The keys* to **the cabinet *has*** become rusty from years of disuse. (**PL → SG**)

One plausible explanation for this asymmetry is the grammatical notion of MARKEDNESS, wherein one marked value of a feature (in this case, plural) is defined by its opposition to another unmarked value (in this case, singular). By tapping into this grammatical notion, the reason for the particular direction of this asymmetry becomes explainable as attraction to the marked plural case in (2a). In (2b), on the other hand, the presence of an unmarked attractor means the verb is less easily misconstrued.

Similarly, the grammatical notion of a syntactic hierarchy has also been shown to be relevant by both Bock and Cutting (1992) and Franck et al. (2002), among others. In the Franck et al. (2002) study, the authors contrasted preambles such as (3) in a sentence production study to determine whether linear distance or syntactic prominence (defined in terms of structural height in a parse tree) contributes more to attraction. In these preambles, the linearly closest NP is not the structurally most prominent NP computed in terms of structural height:

(3) a. *L'-ordinateur* avec **le programme des expériences** . . .
the-computer with the program of.the experiments . . .
"*The computer* with **the program** of **the experiments** . . ."

b. *L'-ordinateur* avec **les programmes** de **l'expérience** . . .
the-computer with the programs of.the experiment . . .
"*The computer* with **the programs** of **the experiment**. . ."

In (3), the PP containing *expérience(s)* is a complement, and therefore structurally contained within the NP headed by *programme(s)*. The authors observed that the syntactically higher noun [*le(s) programme(s)*] has a larger impact on attraction error rates than the linearly closest noun (*des expériences*) leading the authors to conclude that syntactic hierarchical prominence plays a larger role than linear adjacency in modulating attraction rates.

On the other hand, processing constraints clearly matter, as well. Most concretely, attractions are errors, and only appear in

[1]In these examples and throughout the paper, the grammatical controller is in *italics*, the distractor in **bold**, and the erroneous verb in both *italic* and **bold** typeface. We do not mark ungrammatical sentences with a diacritic in this paper, as the nature of the acceptability of *prima facie* ungrammatical sentences is the object of study here.

[2]http://jimwood8.wordpress.com/2012/08/13/non-adjacent-agreement-attraction/.

a subset of observations for any given language community[3]. Moreover, emerging comprehension literature has shown that attraction errors in comprehension only occur in *ungrammatical* utterances, not grammatical ones (see, e.g., Wagers et al., 2009; Tucker and Wagers, 2010; Tanner et al., 2014). Thus, one does not find the comprehension correlates of attraction in examples such as (4):

(4) *The key* to **the cabinets *has*** become rusty from years of disuse.

Despite the fact that the attractor noun phrase mismatches the subject and is plural, reading times at *has* and error rates in speeded grammaticality studies do not reflect difficulty for the parser. Thus, error rates on examples such as (4) are low, and reading times at *has* do not differ from normal reading times for grammatical verbs. The explanation given for this asymmetry by Wagers et al. (2009) is that the attraction in ungrammatical sentences is the result of the parser's attempt to interpret an obviously erroneous verb by searching working memory for a matching noun phrase. Crucially, Wagers and colleagues contrast this with a view wherein grammatical representations are themselves fallible by observing such a view should apply equally in grammatical and ungrammatical utterances. This parsing strategy therefore provides a superior explanation for the grammaticality asymmetry in attraction than a view more wedded to grammatical representation.

1.1. Representations and Processes

A recently emerging hypothesis concerning the proper interpretation of dependency errors takes them to be a failure of the working memory implementation of agreement dependencies (Badecker and Kuminiak, 2007; Badecker and Lewis, 2007; Wagers et al., 2009; Dillon et al., 2013) following a general hypothesis that at least some of the processes involved in language comprehension are underwritten by a kind of skilled memory retrieval (Lewis and Vasishth, 2005; Lewis et al., 2006). Architecturally and programmatically, viewing dependency resolution as skilled memory retrieval allows the development of explicit hypotheses about the relationship between behavioral results and architectural claims about language comprehension insofar as researchers are forced to be explicit about both representational and procedural commitments.

In comprehension-as-retrieval models, some or all agreement morphology on a lexical item triggers a working memory retrieval event wherein the system attempts to find an available controller in a content-addressable memory. In order to do so, a procedural component searches all available chunks (constituents) in memory in parallel and attempts to locate a match along several cue dimensions, with the winner being decided by which element matches along the most dimensions. When the controller matches the agreeing element in all grammatical features, the number of matching retrieval cues will result in a proper retrieval of the true grammatical controller. However, when the controller and agreeing element do not match in all cues, those mismatched cues which the distractor bears can, in some instances, be sufficient to trigger an erroneous retrieval of the distractor, resulting in an attraction error.

The retrieval hypothesis is well-suited to explain the sensitivity of attraction to mismatches in controller and distractor cues (Bock and Miller, 1991; *et seq.*), the absence of attraction-like illusions of ungrammaticality in grammatical utterances (Wagers et al., 2009), and the relative error proportions in various constructions (Dillon et al., 2013). Finally, recently emerging work suggests that memory models are also, when combined with proper representational specifications, well-suited to explaining differing behavioral profiles for at least some different kinds of grammatical dependencies (Dillon et al., 2013; though see Parker, 2014 for some critical discussion).

Linking agreement attraction errors to more general comprehension models provides for some important possibilities for research into both the grammar and parsing. Specifically, as Dillon et al. (2013) demonstrate, in places where experimental data are suggestive of a particular representational commitment in the parser, modeling can provide additional evidence for this commitment when it dovetails with experimental results. Moreover, the success of memory models in accounting for particular experimental results across a range of languages adds to the validity of the models themselves. In order to do this modeling of experimental results, however, researchers must stake particular claims about the relationship between parsing and grammar in order to decide on representations and processes in the models. Any such claims, therefore, help elucidate the connection between parsing *per se*, grammar, and working memory.

By contrast to the memory models, several alternatives have been proposed which view agreement attraction as either grammaticalized alternatives (Kimball and Aissen, 1971; Kayne, 2000; Den Dikken, 2001) or an improper representation driven by feature movement or percolation of number features to incorrect nodes in syntactic trees (Nicol et al., 1997; Vigliocco and Nicol, 1998; Franck et al., 2002; Eberhard et al., 2005). However, as was first pointed out by Wagers et al. (2009), models which eschew the role of memory are only successful insofar as one can identify correlates of their representational claims in all aspects of processing behavior. Grammaticalization models assume that, at the very least, attraction should be possible outside of error contexts, a finding which has yet to be conclusively demonstrated. As for representational models, memory models have been argued to be superior to purely representational approaches in understanding the comprehension of grammatical sentences which contain the structural configurations supporting the creation of erroneous representations. As Wagers et al. (2009) have argued, erroneous representations should be possible in ultimately grammatical utterances, yet experiments designed to test for the presence of "agreement attraction" in grammatical utterances consistently yield null results (see Wagers et al., 2009; Tanner et al., 2014; and our results below). We therefore conclude, with these authors, that memory models provide a better

[3]See any of the previously cited studies for examples of this. Note that this runs counter to the claims of some formal linguists who have occasionally treated these errors as dialectal or idiolectal variation in need of explanation (i.e., Kimball and Aissen, 1971; Kayne, 2000; Den Dikken, 2001). At the very least, one would have to maintain that the data underwriting these studies are dialectically distinct from standard English, as Kimball and Aissen (1971) do.

avenue for exploration in the service of explaining possibly erroneous dependency processing in natural language and couch the study reported here in memory retrieval terms.

1.2. Crosslinguistic Considerations

What working-memory models require, however, is a well-understood theory of the relationship between formal linguistic features usually referenced in linguistic theories of agreement (such as those proposed by Chomsky, 1995, 2000, 2001; and related work) and the cues used in models of working memory tasks. It is therefore conspicuous that the prevailing views on feature-cue mapping have been developed with a comparatively small sample of languages in mind: the majority of studies have examined either Germanic languages such as Dutch, English, and German or Romance languages such as French, Spanish, and Italian (For English, see any of the previously cited works except Franck et al., 2002, among many others. For Dutch, see Hartsuiker et al., 1999; Meyer and Bock, 1999; Bock et al., 2001; Kaan, 2002; Hartsuiker et al., 2003; for German, see Hartsuiker et al., 2003; Häussler, 2009; for Spanish, see Vigliocco et al., 1996; Antón-Méndez et al., 2002; Franck et al., 2008; Lago et al., 2014; for French, see Fayol et al., 1994; Vigliocco et al., 1995; Vigliocco and Franck, 2001; Franck et al., 2002, 2008, 2010 and for Italian, see Vigliocco et al., 1995; Vigliocco and Franck, 1999, 2001; Franck et al., 2006, 2008). The only exceptions to this tendency involve two studies on the Slavic languages Russian (Lorimor et al., 2008) and Slovak (Badecker and Kuminiak, 2007), however even this sample of languages is wholly contained within the larger Indo-European family. To our knowledge, no studies of agreement attraction exist in languages outside Indo-European. A theory of the relationships connecting grammar, parsing, and working memory is ultimately a theory about the implementation of language in the mind, and therefore would benefit from the largest possible cross-linguistic coverage since it is conceivable that there is crosslinguistic variation here.

This lacuna is additionally striking when one considers the possible range of variation in the expression of verbal agreement. Germanic and Romance languages display subject-verb agreement for grammatical number, and while nominals in these languages have formal gender, this gender does not impact the subject-verb agreement system. This is not true of the Slavic languages studied by Badecker and Kuminiak (2007) and Lorimor et al. (2008), where converging evidence seems to suggest that gender does play a role in attraction. However, in these languages, nominal morphology also includes grammatical case-marking, which is shown to play a confounding role insofar as the case a nominal bears helps to disambiguate its grammatical function (for similar evidence in German, see Häussler, 2009). In these languages, it may be possible to set up an attraction configuration involving gender, but grammatical case on the attractor serves to disambiguate its grammatical role in a way which drives down attraction rates. It is thus important to broaden the empirical base of agreement attraction errors by considering their properties in languages outside the handful of well-studied languages in this domain of research, as the restriction to these languages could in principle unduly influence representational commitments made on the basis of particular kinds of verbal agreement paradigms.

A crosslinguistic perspective is an important one for addressing a pressing question in memory models concerning the distinction between a grammatical feature and a processing/memory retrieval cue. While it is clear that theoretical work can identify features utilized by the grammatical system, it is an open question how these features map onto cues which are used in the memory retrieval system. Just because grammar provides a feature as part of a contrast does not mean that the parser must utilize this feature in dependency resolution. Here, again, the memory retrieval models force an explicit commitment insofar as predictions about which constituents in memory are retrieved (as well as the latency of that retrieval) can only be made when one is explicit about the inventory of cues available to the system. Investigating these questions in languages which utilize different grammatical features in distinct ways is therefore a necessary part of understanding the feature-to-cue mapping.

Finally, an additional reason that crosslinguistic consideration is important relates to the way that memory models relate available cues to available activation in the system. Since the eventual retrieval target is the chunk in memory which has the highest activation at the retrieval event, and this activation is itself a function of two things: (1) the number of cues which a chunk shares with the goal and (2) the total number of chunks associated with each individual cue. A corollary of this architecture is that the number of available cues in a language directly modulates the amount of activation in the system. Adding more morphological features to discriminate NPs in memory should, in principle, drive down error rates. It is therefore an open question whether one expects agreement attraction in a language which is sufficiently morphologically rich in its verbal agreement[4]. Understanding the predictions such a system makes as available cues vary crosslinguistically is therefore an important way of validating such architectures more generally. Here, again, we believe testing memory models across the widest variety of languages should be an important research objective.

1.3. The Relevance of Arabic

It is here where Modern Standard Arabic (MSA; also equivalently just "Arabic" in what follows) is particularly well-suited as a language of interest. Arabic is spoken by over 200 million people worldwide and MSA is a *lingua franca* used in writing and formal speech across different regional varieties of spoken Arabic (as well as within-dialect groups). MSA is relevant for agreement attraction studies because it has verbal agreement for grammatical gender for both masculine and feminine subjects, a dual number (Ryding, 2005, pp. 438–444), and case marking which is optional on NPs under particular circumstances (Ryding, 2005, pp. 165–205). These kinds of agreement are in addition to the more standard singular/plural distinction seen in languages such as English and demonstrated for Arabic in (5)[5]:

[4]This is actually true of a broader array of models than just the ones considered here, such as the Competition Model (e.g., MacWhinney, 1987; MacWhinney and Bates, 1989; MacWhinney, 2001).

[5]In this and all subsequent glosses, we use the following abbreviations for grammatical features: MASC = masculine gender, FEM = feminine gender, 3 = third person, PERF = perfect aspect, NOM = nominative case, ACC = accusative case, and COMP = complementizer. Finally, because Arabic orthography is ordered

(5) a. الطالب درس اللغة العربية.

at$^{\varsigma}$-t$^{\varsigma}$aalib daras-a
the-student(.MASC) study-3.MASC.SG.PERF
al-luɣa al-ʕarabiyya.
the-language the-arabic

"The student studied Arabic."

b. الطلاب درسوا اللغة العربية.

at$^{\varsigma}$-t$^{\varsigma}$ulaab daras-uu
the-student(.MASC.PL) study-3.MASC.PL.PERF
al-luɣa al-ʕarabiyya.
the-language the-arabic

"The students studied Arabic."

Additionally, MSA has two distinct strategies for forming plurals on nouns: (1) a plural formed by suffixation, called the "sound" plural (طالبة – طالبات/t$^{\varsigma}$*aaliba*—t$^{\varsigma}$*aalib-aat*, "student ∼ students (fem.)") and (2) a plural formed by ablaut, called the "broken" plural in traditional Arabic grammar (شيخ – شيوخ ʃajx ∼ ʃujuux, "sheikh ∼ sheikhs"). While the latter strategies for pluralization would normally be referred to as "irregular" in the English literature, the broken/ablauting plural strategy is very common in Arabic—if not more common than the sound/suffixing plural strategy (see, e.g., Ryding, 2005, pp. 132–204). For nouns which take suffixes in the plural, these suffixes are absolutely regular: in the feminine there is only /-aat/ (Ryding, 2005, pp. 132–133). For masculine nouns which take suffixing plurals, there are up to two suffixes, /-uun/ for nominative case and /-iin/ for genitive and accusative case (Ryding, 2005, p. 140)[6]. By contrast, the number of broken plural patterns is considerably higher: (Ryding, 2005) lists 26 distinct patterns and (McCarthy and Prince, 1990b), following (Wright, 1889a,b), give 31 patterns. This sound/broken contrast is an important one because it cross-cuts other grammatical concerns in Arabic: what type of case morphology is available for a noun depends on what kind of plural it takes (Ryding, 2005, pp. 165–204); affects theoretical conceptions of morphological process (McCarthy and Prince, 1990b); and may affect lexical access at the word level (Mimouni et al., 1998).

These two types of plurals are of particular interest because they allow investigation of the representation of plurality in both linguistic representation and the working memory system. A recurring question in experimental work on Semitic is to what extent grammatical theories concerning word representation postulate representational constructs which are useful for psycholinguistic theorizing. Specifically, traditional Arabic grammars characterize most words as consisting of a consonantal ROOT (made up of two to five consonants) interleaved among vowels in a so-called prosodic TEMPLATE (see, e.g., Ryding, 2005, pp. 45–50), a characterization which has heavily influenced linguistic theories of the language, as well (see, for example, McCarthy, 1979, 1981; McCarthy and Prince, 1990a,b; Ussishkin, 2000, 2005; Tucker, 2010, 2011; Ussishkin et al., 2015). For instance, (6) gives examples of several distinct words all sharing the root $\sqrt{\text{ktb}}$:

(6) a. كتب/kataba, "he wrote"
b. كاتب/kaataba, "he corresponded"
c. كتاب/kitaab, "a book"
d. اكتب/uktub, "write!"
e. مكتب/maktab, "an office/desk"
f. مكتبة/maktaba, "a library"

Formal Arabic grammar is mostly uniform in its description of Arabic morphology in these root and template terms. However, depending on the part of grammar being considered, psycholinguistic work has found variable evidence for the template, mainly from priming (for Hebrew, see Frost et al., 1997; Deutsch et al., 1998; Frost et al., 2000; for Maltese, see Ussishkin and Twist, 2007; Ussishkin et al., 2011; and for Arabic, see Boudelaa and Marslen-Wilson, 2001, 2004b,a, 2005; Boudelaa et al., 2010; Boudelaa and Marslen-Wilson, 2011). Notably, this reliance on priming has led to most conclusions about the psycholinguistic validity of these representations being confined to lexical decision independent of sentential context. However, one thing which is not addressed in most of the recent work on Semitic morphosyntax is how this root-and-pattern system interacts with the representation of plurality in both grammar and parsing (a notable exception being the early work of McCarthy, 1981, where it is explicitly claimed that templates are morphemes which bear grammatical content). For instance, one can easily wonder, for broken plurals, where the grammatical plural feature is located in the representation and how such a representation translates into use for parsing. Given that there is enough linguistic and psycholinguistic evidence that suggests one should take the broken/ablauting vs. sound/suffixing contrast seriously on Arabic-internal terms, here we attempt to see whether this contrast is informative for diagnosing how the processing system encodes nouns in general and plurality more specifically. While this question is particularly salient for Semitic-internal debates, it is germane to research on morphological representations outside of this language family, as well insofar as other languages have similar representations for morphological features.

1.4. The Present Study

As promising as the grammatical situation is in MSA for probing the mapping between features and cues in agreement dependency resolution, it remains to be seen whether or not agreement attraction exists for the standard number features seen in previous studies. The study reported here took up this question by considering the resolution of agreement dependencies involving plural attractors. In better-studied languages such as English, one finds that plural attractors occasionally condition erroneous plural verbal morphology, as in (7):

(7) *The key* to **the cabinets *are*** rusty from years of disuse.

In production, agreement attraction errors manifest as production of the erroneous verb (Bock and Miller, 1991; *et seq.*),

right-to-left, we do not gloss the Arabic itself; it is included for reference, and a gloss is included for the phonetic transcription.

[6]For both masculine and feminine nouns, the situation is modulated by definiteness, where definiteness is defined as marking with the definite article /al-/. Since all the nouns used in our study were definite, we focus on definite NPs only in this description. See (Ryding, 2005) for ample discussion.

whereas in comprehension attraction errors manifest as facilitation on ungrammatical verbs when attraction configurations are present (Pearlmutter et al., 1999; Wagers et al., 2009; Dillon et al., 2013; Tanner et al., 2014) or as a reduced-amplitude P600 in attraction configurations in ERP research (Tanner et al., 2014).

The study reported below therefore also investigated the representation of number cues across different kinds of plurals in Arabic using self-paced reading while counterbalancing attractor plural type. We predicted the existence of such errors in comprehension in MSA as a facilitation to erroneously plural verbs in the presence of a plural attractor relative to singular attractors in the same context. As for plural type, we were more reserved in our prediction, being unsure as to the theoretical status of plural types in the language. It has previously been observed for English (Bock and Eberhard, 1993) that irregular plural formation (*ox∼oxen*, *mouse∼mice*) on an attractor NP does not condition differential error rates. However, in that production study, the focus was on a language for which ablauting plurals are exceptionally rare and form a small corner of the nominal inventory of the language. In MSA, the relative abundance of ablauting plurals may very well mean differential behavior between suffixing sound plurals and ablauting broken plurals. Any such difference, in turn, would have implications for the mapping between grammatical plural features and plural retrieval cues on NP constituents in working memory.

2. Self-Paced Reading

As a prerequisite for any systematic investigation of the unique properties of Arabic morphology and their effect on agreement attraction, it is first necessary to be sure that attraction errors of the kind documented for other languages occurs in MSA. We think this an especially important contribution given the relative inhospitability of the Arabic agreement system to agreement attraction errors: the system involves a large number of cues (person, number, and gender) which assist the parser in retrieving the correct subject. In order to determine whether attraction errors are possible in MSA, an experiment was designed based upon the relative clause stimuli in the initial (Bock and Miller, 1991) study. The purpose of this experiment was to ensure that subject-verb agreement errors for singular and plural number do occur in a relatively frequently-occurring grammatical configuration that allows for subsequent manipulation of less well-studied number and gender alternations.

We therefore test the Arabic equivalents of a subset of preambles from the Bock and Miller (1991) study on English. Specifically, Bock and Miller (1991) tested production agreement errors elicited after giving participants preambles such as *The boy(s) that liked the snake(s)* … which varied based on the number for the subject [*the boy(s)*] and the local distractor noun [*the snake(s)*]. However, we were also interested in the real-time processing properties of attraction errors, so we investigate comprehension by measuring the reading times for complete versions of these sentences. This allowed us to simultaneously remain close to the original phenomenon in English while simultaneously exploring the comprehension of agreement in Arabic.

2.1. Method

2.1.1. Participants

Participants were 114 native speakers of Arabic from the University of the United Arab Emirates and NYU Abu Dhabi student bodies (113 female; mean age 21.1 years)[7]. All participants had no history of language disorders and read MSA regularly. Each participant provided written informed consent and was compensated for their participation. This experiment was approved by the NYU Abu Dhabi Institutional Review Board and the UAEU Ethics Committee.

2.1.2. Materials

A set of 48 sentences was constructed, each containing a subject relative clause with an animate object modifying the animate subject of a transitive verb. Subject relative clauses were chosen because they are a long-standing example of a configuration which creates agreement attraction errors (e.g., Bock and Miller, 1991) and are relatively common in MSA. In this sense they are a better choice than the canonical *NP—PP* configuration in more memorable examples such as *The key to the cabinets*.... The issue these constructions pose for the present study is that Arabic does not easily allow adverbs to be placed between subject and verb (Tucker, 2011) the inclusion of which was a *desideratum* of our stimuli. This is because, following (Wagers et al., 2009), we wished to insert an adverb or adverbial prepositional phrase between the end of the relative clause and the target main clause verb in order to mitigate plural NP spillover effects into the target region. All the stimuli therefore had the structure *NP1—Complementizer—RC Verb—NP2—Adv/PP—Verb—Continuation*. An example of such a sentence appears in (8):

(8) المترجم الذي ساعد الرئيس أحيانًا يتكلم خمس لغات بفصاحة

ʔal-mutarʒim-u ʔalla ii
the-translator-NOM COMP.MASC.SG
saaʕad-a ʔal-raʔiis-a ʔaħjaanan
helped-3.MASG.SG the-president-ACC often
ja-takallamu xamsata luʁaat-in
3.SG.MASC-speaks five languages-ACC
bi-fasˁaaħatin.
with-fluency

"The translator who helped the president often speaks five languages fluently."

Several constraints guided the construction of these experimental sentences: Firstly, Arabic has a series of prepositions which are only a single syllable/orthographic character and which are written with no space separating them from the complement NP. Only these prepositions were used in constructing adverbial PPs, meaning that the buffer region between distractor NP and target verb was no more than one orthographic word for any sentence.

[7]The discrepancy in gender in this sample is a product of the student body makeup at the UAEU, where the majority of testing was conducted. This university has a 3:1 female-to-male student ratio and has gender-segregated campuses. Testing was conducted on the female side because of the larger number of students, meaning that male students were not able to participate at the UAEU.

Secondly, for any given sentence both the subject and distractor NP were the same grammatical gender (masculine or feminine), and the total number of masculine and feminine gender nouns was balanced across sentences (24 masculine, 24 feminine). We decided not to allow different genders in the same sentence because of the confound introduced by the complementizer in MSA, as it must agree with definite head nouns (Ryding, 2005, pp. 322–324). Because of this, the true subject would receive an additional disambiguating cue from the complementizer's gender. However, the complementizer does inflect for grammatical number, meaning that in our stimuli the true subject receives reinforcement from the singular complementizer in conditions with plural attractors.

Additionally, we sought to vary the kind of plural which the attractor NP takes in the plural conditions. However, grammatical case in Arabic is normally optionally expressed in diacritics which are not written in everyday MSA, with the exception of suffixing masculine plurals, which do show an orthographic distinction between accusative and nominative case (represented by a change in an orthographically obligatory long vowel). In order to avoid adding a potentially disambiguating cue, case-marking, all masculine distractor NPs took broken plurals and all feminine distractor NPs took suffixing plurals. We also opted to conflate gender and plural type because MSA does not furnish a sufficiently large number of broken feminine plurals which refer to animates. This strategy allowed balancing of gender and suffixation in the plural in a grammatically natural way without introducing confounds from orthographically-represented grammatical case. This design allows us to check whether different pluralization processes (ablaut vs. suffixation) influence agreement attraction effects differently, although in our design this is necessarily confounded with grammatical gender.

In addition to gender and plural type, the sentences were also counterbalanced for whether the target verb appeared in the present or past tense. This was done because MSA has two distinct series of affixes for verbal agreement: (1) the present tense, with both a prefix and suffix and (2) the past tense, with suffixes only (see, e.g., Ryding, 2005, pp. 438–444). Counterbalancing in this way allowed conclusions to be drawn about agreement independent of the specific affix series employed. We did assess the effect of tense/aspect in the reading time results presented below and found no effect of the affix series employed.

Finally, stimuli in Arabic must stake a position on the orthographic representation of short vowels. Arabic is written in an alphabet which only represents long vowels, where short vowels are only written in religious texts, poetry, and texts for language-learners. In everyday formal written Arabic, short vowels are sometimes employed when an orthographic string is lexically ambiguous without some short vowel specification or in a way which is not resolvable from sentential context. The effects of adding superfluous or normally unwritten short vowels to Arabic language stimuli is understudied, and therefore a point of particular concern. In our stimuli, we therefore employed minimal diacritics only where lexical ambiguity would result if the diacritics were not used. This is a common scheme for representing diacritic marks in MSA and matches what is seen in everyday formal writing in the Arab world.

For each experimental sentence, four variants were constructed by systematically varying the morphological number of the object of the relative clause (*NP2*, the *attractor* or *distractor*) and the main clause verb (the *Verb*). This resulted in four conditions per sentence which are labeled according to the number of *NP2* and *Verb*: (S)ingular or (P)lural. We call the conditions in which the verb is plural *ungrammatical* conditions, since all subjects were singular in the experimental items. A complete item set appears in **Table 1** and the complete list of sentences appears in the Supplementary Materials.

The 48 sets of four sentences were distributed across four lists in a Latin Square design and combined with 144 grammatical filler items of a similar length in order to distract from the target items. None of the fillers contained the subject relative clause construction contained in the stimuli. This resulted in a filler-to-item ratio of 3:1 with 25% of the sentences being ungrammatical.

In this study we expect several things based upon the previously published studies for Germanic, Romance, and Slavic languages. Specifically, we expect to find a main effect of grammaticality in the critical verb region (RCV) and subsequent regions owing to possible spillover. Moreover, we expect to find a interaction between this factor and the attractor number factor in the critical verb region (possibly including spillovers) driven by slower reading times for the Sg/Ungram condition relative to the Pl/Ungram condition—this is the attraction configuration. Moreover, we expect to find no difference between the two grammatical conditions, Sg/Gram and Pl/Gram, given that no comprehension attraction effects have been observed in the previous literature. Additionally, following the discussion in Wagers et al. (2009), we expect to find a main effect of attractor number alone in the attractor region (NP2), a plural reading time effect noted in that work but not presently well-understood. Finally, we have no *a priori* expectations about the nature of the effect of plural type, but suspect that it is relevant for on-line processing given its centrality in the grammatical and lexical access literature.

2.1.3. Procedure

Subjects were seated comfortably up to eight at a time at a table in a quiet room in front of computers on which the experimental software had been pre-loaded. Sentences were presented using the Linger software (Rhode, 2003) in a self-paced word-by-word moving window paradigm (Just et al., 1982). Each trial begin with the display of a screen containing the sentence masked by dashes (including spaces and punctuation). Each time the participant pressed the space bar, a single word was revealed and the previous word re-masked. All items were presented in the Courier New Arabic font in 28pt bold type. A yes/no comprehension question (not an acceptability judgment) followed each sentence, appearing on the screen all at once. Comprehension questions were designed in such a way that the answer could be provided independent of experimental manipulations—no questions asked about the attractor NP or the main clause verb. None of our comprehension questions required lexical elaboration of the item or difficult semantic processing. A majority of the comprehension questions asked about the relative clause verb or the post-critical region continuation. As an example, the item *The*

TABLE 1 | A complete item set for one sentence for the experiment.

Condition	NP1 R1	Comp R2	RCV R3	NP2 R4	Adv R5	V R6	Continuation R7–R10
Sg/Gram	السفيرة	التي	استضافت	المندوبة	سنوياً	تحدثت	في الأمم المتحدة.
	The ambassador	who	hosted	the diplomat	yearly	spoke(FEM.SG)	at the United Nations
Sg/Ungram	السفيرة	التي	استضافت	المندوبة	سنوياً	تحدثن	في الأمم المتحدة.
	The ambassador	who	hosted	the diplomat	yearly	spoke(FEM.PL)	at the United Nations
Pl/Gram	السفيرة	التي	استضافت	المندوبات	سنوياً	تحدثت	في الأمم المتحدة.
	The ambassador	who	hosted	the diplomats	yearly	spoke(FEM.SG)	at the United Nations
Pl/Ungram	السفيرة	التي	استضافت	المندوبات	سنوياً	تحدثن	في الأمم المتحدة.
	The ambassador	who	hosted	the diplomats	yearly	spoke(FEM.PL)	at the United Nations

student who saw the professor(s) yesterday studied electrical engineering at the university. was followed by the question *Did the student see someone?*. The 'f/ب' key was used for "yes (نعم)" and the 'j/ت' key used for "no (لا)." Onscreen feedback was provided for both correct and incorrect answers. Participants were instructed to read at a natural pace ensuring comprehension and were not alerted to the presence of grammatical errors in the stimuli. The order of sentence presentation within each list was randomized by the experimental software for each participant. Four practice items were presented before the start of the experiment.

2.1.4. Data Analysis

Subjects which were less than 70% accurate on comprehension questions were excluded from further analysis on the grounds that they were not sufficiently attentive to the task; this criterion resulted in the exclusion of 10 subjects. Outliers were handled by Winsorizing the extreme 5% of the data (Ratcliff, 1993). No other exclusion criteria were used.

Data from both the comprehension question responses and remaining region-by-region reaction times were analyzed using mixed effects regression (Baayen et al., 2008). The answers to the comprehension questions were entered into several logistic mixed effects models including experiment, condition, and experimental independent variables (attractor number and grammaticality) as fixed effects and subjects and items as random effects with intercepts only. Self-paced reading data for each region of interest (R4, the attractor region, through R8, the second post-critical verb region) were entered into a linear mixed effects model fit using restricted maximum likelihood estimation with both subjects and items as random effects and several predictors as fixed effects: (1) attractor number, (2) grammaticality, (3) attractor plural type (ablauting/suffixing), (4) item order in the experimental presentation, (5) log frequency of the plural of the attractor according to the arabiCorpus (Parkinson, 2012), (6) word length in characters, (7) the previous region's reading time, and (8) interactions of terms (1–3). Categorical predictors were dummy-coded using the following default values: (1) grammaticality = grammatical, (2) attractor number = singular, and (3) gender/plural type = feminine (sound/suffixing) and neither categorical nor continuous predictors were centered. Our random effects structure was comprised of intercepts for subjects and items. For both the comprehension and reading-time results we used a minimal random effects structure in order to ensure convergence of the models (but see Barr et al., 2013). Degrees of freedom were estimated using the Welch-Satterthwaite approximation in order to calculate a p-value; we therefore report t-values directly instead of $z-$scores or 95% confidence intervals generated by bootstrapping or MCMC sampling. More details on the modeling for the reading time results can be found in the Supplementary Materials.

2.2. Results

2.2.1. Comprehension Question Accuracy

The mean comprehension question accuracy pooled across subjects and items to both experimental items and fillers was 88.2% and was significantly lower for experimental items (80.0%) than for fillers (91.1%) (logistic mixed-effects model $\hat{\beta}_1 = 1.44$; $z = 19.80$; $p < 0.0001$). We believe this lower accuracy to the experimental item comprehension questions is due to errors in the construction of some of the questions themselves. Participants reported confusion over the intent of seven of the questions; with these questions excluded, experimental item accuracy increased to 86.1%. Nevertheless, we exclude data from these items when the comprehension question was answered incorrectly in the reading-time analysis which follows, as this is the most conservative approach.

Accuracy rates for singular attractors were 81.0 $\pm$ 1.2% (with standard errors computed over participant means) for grammatical sentences and 78.8 $\pm$ 1.3% for ungrammatical sentences. For plural attractors, accuracy rates were 82.7 $\pm$ 1.3% for grammatical sentences and 76.0 $\pm$ 1.4% for ungrammatical sentences. The configuration of plural attractor and grammatical verb had a significant impact on question accuracy ($\hat{\beta} = 0.36$; $z = 2.11$; $p = 0.03$) such that participants were more likely to be correct in this condition relative to the attraction configuration of plural attractor and ungrammatical verb.

2.2.2. Self-Paced Reading

The self-paced reading results for all items are presented immediately below. Because of our *a priori* interest in the impact of grammatical and lexical access-related differences in plural

formation type on agreement attraction, we provide some additional results by gender/plural type, as well. In what follows, we focus our reporting on the results of the experimental manipulations of Attractor Number, Grammaticality, and Gender/Plural Type. We do not comment on the presence of effects due to the frequency of the attractor, word length, or previous region's reading time, as these predictors are commonly found to be explanatory in reading time studies and we have nothing to add here to their interpretation as determinants of reading time.

2.2.2.1. All items

The results from the experiment are presented in **Figure 1** and the mixed-effects model results for the attractor region (R4) and critical verb region (R6) appear in **Tables 2**, **3**. Linear mixed-effects model results for all other regions of interest are included in the Supplementary Materials.

The relative clause attractor region (R4) contained a main effect gender/plural type such that masculine attractor NPs were read more slowly than feminine attractor NPs [$\hat{\beta} = 72.00$; $t_{(143.00)} = 2.53$; $p = 0.01$]. Additionally, there was an interaction between gender/plural Type and attractor number [$\hat{\beta} = -62.24$; $t_{(3833.00)} = -2.37$; 0.02] which was driven by significantly longer reading times to plural attractors for feminine attractors [$t_{(207)} = 2.99$; $p = 0.003$; plural mean = 674.80 ms; singular mean = 629.10 ms]. The same was not true of masculine attractors [$t_{(207)} = -1.11$; $p = 0.27$; plural mean = 600.42 ms; singular mean = 615.34 ms]. However, there was no main effect of attractor number alone [$\hat{\beta} = 14.47$; $t_{(757.00)} = 0.65$; $p = 0.51$, *n.s.*; singular mean = 624.87 ms; plural mean = 637.37 ms]. In the adverb region (R5), there were no effects of any of the experimental manipulations (all t's < 1.3).

The main clause verb region (the critical region, R6) showed a main effect of grammaticality such that ungrammatical utterances were read much more slowly than grammatical utterances [$\hat{\beta} = 102.56$; $t_{(3657.00)} = 6.70$; $p < 0.0001$ ungrammatical mean = 651.15 ms; grammatical mean = 575.91 ms]. The main verb region also displayed an interaction of grammaticality and gender/plural type [$\hat{\beta} = -59.85$; $t_{(1157.00)} = -2.50$; $p = 0.01$]. This appeared to be due to a larger grammaticality effect for masculine items (ungrammatical mean = 653.00 ms; grammatical mean = 567.54 ms) than for feminine items (ungrammatical mean = 653.11 ms; grammatical

TABLE 2 | Table of coefficients for a linear mixed effects regression with gender/plural type for the attractor region (R4).

Factor	$\hat{\beta}$	*t*	*df*	*p*
Intercept	**382.05**	**6.01**	**100.00**	**<0.0001**
Attr	14.47	0.65	757.00	0.51
Grammaticality	16.94	0.92	3926.00	0.36
Gender/Plural Type	**72.00**	**2.53**	**143.00**	**0.01**
Item Order	**−1.29**	**−15.18**	**3964.00**	**<0.0001**
Attr Frequency	−14.39	−1.50	104.00	0.14
Length	**35.62**	**4.71**	**72.00**	**<0.0001**
Previous Region RT	**0.17**	**11.99**	**4036.00**	**<0.0001**
Attr × Gram	−27.82	−1.07	3927.00	0.28
Attr × Gender	**−62.24**	**−2.37**	**3833.00**	**0.02**
Gram × Gender	−6.65	−0.26	3929.00	0.80
Attr × Gram × Gender	33.71	0.92	3931.00	0.36

p-values computed using the Welch-Satterthwaite approximation. Predictors significantly different from 0 at $\alpha = 0.05$ highlighted in bold.

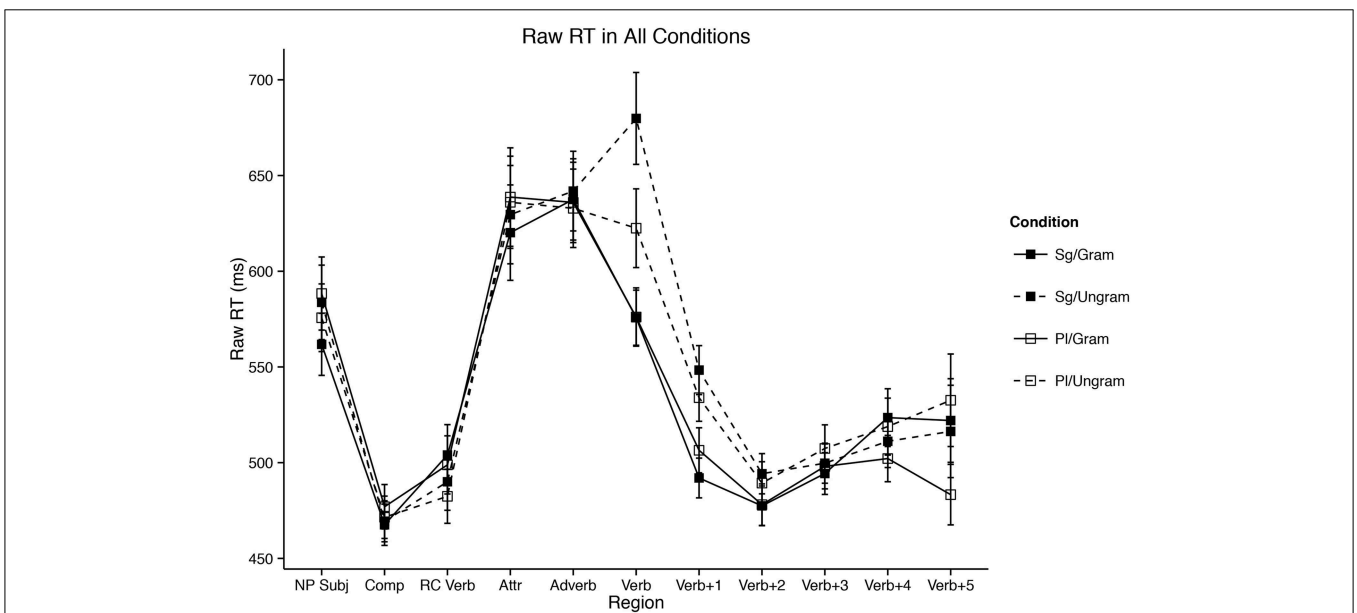


FIGURE 1 | Self-paced reading results. Region by region means segregated by attractor number and verb number. Error bars represent the standard error of the mean computed across subject averages.

TABLE 3 | Table of coefficients for a linear mixed effects regression with gender/plural type for the critical verb region (R6).

Factor	$\hat{\beta}$	*t*	*df*	*p*
Intercept	**557.54**	**13.61**	**122.00**	**<0.0001**
Attr	24.39	1.44	1639.00	0.15
Grammaticality	**102.56**	**6.70**	**3657.00**	**<0.0001**
Gender/Plural Type	−11.90	−0.53	152.00	0.60
Item Order	**−1.38**	**−19.90**	**3961.00**	**<0.0001**
Attr Frequency	16.24	1.90	129.00	0.06
Length	**28.68**	**3.72**	**73.00**	**0.0004**
Previous Region RT	**0.04**	**4.81**	**4062.00**	**<0.0001**
Attr × Gram	**−72.40**	**−3.44**	**3925.00**	**0.0006**
Attr × Gender	−14.50	−0.68	3927.00	0.49
Gram × Gender	**−59.85**	**−2.50**	**1157.00**	**0.01**
Attr × Gram × Gender	39.08	1.31	3928.00	0.19

p-values computed using the Welch-Satterthwaite approximation. Predictors significantly different from 0 at $\alpha = 0.05$ highlighted in bold.

mean = 582.06 ms). Crucially, the main clause verb region also yielded an interaction between attractor number and grammaticality [$\hat{\beta} = -72.40$; $t_{(3925.00)} = -3.44$; $p = 0.0006$]. Planned comparisons revealed that this was driven by an effect of attractor number in the ungrammatical conditions such that plural attractors were read more quickly than singular attractors [$t_{(103)} = 4.48$; $p < 0.0001$; plural mean = 622.48 ms; singular mean = 679.83 ms] but no difference in the grammatical conditions [$t_{(103)} = -0.04$; $p = 0.97$; plural mean = 576.09 ms; singular mean = 575.74 ms]. This agreement attraction interaction did not appear to be modulated by gender/plural type in the main clause region [$\hat{\beta} = 39.08$; $t_{(3928.00)} = 1.31$; $p = 0.19$], though see the following section for some consideration on this finding.

Following the critical main verb, the first spillover region (R7) showed a main effect of attractor number such that plural attractor sentences were read more slowly in R7 than singular attractor sentences [$\hat{\beta} = 27.53$; $t_{(1598.00)} = 2.51$; $p = 0.01$; plural mean = 520.22 ms; singular mean = 520.18 ms], though as the means suggest this effect is not significant in a follow-up comparison [$t_{(207)} = 0.005$; $p > 0.99$]. We believe this effect attributable to our use of dummy coding, as a sum-coded model does not reveal this effect [$\hat{\beta} = 0.87$; $t_{(757.00)} = 0.34$; $p = 0.74$] despite qualitatively different results for all other effects. The effect of grammaticality which began at the main clause verb persisted into the first spillover region, with ungrammatical sentences read more slowly than grammatical sentences [$\hat{\beta} = 70.58$; $t_{(3925.00)} = 7.37$; $p < 0.0001$; ungrammatical mean = 541.16 ms; grammatical mean = 499.24 ms]. Additionally, the attraction interaction of attractor number and grammaticality which began in the previous region persisted into R7 [$\hat{\beta} = -37.92$; $t_{(3924.00)} = -2.81$; $p = 0.005$]. However, in this region this interaction was driven by significantly longer reading times to plural attractors in grammatical conditions [$t_{(103)} = -2.37$; $p = 0.02$; plural mean = 506.48 ms; singular mean = 492.01 ms]. In ungrammatical conditions, plural attractors conditioned faster reading times than singulars, though this effect did not reach significance [$t_{(103)} = 1.58$; $p = 0.11$; plural mean = 533.95 ms; singular mean = 548.36 ms]. Additionally, R7, the first spillover region, also showed a significant interaction of grammaticality and gender [$\hat{\beta} = -30.40$; $t_{(3925.00)} = -2.26$; $p = 0.02$]. This interaction was due to significantly longer reading times to grammatical sentences with masculine attractors than those with feminine attractors [$t_{(207)} = 4.12$; $p < 0.0001$; masculine mean = 515.22; feminine mean = 486.90]. A similar trend was only marginal in the ungrammatical sentences [$t_{(207)} = 1.77$; $p = 0.07$ feminine mean = 536.89 ms; masculine mean = 553.46 ms].

Finally, in the second spillover region, there were no significant effects of any of the experimental manipulations (all *t*'s < 1.85), however the main effect of grammaticality was marginally present [$\hat{\beta} = 14.04$; $t_{(3926.00)} = 1.83$; $p = 0.07$]. This was again because ungrammatical sentences were read longer two words downstream from the main clause verb than grammatical sentences (ungrammatical mean = 491.79 ms; grammatical mean = 477.76 ms).

2.2.2.2. By gender/plural type

Results for the experiment segregated by plural type/gender of the attractor NP are presented in **Figure 2**. While our mixed-effects model presented above did not show a significant interaction of gender/plural type and the attraction effect (the three way interaction of Attractor Number × Grammaticality × Gender/Plural Type was not significant), we had two reasons for investigating the interaction further: (i) *a priori* considerations concerning the grammatical status of plural formation type in MSA (see §1.3, above) and (ii) visual inspection of the difference between the two genders in **Figure 2**. Specifically, we were suspicious of the possibility that feminine items were showing more attraction relative to masculine items, if the latter were indeed displaying attraction at all.

We also suspected that the lack of a significant interaction in our mixed-effects model was partially due to our choice of outlier exclusion method: Winsorizing five percent of the data could have erroneously removed long reading times to critical verbs in the Sg/Ungram and Pl/Ungram conditions—these conditions are fully ungrammatical, and since this is the first reading-time study on MSA, there was no *a priori* way to know the expected size of reading time increases to fully ungrammatical verbs. Such an interpretation is also consistent with an emerging view that agreement attraction effects are driven by reading times in the right tail of the distribution (Staub, 2009, 2010; Lago et al., 2014) It is therefore possible that a 5% cutoff by-region is too conservative and results in the exclusion of data mistaken for outliers. To this end, we ran an identical analysis with no Winsorization. The results of this analysis are qualitatively identical to the analysis presented above, save for the three-way interaction of Attractor Number × Grammaticality × Gender/Plural Type in the main clause verb region (R6); in unwinsorized model, this term emerges as marginal [$\hat{\beta} = 99.13$; $t_{(3926.00)} = 0.05$]. This marginal effect is driven by longer reading times to Sg/Ungrammatical conditions relative to Pl/Ungrammatical conditions in the feminine items [$t_{(103)} = 3.38$; $p = 0.001$; Sg/Ungram mean = 732.05 ms; Pl/Ungram mean = 640.11 ms], a contrast which is not present for masculine items [$t_{(103)} = 0.24$;

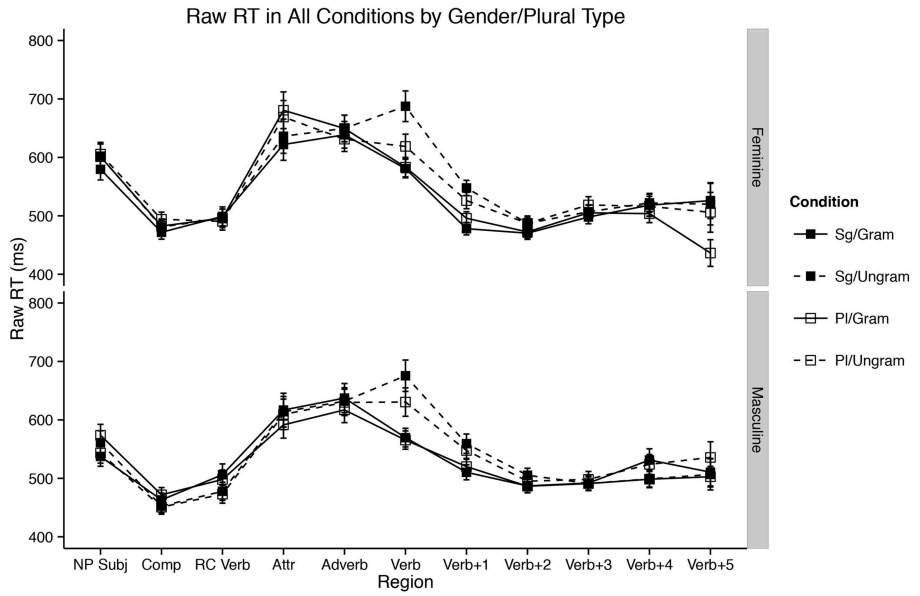


FIGURE 2 | Self-paced reading results. Region by region means segregated by attractor number and verb number for both genders/plural types. Error bars represent the standard error of the mean computed across subject averages.

$p = 0.81$; Sg/Ungram mean = 701.74 ms; Pl/Ungram mean = 694.74].

3. Discussion

The results of our study clearly show that agreement attraction errors can be elicited in the comprehension of written MSA. The results from the critical verb region in this study show that reading times are universally increased in the presence of a grammatically incorrect verb, but that the magnitude of this increase in reading time is modulated by the kinds of non-subject (and therefore, structurally inaccessible for subject–verb agreement) NPs appearing in the preceding context. Specifically, when one of these preceding nouns has features which match the erroneous verb along the dimension the subject does *not* match, then a smaller increase in reading time is observed relative to cases in which no nouns in the preceding context overlap in features with the verb. Alternatively, one can view this effect as a facilitation relative to ungrammatical sentences where the attractor does *not* match the erroneously plural verb. However, it is viewed, this effect is one of the hallmarks of agreement attraction errors.

Another distinguishing feature of agreement attraction phenomena which our data reveal in MSA is the general *absence* of an analogous effect in grammatical utterances. That is, when the verb and subject agree completely in grammatical features, there is no corresponding marginal increase in reading times when a distractor NP bears distinct grammatical features—a plural NP distractor has no effect in the context of a singular subject and verb. We do observe what could be effects of this kind at the first spillover region to a small degree. However, it is worth stepping back to consider the fact that in our study, in general, effects spill over less than they do in languages such as English. We do not have an explanation for this, but note that the agreement attraction effect does not spill over in either the full items analysis or the feminine items analysis for *ungrammatical* utterances. Moreover, the magnitude is suspect: one can assess the magnitude of an attraction effect by subtracting the reading time for plural conditions from the reading time to singular conditions (see §4, below)—what Dillon et al. (2013) call the *Intrusion Effect Size*. For ungrammatical utterances, this will be a positive number (erroneous facilitation to ungrammatical verbs), whereas for grammatical utterances, this would be a negative number (erroneous inhibition to grammatical verbs). At the critical verb region, our observed intrusion effect size is 57.35 ms, whereas in the first spillover region, the observed grammatical intrusion effect is −14.47 ms. We therefore think it safe to conclude that the transient effect in the first spillover region for feminines is not a *bona fide* attraction effect in grammatical utterances. If this logic is correct, Arabic self-paced reading responses to agreement attraction configurations mirror those observed for English in Tanner et al. (2014) and Wagers et al. (2009), but not (Pearlmutter et al., 1999).

Furthermore, our results add another piece to the growing body of evidence that there is something special about the processing of plural NPs in context (Wagers et al., 2009; Tanner et al., 2014). In our data, feminine plural NPs display longer reading times than their singular counterparts in the attractor region, a finding not shared by masculine NPs (see the attractor region in **Figure 2**). Two explanations have been advanced for this finding in the literature: (1) that it is the result of a "plural complexity effect" insofar as it is simply more difficult to process plurals than it is to process singulars, *ceteris paribus* (Wagers et al., 2009) and (2) that it is due to a "plural integration effect" insofar as it is difficult to integrate a semantically plural NP into a context which

features other singular nouns (Tanner et al., 2014, with support from findings in Nicol et al., 1997). Our findings from MSA help to shed some light on this debate. While it is possible to imagine a more nuanced version of the integration story, it is not obvious how to square the simple version of that account with the observation that semantically plural masculine/broken plural NPs do not display the reading time increase shown for feminines—both masculine and feminine plural attractors are semantically plural. If the integration explanation were correct, we might expect integration costs in both cases. While we will not attempt to resolve this fully here, we note that either one must elaborate the complexity story to include consideration of morphological plural formation strategies or return to the complexity suggestions of Wagers et al. (2009). Specifically, if one were to assume that complexity effects were correlated with the salience of plural marking on a noun (see §4, below, for some development of this idea), then we could take complexity to be about integrating plural marking with nominals stems. Alternatively, one could eschew this assumption about the salience of marking and take our data to support neither hypothesis, though we will not develop this idea here[8].

More broadly speaking, the differences between attractor genders/plural types in both the attractor and main clause verb regions are a significant diversion from both our prediction for Arabic and the established facts for English—masculine/broken/ablauting plurals behave distinctly from feminine/sound/suffixing plurals in our data. However, one must be careful in stating *how* this difference manifests. It would be tempting to conclude that attraction occurs with feminine/sound attractors but does not occur with masculine/broken items. This conclusion, while certainly possible, must be made cautiously, as we do not have sufficient evidence in this paper to reject the idea that attraction occurs in both genders/plural types (to wit, the lack of a three-way interaction in the main clause verb region). However, at the very least one could conclude that if attraction is present in the masculine/broken/ablauting items, the effect is much smaller than it is with feminines/sound/suffixing items (**Figure 5**). Again here, the intrusion effect size is instructive: with feminines, the mean agreement intrusion effect is 68.72 ms, whereas for masculines it is 45.00, computed across subject means. While we must be agnostic as to which, one of two things is true in our data: (i) masculine/broken/ablauting items do not display attraction or (ii) they do, but to a smaller degree than feminine/sound/suffixing items.

Even more broadly, we believe our results confirm a growing body of evidence in the literature about the location of agreement attraction effects in the distribution of reaction times to ungrammatical verbs (Staub, 2009, 2010; Lago et al., 2014). That is, previous work by Staub as well as Lago and colleagues has shown that the canonical pattern of agreement attraction in comprehension—facilitation to ungrammatical verbs in the presence of a matching distractor relative to ungrammatical verbs without a matching distractor—is present most strongly in the right tail of reaction time distributions to ungrammatical verbs. This appeared confirmed in our data by the change in the strength of the three-way interaction between Attractor Number, Grammaticality, and Gender as a function of our Winsorization cutoff. This can be seen for three values of Winsorization cutoffs in **Figures 3**, **4**. In both plots, decreasing the amount of data replaced by Winsorization does not change the qualitative pattern of results anywhere except in the shaded region, the critical verb. For the feminine items (**Figure 3**), decreasing the amount of removed data increases the separation between the Sg/Ungram condition and the remaining three conditions. For the masculine items (**Figure 4**), changing the cutoff affects both the Sg/Ungram and Pl/Ungram conditions, moving the two closer together. With no cutoff, the two conditions are identical, i.e., there is no attraction present. We take this to be further evidence that the right tail of reaction time distributions is vitally important for the study of violation responses, such as those seen with agreement attraction[9].

While we believe there is clearly a difference between feminine/suffixing and masculine/ablauting attractor items in our data, in this study this pluralization strategy-based difference is necessarily conflated with gender, both on the attractor NP and the verb itself. This is because of the way our stimuli were designed: all the suffixing plurals in our study were feminine nouns and all the ablauting plurals were masculine nouns. This is because of the grammar of MSA, which affords very few broken/ablauting feminine plurals which refer to animates. Moreover, grammatical case is necessarily orthographically present on masculine sound/suffixing plurals but not feminine sound plurals, making direct comparison somewhat confounded if those NPs were included.

Nevertheless, we find it plausible to tentatively assume that the differential agreement attraction effect across gender/plural type items is brought about by the different pluralization strategies and not by grammatical gender marking on the verb because of the absence of plural-based reading time trend on the attractor NPs for ablauting plurals. While it is conceivable that this lack of an effect is driven by their masculine gender, such an explanation cannot relate the *presence* of the slowdown in feminine/suffixing plurals to similar effects noted for English by Wagers et al. (2009). On the other hand, assuming the strength of the agreement attraction effect is driven by plural type allows for this cross-linguistically and theoretically coherent link as well as a unified explanation of the absence of the NP plural complexity and attraction effects.

[8]An intriguing possibility, raised by a reviewer, is that the lower accuracy rates to comprehension questions in the Pl/Gram condition could be a grammatical agreement attraction effect. We cannot rule this explanation out, but note two things: first, our comprehension questions were constructed to avoid use of lexical items which underwent an experimental manipulation (such as the attractor NP and main clause verb). Therefore, such an effect would have to be driven by the main clause subject, which was invariantly singular. Second, however, not *all* of our comprehension questions asked about this subject, making it difficult to assess this hypothesis in our current data, but the idea is viable for future research.

[9]Another possibility, raised by a reviewer, is that item order matters. Concretely, the idea would be that subjects are susceptible to grammatical effects early in the experiment, with these effects diminishing over time (as the participant begins to realize what is happening in the manipulations). We agree this is a possibility in our data, but have included item order as a predictor precisely so that our experimental effects can be trusted with item order held in abeyance. We hope to return to the issue of item order more concretely in future work.

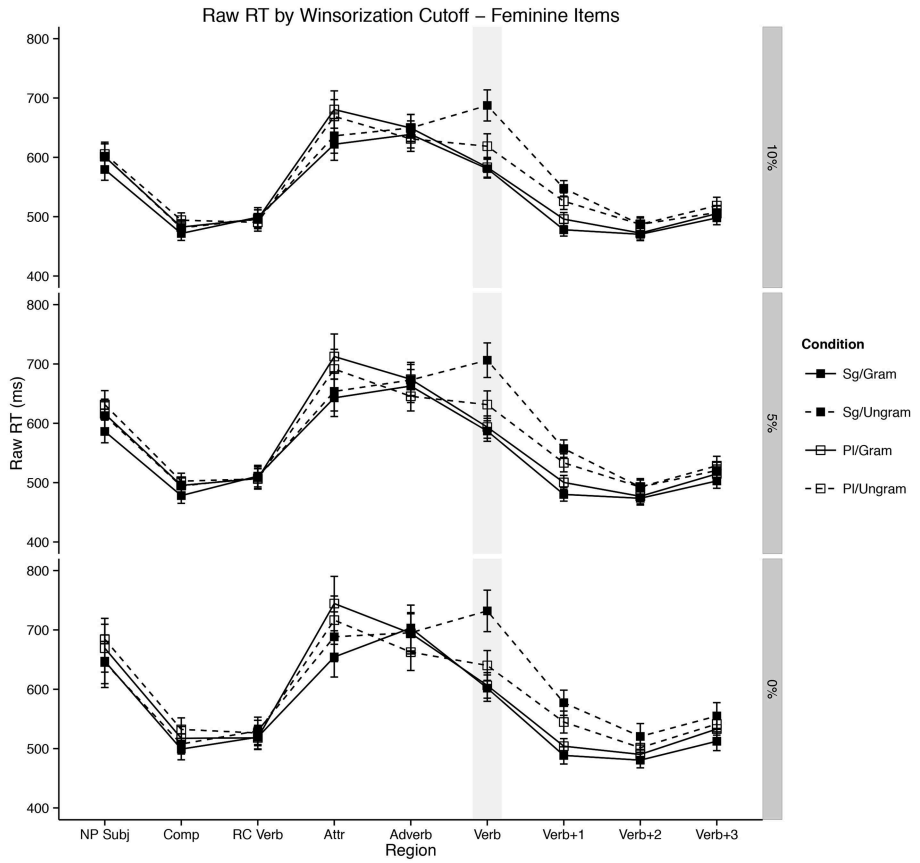


FIGURE 3 | Self-paced reading results. Region by region means of feminine items segregated by attractor number and verb number for three Winsorization cutoff thresholds. Error bars represent the standard error of the mean computed across subject averages.

One might reasonably wonder, at this point, what role, if any, frequency plays in explaining the observed patterns in MSA. In order to address this question, we calculated the token frequency of each attractor NP in the singular and plural form in the *Al-Hayat 1996* sub-corpus of the BYU arabiCorpus (Parkinson, 2012) and entered the plural log frequency values into our mixed effects models. In neither the attractor (R4) nor critical verb (R6) region were these terms significant in the model. However, it is worth noting that both singular [masculine mean $= 1.11$; feminine mean $= -0.18$; $t_{(39.97)} = 6.38$; $p < 0.0001$] and plural [masculine mean $= 0.70$; feminine mean $= -1.14$; $t_{(39.58)} = 9.27$; $p < 0.0001$] nouns did have significantly different log-frequencies for masculine and feminine nouns. A complete table of the frequencies for the attractor nouns in our experimental items appears in the Supplementary Materials.

The question still remains, however, as to what the explanation of this difference between the suffixing and ablauting plurals might be. Here we entertain two possibilities: (1) that the processing system does not have sufficient time for attraction effects to emerge because the system is at floor in the ungrammatical conditions and (2) there is something morphologically distinct about broken/ablauting plurals such that attraction cannot occur because the representation of number with these plurals is fundamentally different.

The first solution is plausible because broken/ablauting plurals in MSA are, on the whole, orthographically shorter than sound/suffixing plurals—usually between one and two characters shorter. Moreover, there is a clear, reliable difference between broken/ablauting and sound/suffixing plurals evident in our data set such that the latter are read around 70 ms slower than the former (see R4 **Figure 2**). The explanation in this approach would be that this shorter reading time is small enough that appreciable agreement attraction effects are not observable in such a short time frame—the system is simply under too much time pressure to reveal these effects and is at the *a priori* floor.

However, we do not believe this to be the correct approach for several reasons. Firstly, the directionality of this change in broken/ablauting plural reading times is in the wrong direction. Attraction in our data is revealed by the difference between plural-attractor plural-verb (Pl/Ungram) and singular-attractor plural-verb (Sg/Ungram) conditions—in both cases the plural verb is ungrammatical but only in the former case does a partially matching attractor lead to decreased reading times. However, broken/ablauting plurals clearly involve *faster* reading times across the board, meaning that the Pl/Ungram condition is

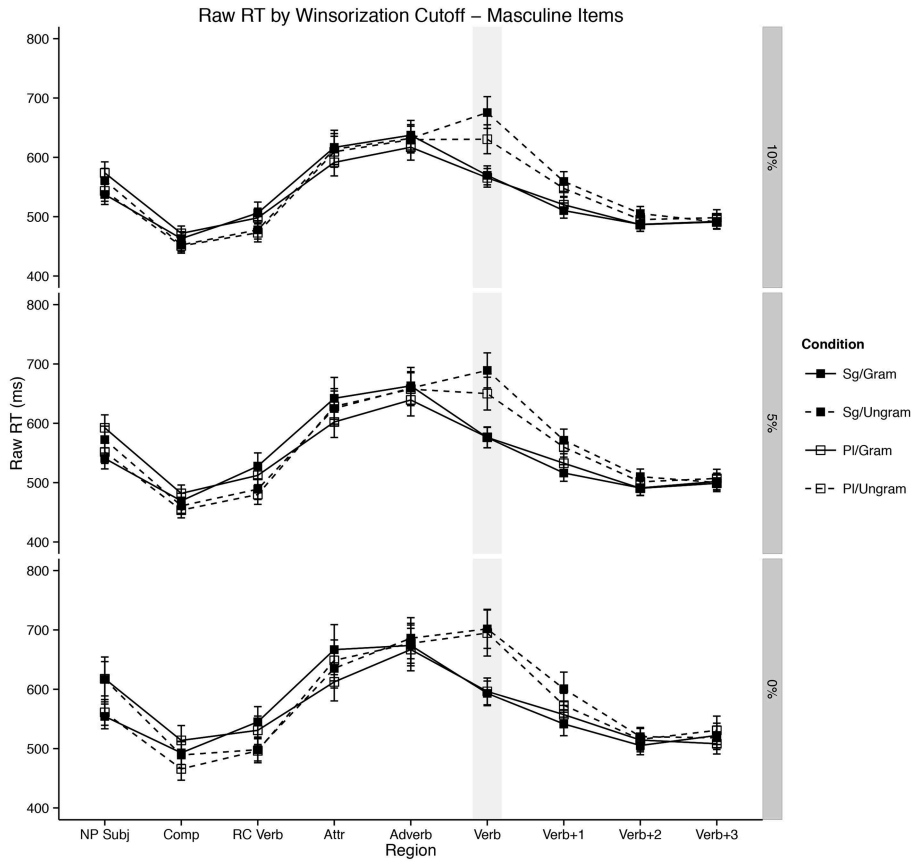


FIGURE 4 | Self-paced reading results. Region by region means of masculine items segregated by attractor number and verb number for three Winsorization cutoff thresholds. Error bars represent the standard error of the mean computed across subject averages.

undergoing an additional reading time decrease when the plural involved is broken/ablauting as opposed to when it is singular—this should increase the magnitude of the attraction effect, not decrease it. Furthermore, in our item set, the difference between mean length of plural and singular items was 1.06 characters for the feminine attractors and 0.72 characters for the masculine attractors, making it hard to specify what role a difference in mean length of 0.3 characters could be playing in a way which accounts for such a large difference between conditions.

Thus, it is reasonable to conclude that nouns with plurals formed by morphologically discontinuous CV-templates *may* drive less agreement attraction, a novel finding in sentence-level reading studies, as far as we know. The split between ablauting and suffixing lends support to the notion that *morphological marking of number* is necessary for agreement attraction to occur in Arabic. The reason for this is that—despite their decreased attraction—broken/ablauting plurals *are* still plurals semantically. Nevertheless, this semantic plurality does not contribute as much as morphological form in driving attraction rates at the critical verb region.

Two things are clear from this limited data set: (1) agreement attraction does occur with attractors in Arabic relative clauses despite the relatively inhospitable grammatical environment relative to non-clausal modifiers such as PPs (Bock and Miller, 1991) and (2) that this effect is modulated by the plural type of the attractor[10]. An immediate follow-up experiment present itself for which preparations are underway: a direct manipulate the gender of the attractor independent of number in order to confirm the argumentation that gender is not the relevant effect in this data.

4. Computational Modeling

Since we take the procedural implementation of agreement dependency resolution to be universal, the important question thus becomes what drives language-specific differences in error profiles and what impact, if any, our findings have on theoretical explanations of agreement attraction. Here we discuss whether or not working-memory models of attraction provide a mechanism for explaining the contrast between broken/ablauting and sound/suffixing plurals seen in our data. The question is one of representation: do explicit models of agreement attraction as

[10]One issue which we have not addressed here is the possibility raised by Gillespie and Pearlmutter (2013) that the inhospitability of relative clauses for attraction is due to lack of consideration for the semantic weight of the relative clause verb and not, say, structural or length differences between relative clauses and PPs. This interpretation is possible for our results and we hope to return to it in future work.

working memory retrieval errors provide a representational way to model the distinction between ablauting and suffixing plurals?

We answer this question by way of computational modeling in the ACT-R system of language comprehension presented by Lewis and Vasishth (2005). Given that well-specified computational models of the sentence processor exist, computational modeling can allow us to evaluate different representational commitments against the results of a system known to accurately model many aspects of working memory and sentence processing. We use ACT-R in particular because of its recent popularity in the sentence processing literature (see Lewis and Vasishth, 2005; Lewis et al., 2006; Dillon et al., 2013) and its requirement that modelers be explicit about representational commitments made for constituents in memory.

One key feature of these models that we believe is implicated by our data is the notion of activation as a zero-sum game across specified retrieval cues. In the ACT-R system, the strength of a particular retrieval cue is proportional to the logarithm of the number of items associated with that cue (Lewis and Vasishth, 2005, p. 381). Given this relationship, one can reduce the strength of, e.g., the number cue for a particular chunk, by removing that chunk's specification for the number cue. This in turn increases the strength of that cue for other chunks in memory which remain specified for number. The result is a reduction in the error rates and retrieval latency intrusion effect size.

In our data, one might therefore consider modeling the ablauting plural attractors with *underspecification*. Underspecification is an approach to the organization of the lexicon wherein certain grammatical features are not present at the lexical level of representation[11]. This approach would therefore remove number from the ablauting attractors and therefore increase the strength of this cue for the true subject. This is a common strategy in the ACT-R language literature for modeling disappearing and reappearing intrusion effects—see (Dillon et al., 2013) for discussion and references in the context of the difference between agreement and reflexive anaphora.

In order to test this idea with Arabic nouns, we need to make some preliminary assumptions. The first of these is that the traditional approach to Arabic grammar which organizes the lexicon in terms of consonantal roots which associate with prosodic templates (see McCarthy, 1981 for the generative approach)[12]. Furthermore, we assume that the prosodic template, despite being morphophonologically abstract, can bear grammatical information for the system—in the case of Arabic nouns, the key feature will be that the template can bear the functional load of number. Finally, we assume that the parser gives access to some form of root/template decomposition during reading, though we will remain agnostic as to the exact mechanism by which this happens.

With this background in mind, we can now ask whether underspecification of grammatical number on the template is an appropriate way to model our data from our ablauting items. Here we consider three distinct models which differ only on their representation of grammatical number as a cue to retrieval:

(9) a. A **fully-specified** model wherein number is a bivalent cue which can take two values: singular and plural

b. A **underspecified NP** model in which nouns appearing in broken plural templates have no specification for number.

c. A **fully-underspecified** model in which number is a fully privative cue that has only one value: plural

The fully-specified model (9a) is meant as a control, a model which accounts for the suffixing data in Arabic and against which we can compare two possible models of ablauting templates. The two models in (9b–c) are two different ways of modeling underspecification in ACT-R, and the viability of either model is the modeling result of interest.

In both the underspecified models (9b–c), underspecification is represented by the absence of a number cue on one or more constituents in memory. In the Underspecified NP model (9b), only NPs which are part of the broken/ablauting plural system lack a number cue; in the Fully Underspecified model (9c), singular verbs *also* lack a number cue. The model in (9c) therefore corresponds to a fully privative number cue system. In either of the underspecified models, representation of a sound/suffixing plural noun simply requires specifying that NP as plural.

To evaluate these models, we ran 10,000 Monte Carlo simulations in ACT-R of each of the four conditions in our experiment with each of the three models (using code first written for Badecker and Kuminiak, 2007; Badecker and Lewis, 2007). ACT-R has several free parameters which must be specified, such as the amount of activation noise present in the system. Instead of computing results across different parameter sets, these parameters were set to the most common values found in the Wong et al. (2010) Online Database of ACT-R Estimated Parameters. While this approach does not provide an argument for the robustness of our results across different parameter values, it does provide for model results using the most neutral parameter specifications.

Our interest in the ACT-R model is in the predictions it makes with respect to a retrieval event triggered at the critical verb which searches for the correct controller of agreement. The model itself provides two dependent measures of interest: (1) the rate of retrieval of each constituent chunk in memory and (2) the latency of retrieval predicted by the model. Both of these dependent measures depend on the schedule of retrievals inputted to the model, which for us included: (1) a retrieval which searches for a NP host for the relative clause at the complementizer, (2) a retrieval which searches for the subject of the embedded clause verb, and (3) a retrieval which searches for a subject of the main clause target verb. (3) is the critical retrieval for us, and all quantitative results we report concern this retrieval.

Figures 6, **7** show activation time-course plots for two of the conditions in our experiment, Pl/Gram and Pl/Ungram, in the Fully Specified model. Pl/Ungram is the attraction condition and Pl/Gram is a control insofar as it involves the same attractor/subject configuration but should yield no attraction. As can be seen in **Figure 6**, the retrieval triggered at the singular

[11]This idea dates back to Trubetskoy's (1958) conception of the lexicon, and the modern instantiation of this notion first appears in Halle (1959).

[12]However, since our data requires no commitment to the position of vowels in this model, we will assume they are part of the prosodic template itself, *contra* (McCarthy, 1981).

main clause verb involves no increased activation of the attractor, whereas when the verb is plural (**Figure 7**), the attractor receives a boost in activation which corresponds to the attraction effect. This is a correct result for agreement attraction insofar as the increased activation translates to a proportional increase in incorrect retrievals of the attractor.

Turning now to error rates, **Table 4** shows the percentage of retrievals in which the true subject or attractor is retrieved at the main clause verb across the three model types[13]. Crucially, **Table 4** shows a marked increase in error rates in the Pl/Ungram condition in the Fully Specified model; this is a predicted attraction effect. Notably, however, either of the underspecification models cause this attraction rate to fall off considerably, decreasing from 24.15 to 6.30% in the Underspecified NP model and from 24.15 to 5.88% for the fully underspecified model. Moreover, in both the underspecified models the error rate is flat across all four experimental conditions. This is a prediction of no or very little attraction effect in for broken/ablauting plurals in these models.

Moving to predictions more analogous to our results, the ACT-R model also furnishes latencies to retrieval of any chunk, and these latencies can be used to predict the size of the agreement attraction or intrusion effect, exactly as is shown in **Figure 5** for our data. **Figure 8** shows the predictions of the ACT-R model across the three model types in (9). Starting with the Fully Specified Model, **Figure 8** shows that the system predicts a large intrusion effect for ungrammatical utterances, exactly as we observe in our data. Turning to the two underspecification models, we observe a flattening of this effect across grammaticality. Both the Underspecified NP and Fully Underspecified models predict no obvious difference between grammatical and ungrammatical conditions.

How one views the success of these modeling results depends on one's interpretation of the empirical results in our study. If one assumes that the masculine/broken/ablauting plural attractors cause a *smaller* attraction effect than the feminine/sound/suffixing attractors, then these modeling results point to a weakness in the representational commitments or architecture of the computational model. Specifically, the dependency between number of items associated with a cue and cue strength means that the reduction in cue strength given by underspecification is all-or-nothing. What is required here is another mechanism for allowing cue strength to modulate in a continuous way—one possibility couched entirely in the memory architecture is to assume that the number cues present on broken/ablauting plurals are somehow different from the number cues present on sound/suffixing plurals in a way which leads the number cues on masculine items to be confusable (Jäger et al.,

[13]Note that these rates do not sum to 100 because the two NPs are not the only constituents in memory, and some small percentage of the time the system retrieves a nonsensical constituent such as a VP or CP. We ignore those results here.

TABLE 4 | ACT-R Retrieval rates for the target/attractor NPs in 10,000 model runs.

	Condition			
Model Type	**Sg/Gram**	**Sg/Ungram**	**Pl/Gram**	**Pl/Ungram**
Fully Specified	90.82/6.15%	90.90/6.29%	94.95/1.63%	74.07/24.15%
Underspecified NP	95.09/1.64%	91.07/6.53%	94.65/2.00%	90.85/6.30%
Fully Underspecified	93.90/4.63%	91.99/5.60%	93.50/4.98%	91.89/5.88%

A fully specified model takes plurality to be a bivalent cue with possible values [SG] and [PL]; a NP underspecified model considers only ablauting plurals as underspecified for number; a fully underspecified model considers all non-plural constituents underspecified for number.

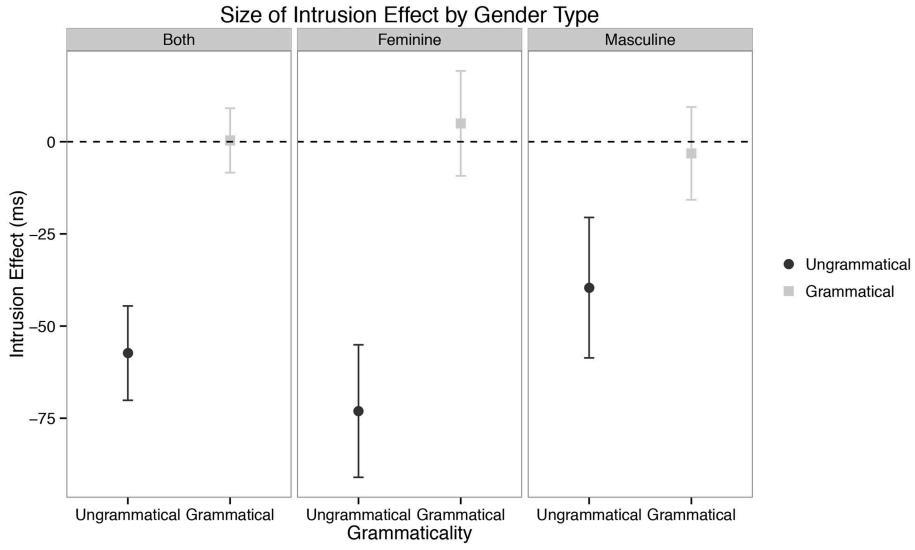


FIGURE 5 | Actual sizes of intrusion effects across different grammaticalities and gender/plural types in the self-paced reading data. Intrusion effects are computed by subtracting retrieval times in plural attractor conditions from times in singular attractor conditions. Error bars represent the standard error of the mean computed across subject averages.

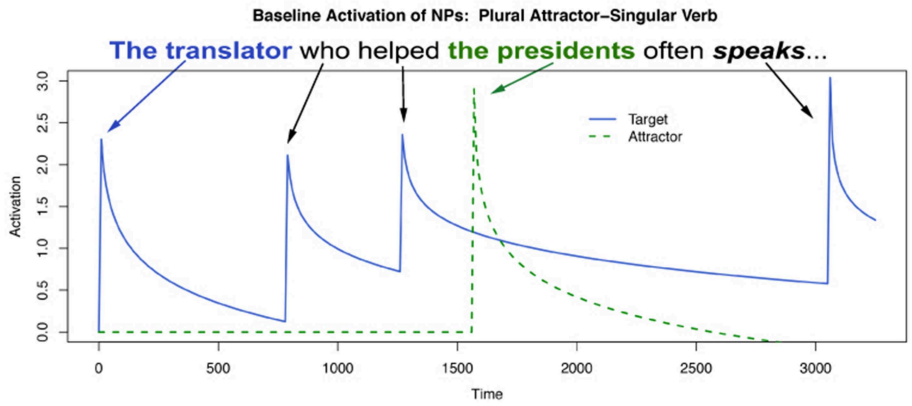


FIGURE 6 | Model activation values over time for an ACT-R model of grammatical attraction conditions.

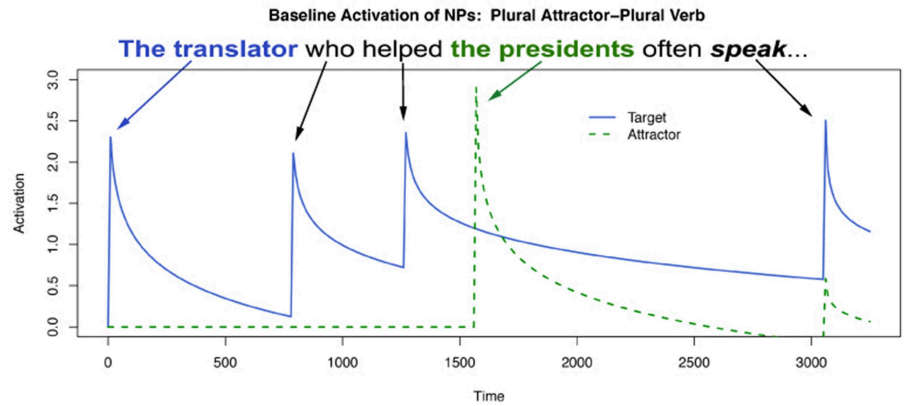


FIGURE 7 | Model activation values over time for an ACT-R model of ungrammatical attraction conditions.

2014) with the number cue present on the main clause verb. In a system where matching is not all-or-nothing, a confusable number cue on the masculine items would lead to lower, but not completely absent, rates of attraction. We do not implement this here because of the relative novelty of the confusability proposal as well as the focus of this paper, but note that it is an intriguing possibility.

On the other hand, if one views our results as showing that masculine/broken/ablauting attractors lead to a complete *absence* of attraction effects, then our modeling exercise here can be taken to show the limited utility of underspecification in the retrieval system (but not the grammar, or the mapping between features and cues; see below) Specifically, our results would show that the model matches the data reasonably well if one assumes that underspecification is the operative difference between ablauting and suffixing plurals insofar as the former are underspecified for number. However, it is important to step back and ask why this is: it is not surprising that a model taught to ignore number (via the absence of number cues on masculines) would yield results that are invariant for plurality. The question should be whether such a state of affairs is congruent with theories of grammar or the mapping between representations used in grammar and representations used in processing[14]. This, however, is a significant shift in perspective: it requires examining the consequences of assuming that number features are not fully specified on ablauting plurals. However, this is not a innocuous assumption as it requires complicating the relationship between grammatical features and retrieval cues; we return to it in the general discussion.

5. General Discussion

5.1. Universal Procedural Components

The results of our study suggest that agreement comprehension errors occur in Arabic similar in broad strokes to the way they occur in Slavic, Romance, and Germanic languages. Specifically, our results show that plural attractors in subject relative clauses can spuriously attract agreement in such a way that an erroneously plural verb can be read as grammatical some percentage of the time. This is an especially striking result given

[14]We thank an anonymous reviewer for forcing us to be clearer about this point.

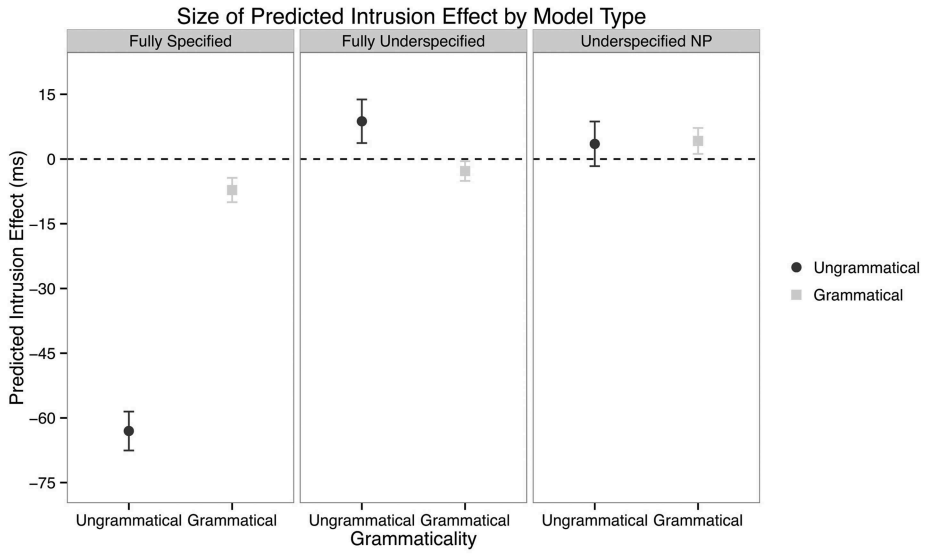


FIGURE 8 | Predicted sizes of intrusion effects for three different ACT-R models. Intrusion effects are computed by subtracting retrieval times in plural attractor conditions from times in singular attractor conditions. Error bars represent the standard error of the mean of 10,000 Monte Carlo trials. A normal model takes number cues to be bivalent, an underspecified NP model takes number to be privative on the broken plural template NPs only, and a fully underspecified model takes number to be privative on all constituents.

two properties of our stimuli which mitigate against high attraction rates: 1) the use of a distractor in a relative clause and 2) the agreeing status of the complementizer in MSA. While other studies such as Dillon et al. (2013) have demonstrated that relative clauses can still contribute to attraction at a possibly lower rate, combining the presence of these relative clauses with the disambiguating cues provided by the complementizer yields an environment where one could imagine that error rates are driven to floor or precluded altogether; nevertheless, this is not what happens in MSA.

The first property has historically been shown to drive down error rates, starting with the original study by Bock and Miller (1991) (see also Bock and Cutting, 1992). In their Experiment 2 using single-clause stimuli where the attractor was contained inside a modifying prepositional phrase, the error rate was 2.39% vs. an error rate of 1.80% for Experiment 3 using bi-clausal stimuli with the attractor inside the embedded relative clause. Note, additionally, that error rates are low across the board due to the sentence-completion task employed in that study, a task which usually results in very low error rates. There are numerous ways to model such a near-halving of the error rates, but a common approach is to assume that subject-hood or clause-mate status is relevant for cue-based retrieval—when this feature is shared by the true subject and critical verb, activation of the correct NP is boosted at the cost of the activation of the attractor NP.

This activation benefit of the true subject conferred by the clause-mate is augmented by the fact that complementizers in MSA necessarily agree in grammatical number and gender with definite NPs to which they are attached or with which the gap in the relative clause is co-construed (Ryding, 2005, pp. 322–323). If one assumes that constituent activation levels are augmented when a relative clause is attached to a head noun, then this must occur at the complementizer position in MSA. Moreover, this retrieval event necessarily includes a number (and gender) cue, unlike the equivalent retrieval event in languages without an inflecting complementizer, such as English. Thus, by the time the critical verb is encountered, more temporal decay of activation of the true subject should have occurred in English than in Arabic. This, in turn, should imply smaller error rates in Arabic than in English, since the number and gender cues on the complementizer reinforce the activation of the proper subject NP. Even in model-neutral terms, something like this is expected, since the bare fact is that the complementizer provides the speaker with additional cues to the proper subject in Arabic, but not in English.

Nevertheless, these potentially mitigating factors did not result in a complete absence of agreement attraction errors in Arabic. In both perfect and imperfect aspect, each with distinct verbal agreement affixes, attraction by the plural attractors is clearly evident in the decreased reaction time in the ungrammatical sentences with plural attractors which match the verb relative to cases where the attractor does not match the verb. Such a result is consistent with the working-memory model of agreement attraction which views the procedural underpinning of these errors as a universal part of the language comprehension system. In this model, decreased reaction times in ungrammatical sentences with matching attractors are driven by partial cue overlap between the attractor and the erroneous verb—a result which is driven by the fundamental architecture of the memory system.

5.2. Representing Plurality

However, one marked difference between our data and those reported for other languages is the role of the morpheme which carries plural marking. Our results indicate a difference between plurals formed by suffixation and those formed by ablaut in

the size or presence of the attraction effect. This is a novel observation in the agreement attraction literature for any language, as far as we are aware. In §4 we noted that the representational implications of this result for the retrieval system depend in part on the assessment of its nature: if the difference is just one of quantity, then complications could be made to the way that plurality is represented solely in the memory system (*cf.* the discussion of cue confusability). If, on the other hand, one takes the difference to be qualitative, then complications need to be in the feature-cue mapping algorithm. It is that complication which we explore in this section.

In the case where the difference between feminine/sound/suffixing and masculine/broken/ablauting plural attractors is taken to be large, then one needs to articulate the way in which grammatical features map onto retrieval cues. Recall that underspecifying a broken plural for a plural feature results in a model which matches data in which no attraction occurs in broken plural items reasonably well. However, the question would then be how to articulate the relationship between grammatical plurality and the absence of an effect of plurality in the retrieval system. One simplistic option would be to claim that semantic plurality does not not contribute to retrieval interference for plural cues expressed morphologically on the target verb. However, this simple idea is unlikely to be the entire story given the finding in the literature that semantic overlap contributes to retrieval interference elsewhere (for a close parallel to our data in English, see Bock and Eberhard, 1993; for an overview of semantic contributions to interference, see the overview and references in Van Dyke and Johns, 2012). A more nuanced view would take features from various components of formal grammar to contribute additively to cues in the retrieval system[15].

In an additive approach, one could imagine that semantic and morphological features combined underwrite what is ultimately expressed as a plural cue in on NP chunks in memory. One could then specify that the morphology of these broken plurals contributes less to the sum that ultimately makes up plural cue values for constituents in memory such that broken plurals appear to the memory system as less plural than the sound plurals. Our results would then speak to the nature of the weights of various featural components that underwrite such an additively composite cue such that semantic plurality alone is not sufficient to yield a plural cue that causes measurable agreement attraction. Underspecification can then be seen as coherent insofar as only the morphological component of an additive cue is underspecified. This allows for a way to understand our data in a theoretically meaningful way.

More generally and independently of any one interpretation of the models of feature-cue mapping involved, the results here suggest some important conclusions about morphological representation in general and Arabic templates in particular. Regardless of the model-theoretic interpretation of these results, one fact is clear: a discontinuous and/or abstract morphological constituent modulates error rates related to plural nouns in Arabic. This is important because it underscores the morphological contribution of discontinuous alterations in form in the language. Not only is this a point which is important for language-independent theorizing, but it is a point currently under contention in the Semitic-specific priming literature. The results of this study suggest that whatever the correct representational view of the CV-template is, it must minimally be allowed to augment plurality-driven effects in reading comprehension.

[15]We thank an anonymous reviewer for this intriguing suggestion.

6. Conclusion

We have demonstrated that agreement attraction errors exist in MSA in a configuration which is relatively inhospitable to the presence of such mistakes: subject relative clauses with an agreeing complementizer in a morphologically rich language. Furthermore, we showed that MSA, like English, has a plural complexity cost associated with reading suffixed plural NPs. However, we also showed that MSA differs from English in important ways concerning the nature of plural formation. Specifically, we showed that plurals formed by suffixation strongly attract agreement, whereas plurals formed by ablaut/internal vowel change do so at greatly reduced rates, if at all. Moreover, we have suggested that Arabic also provides evidence that agreement attraction effects are driven mostly by observations in the right tail of the reaction time distribution. Finally, we have provided model evidence which suggests that morphologically discontinuous plural forms in MSA require some elaboration of the way grammatical features are translated into processing cues for the memory retrieval system. Finally, we discussed how these results suggest a somewhat form-driven comprehension mechanism for agreement resolution, provided that such a model allows discontinuous form-based differences to modulate comprehension of agreement dependencies.

Acknowledgments

The authors wish to thank Eias Al-Daman, Souad Al-Helou, Meera Saeed Al-Kaabi, Esma Mansouri, and Samer Nehme for assistance with stimuli creation and Tommi Leung for help with participant recruitment. Thanks to Joseph King, Stephen Politzer-Ahles, Kevin Schluter, Shravan Vasishth, Matt Wagers, audiences at the 20th Architectures and Mechanisms for Language Processing (AMLaP) in Edinburgh, Scotland and the 2015 Annual Meeting of the Linguistic Society of America for comments on previous versions of this work. Additionally, thanks to two anonymous reviewers for comments which led to a considerable improvement in this paper. Finally, thanks to Rick Lewis for the creation of the code used for ACT-R modeling, Brian Dillon for sharing his modifications thereto, and Dave Kush for substantial help with debugging. None of these people are responsible for errors in this manuscript, the fault of which lies solely with the authors.

Supplementary Material

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fpsyg.2015.00347/abstract

References

Antón-Méndez, I., Nicol, J. L., and Garrett, M. F. (2002). The relation between gender and number agreement processing. *Syntax* 5, 1–25. doi: 10.1111/1467-9612.00045

Baayen, R., Davidson, D., and Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *J. Mem. Lang.* 59, 390–412. doi: 10.1016/j.jml.2007.12.005

Badecker, W., and Kuminiak, F. (2007). Morphology, agreement and working memory retrieval in sentence production: evidence from gender and case in Slovak. *J. Mem. Lang.* 56, 65–85. doi: 10.1016/j.jml.2006.08.004

Badecker, W., and Lewis, R. (2007). "A new theory and computational model of working memory in sentence production: agreement errors as failues of cue-based retrieval," in *Paper Presented at the 20th Annual CUNY Sentence Processing Conference* (San Diego, CA: La Jolla).

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: keep it maximal. *J. Mem. Lang.* 68, 255–278. doi: 10.1016/j.jml.2012.11.001

Bock, K., and Cutting, J. (1992). Regulating mental energy: performance units in language production. *J. Mem. Lang.* 31, 99–127. doi: 10.1016/0749-596X(92)90007-K

Bock, K., and Eberhard, K. M. (1993). Meaning, sound and syntax in English number agreement. *Lang. Cogn. Process.* 8, 57–99. doi: 10.1080/01690969308406949

Bock, K., Eberhard, K. M., Cutting, J. C., Meyer, A. S., and Schriefers, H. (2001). Some attractions of verb agreement. *Cogn. Psychol.* 43, 83–128. doi: 10.1006/cogp.2001.0753

Bock, K., and Miller, C. (1991). Broken agreement. *Cogn. Psychol.* 23, 45–93. doi: 10.1016/0010-0285(91)90003-7

Boudelaa, S., and Marslen-Wilson, W. D. (2001). Morphological units in the arabic mental lexicon. *Cognition* 81, 65–92. doi: 10.1016/S0010-0277(01)00119-6

Boudelaa, S., and Marslen-Wilson, W. D. (2004a). Abstract morphemes and lexical representation: the CV-skeleton in Arabic. *Cognition* 92, 271–303. doi: 10.1016/j.cognition.2003.08.003

Boudelaa, S., and Marslen-Wilson, W. D. (2004b). Allomorphic variation in arabic: implications for lexical processing and representation. *Brain Lang.* 90, 106–116. doi: 10.1016/S0093-934X(03)00424-3

Boudelaa, S., and Marslen-Wilson, W. D. (2005). Discontinuous morphology in time: incremental masked priming in Arabic. *Lang. Cogn. Process.* 20, 207–260. doi: 10.1080/01690960444000106

Boudelaa, S., and Marslen-Wilson, W. D. (2011). Productivity and priming: morphemic decomposition in Arabic. *Lang. Cogn. Process.* 26, 624–652. doi: 10.1080/01690965.2010.521022

Boudelaa, S., Pulvermüller, F., Hauk, O., Shtyrov, Y., and Marslen-Wilson, W. (2010). Arabic morphology in the neural language system. *J. Cogn. Neurosci.* 22, 998–1010. doi: 10.1162/jocn.2009.21273

Chomsky, N. (1995). *The Minimalist Program*. Cambridge, MA: MIT Press.

Chomsky, N. (2000). "Minimalist inquiries," in *Step by Step: Essays on Minimalist Sytnax in Honor of Howard Lasnik, Chapter 3,* eds R. Martin, D. Michaels, and J. Uriagereka (Cambridge, MA: MIT Press), 89–156.

Chomsky, N. (2001). "Derivation by phase," in *Ken Hale: A Life in Language*, ed M. Kenstowicz (Cambridge, MA: MIT Press), 1–52.

Den Dikken, M. (2001). Pluringulars, pronouns and quirky agreement. *Linguist. Rev.* 18, 19–41. doi: 10.1515/tlir.18.1.19

Deutsch, A., Frost, R., and Forster, K. (1998). Verbs and nouns are organized and accessed differently in the mental Lexicon: evidence from Hebrew. *J. Exp. Psychol. Learn. Mem. Cogn.* 24, 1238–1254. doi: 10.1037/0278-7393.24.5.1238

Dillon, B., Mishler, A., Sloggett, S., and Phillips, C. (2013). Contrasting intrusion profiles for agreement and anaphora: experimental and modeling evidence. *J. Mem. Lang.* 69, 85–103. doi: 10.1016/j.jml.2013.04.003

Eberhard, K., Cutting, J., and Bock, K. (2005). Making syntax of sense: number agreement in sentence production. *Psychol. Rev.* 112, 531–559. doi: 10.1037/0033-295X.112.3.531

Fayol, M., Largy, P., and Lemaire, P. (1994). When cognitive overload enhances subject–verb agreement errors: a study in French written language. *Q. J. Exp. Psychol.* 47, 437–464. doi: 10.1080/14640749408401119

Francis, W. N. (1986). Proximity concord in English. *J. Engl. Linguist.* 19, 309–317. doi: 10.1177/007542428601900212

Franck, J., Lassi, G., Frauenfelder, U., and Rizzi, L. (2006). Agreement and movement: a syntactic analysis of attraction. *Cognition* 101, 173–216. doi: 10.1016/j.cognition.2005.10.003

Franck, J., Soare, G., Frauenfelder, U. H., and Rizzi, L. (2010). Object interference in subject-verb agreement: the role of intermediate traces of movement. *J. Mem. Lang.* 62, 166–182. doi: 10.1016/j.jml.2009.11.001

Franck, J., Vigliocco, G., Antón-Méndez, I., Collina, S., and Frauenfelder, U. (2008). The interplay of syntax and form in sentence production: a cross-linguistic study of form effects on agreement. *Lang. Cogn. Process.* 23, 329–374. doi: 10.1080/01690960701467993

Franck, J., Vigliocco, G., and Nicol, J. (2002). Subject-verb agreement errors in French and English: the role of syntactic hierarchy. *Lang. Cogn. Process.* 17, 371–404. doi: 10.1080/01690960143000254

Frost, R., Deutsch, A., and Forster, K. (2000). Decomposing morphologically complex words in a nonlinear morphology. *J. Exp. Psychol. Learn. Mem. Cogn.* 26, 751. doi: 10.1037//0278-7393.26.3.751

Frost, R., Forster, K., and Deutsch, A. (1997). What can we learn from the morphology of Hebrew? A masked-priming investigation of morphological representation. *J. Exp. Psychol. Learn. Mem. Cogn.* 23, 829–856. doi: 10.1037/0278-7393.23.4.829

Gillespie, M., and Pearlmutter, N. J. (2013). Against structural constraints in subject–verb agreement production. *J. Exp. Psychol. Learn. Mem. Cogn.* 39, 515. doi: 10.1037/a0029005

Halle, M. (1959). *The Sound Pattern of Russian*. The Hague: Mouton & Company.

Hartsuiker, R. J., Antón-Méndez, I., and van Zee, M. (2001). Object attraction in subject-verb agreement constructions. *J. Mem. Lang.* 45, 546–572. doi: 10.1006/jmla.2000.2787

Hartsuiker, R. J., Kolk, H. H. J., and Huinck, W. J. (1999). Agrammatic production of subject–verb agreement: the effect of conceptual number. *Brain Lang.* 69, 119–160. doi: 10.1006/brln.1999.2059

Hartsuiker, R. J., Schriefers, H., Bock, K., and Kikstra, G. (2003). Morphophonological influences on the construction of subject-verb agreement. *Mem. Cogn.* 31, 1316. doi: 10.3758/BF03195814

Haskell, T., and MacDonald, M. (2005). Constituent structure and linear order in language production: evidence from subject-verb agreement. *J. Exp. Psychol. Learn. Mem. Cogn.* 31, 891–904. doi: 10.1037/0278-7393.31.5.891

Häussler, J. (2009). *The Emergence of Agreement Attraction Errors During Sentence Comprehension.* Ph.D. thesis, Universität Konstanz, Konstanz, Germany.

Jäger, L., Engelmann, F., and Vasishth, S. (2014). "Inhibitory interference in reflexives: evidence for cue confusability," in *Poster Presnted at the 20th Architectures and Mechanisms for Language Processing Conference (AMLaP),* (Edinburgh, Scotland).

Jespersen, O. (1924). *The Philosophy of Grammar*. London: Allen & Unwin.

Just, M., Carpenter, P., and Woolley, J. (1982). Paradigms and processes in reading comprehension. *J. Exp. Psychol.* 111, 228–238.

Kaan, E. (2002). Investigating the effects of distance and number interference in processing subject-verb dependencies: an ERP study. *J. Psycholinguist. Res.* 31, 165–193. doi: 10.1023/A:1014978917769

Kayne, R. S. (2000). *Parameters and Universals.* New York, NY: Oxford UP.

Kimball, J., and Aissen, J. (1971). I think, you think, he think. *Linguist. Inq.* 2, 241–246.

Kuznetsova, A., Bruun Brockhoff, P., and Haubo Bojesen Christensen, R. (2014). *lmerTest: Tests for Random and Fixed Effects for Linear Mixed Effects Models.*

Lago, S., Shalom, D. E., Sigman, M., Lau, E. F., and Phillips, C. (2014). Agreement in spanish comprehension. MS., University of Maryland and University of Buenos Aires.

Lewis, R. L., and Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cogn. Sci.* 29, 375–419. doi: 10.1207/s15516709cog0000_25

Lewis, R. L., Vasishth, S., and Van Dyke, J. A. (2006). Computational principles of working memory in sentence comprehension. *Trends Cogn. Sci.* 10, 447–454. doi: 10.1016/j.tics.2006.08.007

Lorimor, H., Bock, K., Zalkind, E., Sheyman, A., and Beard, R. (2008). Agreement and attraction in Russian. *Lang. Cogn. Process.* 23, 769–799. doi: 10.1080/01690960701774182

MacWhinney, B. (1987). "Cues, competition, and learning," in *Wisconsin Papers on Language Development*, ed J. Miller (Madison, WI: University of Wisconsin Press).

MacWhinney, B. (2001). "The competition model: the input, the context, and the brain," in *Cognition and Second Language Instruction*, ed P. Robinson (New York, NY: Cambridge University Press), 69–90.

MacWhinney, B., and Bates, E. (1989). *The Cross-Linguistic Study of Sentence Processing*. New York, NY: Cambridge University Press.

McCarthy, J. J. (1979). *Formal Problems in Semitic Phonology and Morphology*. Ph.D. thesis, Massachusetts Institute of Technology.

McCarthy, J. J. (1981). A prosodic theory of nonconcatenative morphology. *Linguist. Inq.* 12, 373–418.

McCarthy, J. J., and Prince, A. (1990a). "Prosodic morphology and templatic morphology," in *Perspectives on Arabic Linguistics II*, eds M. Eid and J. J. McCarthy (Amsterdam: John Benjamins), 1–54.

McCarthy, J. J., and Prince, A. S. (1990b). Foot and word in prosodic morphology: the Arabic broken plural. *Nat. Lang. Linguist. Theory* 8, 209–283.

Meyer, A. S., and Bock, K. (1999). Representations and processes in the production of pronouns: some perspectives from Dutch. *J. Mem. Lang.* 41, 281–301.

Mimouni, Z., Kehayia, E., and Jarema, G. (1998). The mental representation of singular and plural nouns in Algerian Arabic as revealed through auditory priming in agrammatic aphasic patients. *Brain Lang.* 61, 63–87.

Nicol, J. L., Forster, K. I., and Veres, C. (1997). Subject–verb agreement processes in comprehension. *J. Mem. Lang.* 36, 569–87.

Parker, D. (2014). *The Cognitive Basis for Encoding and Navigating Linguistic Structure*. Ph.D. thesis, University of Maryland at College Park.

Parkinson, D. B. (2012). ArabiCorpus. Available online at: http://arabicorpus.byu.edu/

Pearlmutter, N. J., Garnsey, S. M., and Bock, K. (1999). Agreement processes in sentence comprehension. *J. Mem. Lang.* 41, 427–456.

Quirk, R., Greenbaum, S., Leech, G., and Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. New York, NY: Longman.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychol. Bull.* 114, 510–532.

Rhode, D. (2003). *Linger: A Flexible Platform for Language Processing Experiments, Version 2.94.* Cambridge, MA: MIT Press.

Ryding, K. C. (2005). *A Reference Grammar of Modern Standard Arabic*. Cambridge, UK: Cambridge UP.

Staub, A. (2009). On the interpretation of the number attraction effect: response time evidence. *J. Mem. Lang.* 60, 308–327. doi: 10.1016/j.jml.2008.11.002

Staub, A. (2010). Response time distributional evidence for distinct varieties of number attraction. *Cognition* 114, 447–454. doi: 10.1016/j.cognition.2009.11.003

Tanner, D., Nicol, J. L., and Brehm, L. (2014). The time-course of feature interference in agreement comprehension: multiple mechanisms and asymmetrical attraction. *J. Mem. Lang.* 76, 195–215. doi: 10.1016/j.jml.2014.07.003

Thornton, R., and MacDonald, M. C. (2003). Plausibility and grammatical agreement. *J. Mem. Lang.* 48, 740–759. doi: 10.1016/S0749-596X(03)00003-2

Trubetskoy, N. (1939/1958). *Grundzüge der Phonologie*. Göttingen: Vandenhoeck & Ruprecht.

Tucker, M. A. (2011). "The morphosyntax of the Arabic verb," in *Morphology at Santa Cruz: Papers in Honor of Jorge Hankamer*, eds N. LaCara, A. Thompson, and M. A. Tucker (Santa Cruz, CA: LRC Publications), 177–211.

Tucker, M. A. (2010). Roots and prosody: the Iraqi Arabic derivational verb. *Res. Linguist. Vincennes Racine Radic.* 39, 31–68.

Tucker, M. A., and Wagers, M. W. (2010). "Agreement attraction in infinitival complements," in *UCSC LRC Laboratory Report 2010*, ed G. McGuire (Santa Cruz, CA: LRC Publications), 23–43.

Ussishkin, A. (2000). *The Emergence of Fixed Prosody*. Ph.D. thesis, University of California, Santa Cruz.

Ussishkin, A. (2005). A fixed prosodic theory of nonconcatenative templatic morphology. *Nat. Lang. Linguist. Theory* 23, 169–218. doi: 10.1007/s11049-003-7790-8

Ussishkin, A., Dawson, C., Wedel, A., and Schluter, K. (2015). Auditory masked priming in Maltese spoken word recognition. *Lang. Cogn. Neurosci.* doi: 10.1080/23273798.2015.1005635

Ussishkin, A., and Twist, A. (2007). "Auditory and visual lexical decision in Maltese," in *Introducing Maltese Linguistics: Papers from the 1st International Conference on Maltese Linguistics*, eds B. Comrie, R. Fabri, E. Hume, M. Mifsud, T. Stolz, and M. Vanhove (Amsterdam; Philadelphia: John Benjamins Publishing Company), 233–249.

Ussishkin, A., Wedel, A., Schluter, K., and Dawson, C. (2011). "Supraliminal and subliminal root and binyan priming in Maltese," in *Talk Presented at GĂDÂęILM 3rd Conference on Maltese Linguistics* (Valletta: University of Malta).

Van Dyke, J. A., and Johns, C. L. (2012). Memory interference as a determinant of language comprehension. *Lang. Linguist. Compass* 6, 193–211. doi: 10.1002/lnc3.330

Vigliocco, G., Butterworth, B., and Garrett, M. F. (1996). Subject-verb agreement in Spanish and English: differences in the role of conceptual constraints. *Cognition* 61, 261–298. doi: 10.1016/S0010-0277(96)00713-5

Vigliocco, G., Butterworth, B., and Semenza, C. (1995). Constructing subject-verb agreement in speech: the role of semantic and morphological factors. *J. Mem. Lang.* 34, 186–215. doi: 10.1006/jmla.1995.1009

Vigliocco, G., and Franck, J. (1999). When sex and syntax go hand in hand: gender agreement in language production. *J. Mem. Lang.* 40, 455–478.

Vigliocco, G., and Franck, J. (2001). When sex affects syntax: contextual influences in sentence production. *J. Mem. Lang.* 45, 368–390. doi: 10.1006/jmla.2000.2774

Vigliocco, G., and Nicol, J. (1998). Separating hierarchical relations and word order in language production: is proximity concord syntactic or linear? *Cognition* 68, B13–B29.

Wagers, M. W., Lau, E. F., and Phillips, C. (2009). Agreement attraction in comprehension: representations and processes. *J. Mem. Lang.* 61, 206–237. doi: 10.1016/j.jml.2009.04.002

Wong, T. J., Cokely, E. T., and Schooler, L. J. (2010). "An online database of ACT-R parameters: towards a transparent community-based approach to model development," in *Proceedings of the 10th International Conference on Cognitive Modeling*, eds D. D. Salvucci and G. Gunzelmann (Philadelphia, PA: Drexel UP), 282–286.

Wright, W. (1889a). *A Grammar of the Arabic Language*, Vol. I. Cambridge: Cambridge University Press.

Wright, W. (1889b). *A Grammar of the Arabic Language*, Vol. II. Cambridge: Cambridge University Press.

Zandvoort, R. (1961). *Varia Syntactica. Language and Society: Essays Presented to Arthur M. Jensen on his Seventieth Birthday*. Copenhagen: Det Berlingske Bogtrykkeri.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2015 Tucker, Idrissi and Almeida. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 04 November 2016
doi: 10.3389/fpsyg.2016.01651

Gender Agreement Attraction in Russian: Production and Comprehension Evidence

Natalia Slioussar[1,2]* *and Anton Malko*[3]

[1] School of Linguistics, Higher School of Economics, Moscow, Russia, [2] Faculty of Liberal Arts and Sciences, Saint-Petersburg State University, Saint-Petersburg, Russia, [3] Department of Linguistics, University of Maryland, College Park, MD, USA

OPEN ACCESS

Edited by:
Matthew Wagers, University of California, Santa Cruz, USA

Reviewed by:
Yulia Esaulova, University of Potsdam, Germany
Franc Marusic, University of Nova Gorica, Slovenia

***Correspondence:**
Natalia Slioussar
slioussar@gmail.com

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: *21 August 2015*
Accepted: *10 October 2016*
Published: *04 November 2016*

Citation:
Slioussar N and Malko A (2016) Gender Agreement Attraction in Russian: Production and Comprehension Evidence. Front. Psychol. 7:1651. doi: 10.3389/fpsyg.2016.01651

Agreement attraction errors (such as the number error in the example "The key to the cabinets are rusty") have been the object of many studies in the last 20 years. So far, almost all production experiments and all comprehension experiments looked at binary features (primarily at number in Germanic, Romance, and some other languages, in several cases at gender in Romance languages). Among other things, it was noted that both in production and in comprehension, attraction effects are much stronger for some feature combinations than for the others: they can be observed in the sentences with singular heads and plural dependent nouns (e.g.,"The key to the cabinets..."), but not in the sentences with plural heads and singular dependent nouns (e.g., "The keys to the cabinet..."). Almost all proposed explanations of this asymmetry appeal to feature markedness, but existing findings do not allow teasing different approaches to markedness apart. We report the results of four experiments (one on production and three on comprehension) studying subject-verb gender agreement in Russian, a language with three genders. Firstly, we found attraction effects both in production and in comprehension, but, unlike in the case of number agreement, they were not parallel (in production, feminine gender triggered strongest effects, while neuter triggered weakest effects, while in comprehension, masculine triggered weakest effects). Secondly, in the comprehension experiments attraction was observed for all dependent noun genders, but only for a subset of head noun genders. This goes against the traditional assumption that the features of the dependent noun are crucial for attraction, showing the features of the head are more important. We demonstrate that this approach can be extended to previous findings on attraction and that there exists other evidence for it. In total, these findings let us reconsider the question which properties of features are crucial for agreement attraction in production and in comprehension.

Keywords: agreement, gender, attraction, production, comprehension, Russian

1. INTRODUCTION

1.1. The Phenomenon of Agreement Attraction

Grammatical agreement is one of the most basic linguistic operations. It is well-known, however, that it is not always accurate. In the last 20 years many studies have looked at so-called agreement attraction errors, exemplified in (1). In (1a) the verb agrees not with the head of the subject NP *key*[1], but with another, embedded NP *cabinets* (we will further call such NPs "attractors"). In (1b) the verb in a relative clause agrees with the subject of the matrix clause.

(1) a. The key to the cabinets **were** rusty (Bock and Miller, 1991).

b. The musicians who the reviewer **praise** so highly will probably win a Grammy (Wagers et al., 2009).

Agreement attraction errors are observed in spontaneous speech and in well-edited texts. They have also been studied experimentally, mostly in English, but also in French, Spanish, Italian, Dutch, German, and some other languages (Bock and Miller, 1991; Vigliocco et al., 1995, 1996; Pearlmutter et al., 1999; Anton-Mendez et al., 2002; Hartsuiker et al., 2003, to name just a few). The first accounts suggested that the verb simply agrees with the linearly closest noun (Jespersen, 1924; Quirk et al., 1972; Francis, 1986, a.o.). However, later studies demonstrated that agreement attraction is a structural phenomenon. For example, Vigliocco and Nicol (1998) showed that people make attraction errors producing questions, e.g., "Are the helicopter for the flights safe?." Various factors that influence attraction have also been identified. However, the overwhelming majority of studies focused on number agreement in the languages where number has only two values: singular and plural. It is not clear to what extent these results can be generalized to other cases.

In this paper, we analyze subject-predicate gender agreement. Gender attraction has been investigated only in a few studies, and mostly in Romance languages, which have two genders. We report one production and three comprehension experiments on Russian, a language with three genders. To the best of our knowledge, this is the first comprehension study looking at agreement attraction in a non-binary category. Below we present several findings from the research on number agreement, which will be most important for our study, and different accounts of attraction. Next, we review the few existing studies on gender attraction, providing rationale for the present work.

1.1.1. Plural Markedness Effect

In all studied languages, attraction effects were found to be asymmetric. They can be observed when the head is singular, and the attractor is plural [as in (1) above], but are much weaker or virtually non-existent in the opposite configuration. In the majority of agreement attraction studies, this asymmetry is explained in terms of feature markedness. Plural is assumed to be the marked value of number feature[2], and the asymmetry is attributed to the fact that attractors with a marked feature are more disruptive. Hence it is known under the name of "plural markedness effect."

However, the concept of markedness is not widely agreed upon. Different authors adopt different theoretical approaches and different tests to determine marked and unmarked feature values [including frequency, presence of a non-zero affix, default use of a form (e.g., in impersonal sentences), various semantic tests etc.; see Haspelmath, 2006]. It is impossible to evaluate them looking only at singular and plural. To figure out which of these properties may be relevant for the asymmetry between feature values (and whether it makes sense to attribute it to markedness in a particular theoretical framework), it is crucial to look at other features systems. As we will show below, Russian gender is interesting in this respect because the results of different markedness tests do not converge, letting us tease several approaches apart.

1.1.2. Parallel Results in Production and Comprehension

Experimental studies demonstrated that attraction exists not only in production, but also in comprehension. In production it manifests itself as agreement errors. In comprehension attraction errors have been observed to trigger more grammaticality judgment mistakes and to provoke less pronounced effects in reading time and EEG studies than other agreement errors. In other words, people perceive ungrammatical sentences as if they were grammatical or had a minor violation. This is often called a "grammaticality illusion."

The results from production and comprehension are largely parallel (in particular, significant attraction effects are observed only with plural attractors). This is often used to conclude that the mechanism of attraction is the same in both modalities. We will come back to this problem discussing our findings because we did not observe parallelism that we expected based on the previous studies.

1.1.3. Debate on Ungrammaticality Illusions

We just mentioned that in comprehension, attraction causes grammaticality illusions, making ungrammatical sentences more acceptable. Can it also lead to ungrammaticality illusions, and make grammatical sentences less acceptable? For example, if people tend to miss agreement errors in sentences like (2a), do they sometimes see non-existent errors in sentences like (2b)? As we show below, different approaches to attraction make opposing predictions about ungrammaticality illusions, so this is an important question.

(2) a. The key to the cabinets were rusty.

b. The key to the cabinets was rusty.

Several studies (e.g., Nicol et al., 1997; Pearlmutter et al., 1999) suggested that ungrammaticality illusions do arise. However, Wagers et al. (2009) demonstrated that at least on-line findings may be artifactual (they might be due to the fact that processing

[1]Here and further, the following standard symbols are used: N, noun; NP, noun phrase; P, preposition; PP, prepositional phrase; V, verb; M, masculine gender; F, feminine; N, neuter.

[2]Notably, in semantics there is an ongoing debate whether singular or plural is the default (e.g., Sauerland et al., 2005; Farkas and de Swart, 2010).

plural nouns carries an additional cost compared to singular ones, not to any aspects of subject-verb agreement processing). This hypothesis can be tested by analyzing some cases where this problem does not apply, and we do so in the present study looking at gender agreement [3].

1.1.4. The Role of Morphophonology

Hartsuiker et al. (2003) showed that when the form of the attractor is morphologically ambiguous and coincides with nominative, the rate of attraction errors increases. They compared German sentences like (3a,3b). People made more errors in (3a), where the attractor (*die Demonstrationen*) is ambiguous between accusative and nominative, compared to (3b), where the attractor (*den Demonstrationen*) is unambiguously dative. We do not explore the role of morphophonology in the present study, but take this factor into account. Several studies also demonstrated that heads with regular inflections are more resistant to attraction, but no similar effects were observed for attractors (e.g., Bock and Eberhard, 1993; Vigliocco et al., 1995).

(3) a. die Stellungnahme gegen die
the$_{F.NOM.SG}$ position against the$_{F.ACC.PL}$
Demonstrationen
demonstrations

b. die Stellungnahme zu den
the$_{F.NOM.SG}$ position on the$_{DAT.PL}$
Demonstrationen
demonstrations

1.2. Models of Agreement Attraction

There exist two major approaches to agreement attraction. Here they will be referred to as the "representational account" and the "retrieval account." Models that belong to the **representational account** share one crucial assumption: agreement attraction takes place because the mental representation of the number feature on the subject NP is faulty or ambiguous (Nicol et al., 1997; Vigliocco and Nicol, 1998; Franck et al., 2002; Eberhard et al., 2005; Staub, 2009, 2010; Brehm and Bock, 2013). In some models, it is assumed that syntactic features can "percolate" or otherwise move to neighboring nodes: for example, sometimes number features from the embedded NP percolate to the subject NP (which normally has the same number marking as its head).

Another model known as Marking and Morphing (Eberhard et al., 2005) postulates that the number value of the subject NP is a continuum, i.e., it can be more or less plural. For example, if a subject NP contains a singular head and a plural dependent NP it is more plural than a subject NP with a singular modifier. A subject NP that is formally singular, but refers to a collective entity is more plural than the ones referring to singular entities. The more plural the subject NP, the higher the possibility of choosing a plural verb. In such accounts there is no way to avoid ungrammaticality illusions: if the agreement controller can be mis-construed or ambiguous, there is no way to restrict such mis-construals to only ungrammatical sentences. They happen even before we encounter the verb, i.e., even before it is clear whether the sentence is or is not grammatical.

Now let us turn to the **retrieval account** (Solomon and Pearlmutter, 2004; Lewis and Vasishth, 2005; Badecker and Kuminiak, 2007; Badecker and Lewis, 2007; Wagers et al., 2009; Dillon et al., 2013). Research on memory suggests that the amount of material a person can hold in a ready-to-process state is extremely limited (McElree, 2006; Cowan, 2001). Thus, it can be hypothesized that when we reach an agreeing predicate, the subject needs to be reactivated. This reactivation can be done via so-called cue-based retrieval (Lewis and Vasishth, 2005; McElree, 2006): we query the memory with a set of cues (e.g.,"number: plural," "case: nominative" etc.) and select an element that matches the maximum number of cues.

This process is not error-free, and the retrieval account argues that attraction arises at this stage. For example, in a sentence like "The key to the cabinets are rusty" the form of the verb suggests that we need to look for an NP with the features "subject" and "plural." However, no NP perfectly satisfies these conditions: *key* is the subject, but is not plural, and *cabinets* is plural, but is not the subject. It is hypothesized that in such conditions we may mistakenly select the wrong NP . The retrieval approach predicts the absence of ungrammaticality illusions: if a sentence is grammatical, the true subject is a *perfect* match and will always be selected. Thus, unlike in the representational account, there is nothing wrong or ambiguous in the syntactic structure, errors are access failures. Such cases with several elements competing for retrieval are an instance of "retrieval interference." Other examples are discussed in Van Dyke and Johns (2012).

1.3. Studies of Gender Agreement Attraction

Relatively few studies of gender agreement attraction have been conducted so far. Their results do not always converge, but one thing seems to be certain: attraction effects are present. They have been observed in several experiments on different languages.

1.3.1. Previous Studies on Languages with Two Genders

As far as we know, the first attempt to induce gender agreement attraction was made in the production study on Italian by Vigliocco et al. (1995). Virtually no evidence of attraction was found: out of 1920 responses only four (0.2%) contained a gender error. However, in a later study Vigliocco and Franck (1999) observed gender agreement attraction in Italian.

Vigliocco and Franck carried out four production experiments: two on Italian and two on French. Both languages have two genders: masculine and feminine. In all experiments, participants saw a masculine and a feminine adjective at the same time (one above the other) and then a noun phrase, and had to combine them saying the resulting sentence aloud. The gender of the head and the attractor were manipulated. When the genders

[3]In production, looking for symmetric effects in ungrammatical and grammatical sentences is less straightforward. However, several authors suggested not only counting errors, but also measuring RTs during elicitation tasks (e.g., Staub, 2009, 2010; Brehm and Bock, 2013). They demonstrated that participants slow down when the subject contains a singular head and a plural attractor both when they eventually answer correctly and when they do not [to be precise, Staub observed this for the subjects containing a PP attractor, but not for the subjects contained within relative clauses, as in (1b)].

mismatched, people were found to make more agreement errors. In Italian, there was no significant difference between FM and MF conditions[4]. In French, more errors were made in FM conditions (the difference was significant in Experiment 2 and marginally significant in Experiment 4). Whether the head gender was purely grammatical (on inanimate nouns) or conceptual (on animate nouns) also played a role. Participants made fewer errors in the latter case. Thus, semantic factors do enter the picture in case of gender agreement attraction, but, as far as we can judge, only to suppress it (on the contrary, conceptual numerosity can increase the number agreement attraction rate).

The observed pattern of attraction errors was different from number agreement studies. Firstly, a significant number of errors was made in all mismatch conditions, while in case of number agreement, the error rate in the conditions with plural heads and singular attractors was very low, often the same as the error rate without attraction. Secondly, both in French and in Italian, masculine is used as the grammatical default (for example, it appears in impersonal constructions and in the cases where the predicate must agree with several masculine and feminine nouns) and is more frequent. So the pattern observed in French (more errors in FM conditions) is the reverse of the number agreement attraction pattern found across languages.

The authors concluded that feature markedness does not matter for gender agreement and outlined an explanation based on inflectional differences between Italian and French. However, this explanation was undermined by Anton-Mendez et al. (2002) who conducted a production study on Spanish. Spanish is similar to Italian in terms of adjectival inflections, but the results were the same as in French. In addition to that, Vigliocco and Zilli (1999) and Franck et al. (2008) demonstrated in a number of experiments on Italian, Spanish, and French that the morphophonological properties of the head influence the error rate in gender agreement attraction. As in the studies of number agreement attraction, there were fewer errors when heads had regular inflections, but no similar effects were found for attractors.

We could find only two studies examining gender agreement attraction in comprehension: Acuña-Fariña et al. (2014) and Martin et al. (2014). Both looked at Spanish, eye-tracking was used in the first and ERPs in the second. Attraction effects were detected, but no differences between M and F genders were reported.

1.3.2. Previous Studies on Languages with Three Genders

Badecker and Kuminiak (2007) (henceforth, B&K) report results of three production experiments on Slovak. Slovak has three genders: masculine, feminine, and neuter. M is the most frequent, N is the least frequent, but is used in impersonal constructions. In all experiments, participants were given subject NPs (often called "preambles') and asked to generate complete sentences. In Experiment 1, B&K compared the number of errors in two groups of conditions: MM, MF, FF, FM and MM, MN, NN, NM. As in the previous studies, there were significantly more errors in mismatch conditions than in match conditions. But the pattern was different: there were more errors in the MF condition compared to the FM and in the NM compared to the MN.

Experiment 2 confirmed the results of Experiment 1 (it contained MM, MF, FF, and FM conditions and was designed to test the role of morphophonological factors). In Experiment 3, NN, NM, and NF conditions were compared. NM and NF preambles provoked more errors than NN preambles; but the number of errors in NM and NF conditions was comparable. Explaining this pattern, B&K adopt an optimality-theoretic approach and argue that there is no single markedness hierarchy in the Slovak gender system (such as N < M < F), but markedness is defined in pairs (N < M, N < F, M < F). Among other things, the results of this study show that frequency does not play a role for feature asymmetries.

Another production experiment was conducted on Russian (Lorimor et al., 2008). The authors manipulated both the number and the gender of heads and attractors (only M and F genders were used). In all trials, participants saw *and* heard the predicate and then saw the preamble. Their task was to construct a sentence using these two parts and to say it aloud. Out of 1155 answers where gender agreement was necessary (in Russian, as well as in Slovak, verbs agree in gender only in past tense singular forms), only seven (0.6%) contained an agreement error. Based on this, the authors concluded that gender agreement attraction does not exist in Russian.

To summarize, in all gender agreement attraction studies, if any effects are observed, error rates in all mismatch conditions are higher than in match conditions (unlike in number attraction studies, where significant effects are found only in one mismatch condition: with singular heads and plural attractors). Otherwise, the results of gender agreement studies are different: larger effects are found in the FM condition (compared to the MF condition) in Spanish and French, and in the MF and NM conditions (compared to the FM and MN conditions) in Slovak. The results from Slovak are closer to the pattern observed for number, if we assume that feminine and masculine genders and plural number are marked.

Out of several approaches to attraction outlined above, the existence of gender agreement attraction is hardly compatible with the Marking and Morphing model, primarily because in the absolute majority of cases, gender features are semantically empty. Moreover, even if we take nouns with conceptual gender, as *mal'čik* "boy$_M$" or *sestra* "sister$_F$" in Russian, it makes little sense to assume that, for example, having an M dependent NP could make an F noun "more masculine." Notably, we do not want to say that the existence of attraction with semantically empty features implies that conceptual numerosity cannot play any role for number agreement attraction - various experimental findings clearly indicate that it does (e.g., Bock and Cutting, 1992; Eberhard, 1999; Haskell and MacDonald, 2005; Mirkovic and MacDonald, 2013). We would only like to stress that attraction is possible without any semantic effects of this sort and therefore should result from some process that does not depend on them

[4]In combinations like MFM the first letter shows the gender of the head, the second letter - the gender of the attractor, the third letter (if present) the gender of the predicate.

(e.g., from the formal properties of features). Semantic effects can be added to the picture, but this is optional.

1.4. The Present Study

Apparently, gender agreement attraction errors are more difficult to induce than number errors. For example, Vigliocco et al. (1995) did not observe them in Italian, although they were found in subsequent experiments. So we decided to run another production experiment on Russian replicating B&K's first experiment on Slovak (which, in terms of its gender system, is very close to Russian). Our goal was to see whether any attraction errors would be induced, and, if yes, whether the pattern would be similar to B&K's study or to what has been observed for French, Spanish, or Italian. We also planned comprehension experiments because no existing studies had looked at comprehension in a language with three genders. We were particularly interested to find out whether production and comprehension results would be parallel and whether ungrammaticality illusions would be found. Before we move on to the experiments, let us present a brief overview of the Russian gender system.

1.4.1. Russian Gender System

Russian nouns are inflected for number and case, and the ones that have the same endings in the majority of forms are grouped into declension classes. Russian has three declension classes for nouns (and a separate class for substantivized adjectives). The first class includes almost all M nouns (they have zero endings in nominative singular, like *mal'čik* "boy") and all N nouns (they have *-o* or *-e* endings, like *okno* "window"). These M and N nouns use the same set of endings in all cases except for genitive plural and nominative and accusative in singular and plural (in plural, all declension classes have the same endings in dative, instrumental and locative). The second class includes the majority of F nouns (they end in *-a* or *-ja*, like *devočka* "girl") and a small group of animate M nouns with the same endings, like *mužčina* "man." The third class includes F nouns with zero endings in nominative singular, like *doč'* "daughter." In addition to that, there are some irregular and uninflected nouns.

Thus, in most cases, it is impossible to determine the gender of the noun unambiguously looking at the noun itself, and, at least prima facie, we cannot speak of something like morphological markedness in the noun system. Let us add that M nouns are the most frequent and N nouns are the least frequent. M nouns constitute about a half of the lexicon, F nouns - about 30–35%, N nouns are the rest (Yanovich and Fedorova, 2006; Slioussar and Samoilova, 2014).

Gender agreement can be observed only in singular, on adjectives, participles and past tense verb forms. Russian adjectives and participles have so-called full forms (used attributively and predicatively) and short forms (used only in predicates and inflected for number and gender, but not for case). M form is the citation form (i.e., the form would appear in dictionaries, grammatical descriptions etc.).

Verb forms and short forms of adjectives and participles have zero endings in M gender (e.g., *byl* "was$_M$" - *byla* "was$_F$" - *bylo* "was$_N$"), otherwise all forms have non-zero endings (e.g., *krasivyj* "beautiful$_{M.NOM.SG}$" - *krasivaja* "beautiful$_{F.NOM.SG}$" - *krasivoe* "beautiful$_{N.NOM.SG}$"). Thus, we cannot say that M forms are morphologically unmarked, even if we limit ourselves to predicates. In impersonal sentences, where unmarked forms are expected, N predicates are used, as (4) shows.

(4) Svetalo.
dawn$_{PST.N.SG}$
It dawned.

As for gender conflict resolution, another classical test for markedness, it is of limited use in Russian because there is no gender agreement in plural. Gender conflict resolution can be observed only in constructions like "X and Y each did something." We conducted an informal questionnaire, asking about 30 native speakers.

As we discuss below, acceptability of such sentences differs depending on animacy of the nouns and the genders that are combined, and there is substantial individual variation among speakers. However, one crucial generalization can be made: examples with the feminine or neuter forms of *každyj* "each' are never found even marginally acceptable, only some examples with the masculine forms are.

Firstly, let us consider sentences with M and F nouns, like in (5). Not all speakers of Russian find these examples acceptable, but for those who do, this construction sounds better with human animates (5a) than with non-human animates (5b). Nobody accepts this construction with inanimate nouns, as in 6a), although they can be used in such sentences if both nouns are of the same gender, as in (6b)[5].

(5) a. Mužčina i ženščina každyj
man$_{M.NOM.SG}$ and woman$_{F.NOM.SG}$ each$_{M.NOM.SG}$
sjeli po jabloku
ate$_{PST.PL}$ PREP$_{DISTR}$ apple$_{DAT.SG}$

b. Jož i svin'ja
hedgehog$_{M.NOM.SG}$ and swine$_{F.NOM.SG}$
každyj sjeli po jabloku.
each$_{M.NOM.SG}$ ate$_{PST.PL}$ PREP$_{DISTR}$ apple$_{DAT.SG}$

(6) a. Divan i krovat' každyj
sofa$_{M.NOM.SG}$ and bed$_{F.NOM.SG}$ each$_{M.NOM.SG}$
stoili celoe sostojanie.
cost$_{PST.PL}$ whole$_{ACC.SG}$ fortune$_{ACC.SG}$

b. Kušetka i krovat' každaja
couch$_{F.NOM.SG}$ and bed$_{F.NOM.SG}$ each$_{F.NOM.SG}$
stoili celoe sostojanie.
cost$_{PST.PL}$ whole$_{ACC.SG}$ fortune$_{ACC.SG}$

Now let us look at M and N nouns. More than half of the speakers we asked rejected this construction even with animate human nouns (7a) as ungrammatical, but those who accepted it used masculine form. All our informants rejected examples with non-human animates like (7b) or

[5]Since acceptability ratings for some sentences vary from speaker to speaker, we do not mark any of the examples below with asterisks or question marks used to indicate ungrammaticality or marginal acceptability.

found them only marginally acceptable. This might be at least partly due to independent factors (the relevant neuter words, like *mlekopitajuščee* "mammal," *životnoe* "animal," *nasekomoe* "insect," tend to be abstract), but is still telling.

(7) a. Voin i dit'a každyj
warrior$_{M.NOM.SG}$ and child$_{N.NOM.SG}$ each$_{M.NOM.SG}$
sjeli po jabloku.
ate$_{PST.PL}$ PREP$_{DISTR}$ apple$_{DAT.SG}$

b. Gryzun i nasekomoe každyj
rodent$_{M.NOM.SG}$ and insect$_{N.NOM.SG}$ each$_{M.NOM.SG}$
vypili po kaple.
drank$_{PST.PL}$ PREP$_{DISTR}$ drop$_{DAT.SG}$

Finally, such constructions with F and N nouns, as in (8), were rejected by most of our informants. The few people who accepted them again preferred the masculine form.

(8) Ženščina i dit'a každyj
woman$_{F.NOM.SG}$ and child$_{N.NOM.SG}$ each$_{M.NOM.SG}$
sjeli po jabloku.
ate$_{PST.PL}$ PREP$_{DISTR}$ apple$_{DAT.SG}$

Let us add that M nouns are used to refer to groups of people of mixed or uncertain gender, or to an arbitrary member of such groups. This generalization is discussed by Yanovich (2012) who shows that it does not hold for animals. For example, the word *sobaka* "dog" is feminine. There are specific words to denote male and female dogs, but they are much more often used as swearwords, like the English *bitch*. To sum up, N appears to be the grammatical default as the gender used in impersonal constructions, while all cases where M is used as the standard option are limited to the nouns denoting humans and sometimes other animates. In all our experiments, we used only inanimate nouns as heads and attractors (we wanted to avoid additional factors before the general picture becomes clear)[6].

2. EXPERIMENT 1

Experiment 1 was designed to check whether the findings of Badecker and Kuminiak (2007) would be replicated in Russian, which is very close to Slovak in the relevant part of the grammar. In particular, both languages have three genders, M is the most frequent, N is the least frequent, but is used in impersonal sentences. There are no articles. Gender agreement can be observed on adjectives and participles (in singular) and on verbs (in past tense singular). The system of declensions is very similar as well.

2.1. Participants

Thirty native speakers of Russian (8 male, 22 female) participated in Experiment 1. Ages ranged from 18 to 50 (mean age 28.7, *SD* 9.4). No participant took part in more than one experiment. All experiments reported in this paper were carried out in accordance with the Declaration of Helsinki and the existing Russian and international regulations concerning ethics in research. All participants provided informed consent. They were tested at the Laboratory for Cognitive Studies of Saint-Petersburg State University.

2.2. Materials

In this experiment, participants first saw a predicate, then on the next slide a subject at which point they were asked to produce a complete sentence. In half of the cases, predicates did not agree with the subject in gender, and participants were asked to modify them. Like in B&K's study, subject noun phrases were always built according to the following schema: NP_1–preposition–NP_2, e.g., *okno vo dvor* "window$_{N.SG}$ to yard$_{M.SG}$." NP_1 was always in nominative singular, NP_2 was in accusative singular. We selected inanimate nouns that have the same form in accusative and nominative, since this was shown to inflate the error rate (Badecker and Kuminiak, 2007). As in many other agreement attraction studies, we had both adjunct and argument PPs.

The predicates always consisted of two words: the copula *byt'* "to be" in the past tense (where gender agreement can be observed) and an adjective or participle. We opted for such predicates because they are short and do not contain any objects or other nouns that could cause additional disturbance of subject-predicate agreement (initially, we wanted to use single verbs, but could not come up with such predicates for all experimental stimuli). Adjectives and participles were always in instrumental singular form[7].

The genders of NP_1 and NP_2 were manipulated. As **Table 1** shows, these two factors were not fully crossed. Like in B&K's Experiment 1, we used only seven out of nine possible combinations of genders. Additionally, we manipulated the agreement marking on the predicate[8]. Sample stimuli in conditions 1-4 in **Table 1** represent one set: two variants of the subject NP (one head and two different dependent nouns, or attractors) and two variants of the predicate (matched or mismatched in gender with the subject). We constructed 48 sets, 12 for each of the four combinations of conditions. This approach to the construction of materials (one head noun and several attractors of different genders, plus a grammatical and an ungrammatical version of the predicate) holds for all experiments in this article. All materials are listed in Appendices in Supplementary Material.

In addition to that, we constructed 100 fillers, also consisting of a predicate and a subject. Subject NPs had singular or plural

[6] Vigliocco and Franck (1999) demonstrated that the gender agreement error rate was lower when the gender of the head noun was conceptual, rather than purely grammatical, but we would not expect markedness patterns to be reversed in such cases.

[7] As we explained in the introduction, participles, adjectives, and nouns in predicates can appear either in nominative or in instrumental, and adjectives and participles also have short forms used only in predicates and inflected for gender and number, but not for case. Often only one variant is grammatical, but sometimes two or even three are, or one is fine, while the others are marginally acceptable. Meaning nuances associated with them can be very subtle. It will suffice to say that we chose instrumental forms because, unlike nominative and short forms, they suited all our stimuli. But, if the participants occasionally responded with nominative or short forms, we did not count this as a mistake.

[8] We opted for this design primarily to facilitate the comparison with B&K's study. In addition to that, when we were pretesting the experiment, we found that the experimental session was relatively short, but very intense because we prompted the participants to respond very fast. We concluded that making it 1.5 times longer to fully cross the two factors could make it too taxing.

TABLE 1 | Gender combinations used in Experiment 1.

Condition	NP_1 gender	NP_2 gender	Predicate gender	Example
1 / 2	M	M	M / F	*byl prosročennym / byla prosročennoj + recept na porošok* ($was_{M.SG}$ $expired_{M.SG}$ / $was_{F.SG}$ $expired_{F.SG}$ + $prescription_{M.NOM.SG}$ for $powder_{M.ACC.SG}$)
3 / 4	M	F	M / F	*byl prosročennym / byla prosročennoj + recept na maz'* ($was_{M.SG}$ $expired_{M.SG}$ / $was_{F.SG}$ $expired_{F.SG}$ + $prescription_{M.NOM.SG}$ for $ointment_{F.ACC.SG}$)
5 / 6	F	F	F / M	
7 / 8	F	M	F / M	
9 / 10	M	M	M / N	
11 / 12	M	N	M / N	
13 / 14	N	N	N / M	*byl otkrytym / bylo otkrytym + okno v pole* ($was_{M.SG}$ $opened_{M.SG}$ / $was_{N.SG}$ $opened_{N.SG}$ + $window_{N.NOM.SG}$ to $field_{N.ACC.SG}$)
15 / 16	N	M	N / M	*byl otkrytym / bylo otkrytym + okno vo dvor* ($was_{M.SG}$ $opened_{M.SG}$ / $was_{N.SG}$ $opened_{N.SG}$ + $window_{N.NOM.SG}$ to $yard_{M.ACC.SG}$)

heads and adjectival or prepositional modifiers (the NPs inside these PPs were not in accusative). Predicates were similar to the ones in target stimuli and did not agree with subjects in gender in one third of the cases.

Each participant saw only one target stimulus from each set. Consequently, we had four experimental lists with 148 items (48 stimuli and 100 fillers). The number of conditions was balanced for every list. Thus, every participant saw three target items per condition: for example, three FF stimuli (having an F head and an F attractor) with a matched F predicate, three FF stimuli with a mismatched M predicate etc. All lists began with ten fillers, and then fillers and experimental items were presented in a pseudo-random order, with the constraint that no more than two experimental items occur consecutively.

2.3. Procedure

In a pilot experiment, we used the same procedure as in B&K 's study: participants listened to preambles and were asked to generate complete sentences. But after running six subjects, we did not get any attraction errors. This can be explained by the fact that such errors are in general relatively infrequent. In B&K 's study, they occurred in 3% cases on average. Since the number of errors varies from subject to subject, the probability to elicit no errors from several people in a row is considerably high. However, we decided to switch to a different method in the main experiment in hope to elicit more errors.

The experiment was run on a Macintosh computer using PsyScope software (Cohen et al., 1993). In every trial, participants saw on the computer screen a fixation point (for 300 ms), then a predicate (for 800 ms), and then a subject NP (for 800 ms). Their task was to combine the predicate and the subject in a grammatical sentence and to say it aloud. If the predicate did not agree with the subject, participants were instructed to modify the predicate. Before the main session started, the experimenter explained the task on two sample items (saying that participants would see two phrases and would be asked to combine them into a correct sentence as fast as possible, i.e., without explicitly mentioning gender agreement). Then there were four practice items.

To encourage participants to respond faster, a time counter appeared on the screen after both the predicate and the subject were presented. As soon as the participant responded, the experimenter pressed a key, and the next trial started. All participants' responses were tape-recorded. An experimental session lasted around 7.5 min.

2.4. Results

The participants' responses were transcribed, and each of them was assigned into one of the following categories:

i. Correct response: the sentence is grammatical, the subject and the predicate provided as stimuli are repeated faithfully.
ii. Agreement error: the sentence is correct except for a gender agreement error.
iii. Repetition error: the sentence is grammatical, but the subject or the predicate is repeated incorrectly (for example, the word *krem* "cream" was used instead of the word *maz'* "ointment").
iv. A combination of a repetition error and an agreement error.
v. Incomplete response: the participant utters only a part of the sentence or says nothing at all.
vi. A combination of an incomplete response and an agreement error: the sentence is incomplete, but a verb, a participle or an adjective was uttered and did not agree with the subject (cf. 9a – 9b).

(9) a. Recept na maz' byla …
recipe$_{M.NOM.SG}$ for ointment$_{F.ACC.SG}$ was$_{F.SG}$ …

b. Recept na maz'
recipe$_{M.NOM.SG}$ for ointment$_{F.ACC.SG}$
prosročennaja…
expired$_{F.SG}$ …

Errors in subject-verb gender agreement were the only grammar errors participants made, all other errors involved incorrectly repeating or omitting lexical material (we did not expect any

TABLE 2 | The distribution of responses in Experiment 1.

Correct responses	1018 (71.2%)
Agreement errors	61 (4.3%), 8 of them partial[a]
Repetition errors	111 (7.8%)
Repetition and agreement errors	9 (0.6%), 3 of them partial
Incomplete responses	224 (15.7%)
Incomplete responses with agreement errors	7 (0.5%)

[a]*In most sentences with agreement errors, both the verb and the adjective or participle were in a wrong form. They are components of a complex predicate, so we counted this as one error (note that counting them as two errors instead would not affect the outcome, because the differences between the relevant conditions would only be inflated). However, in several cases only one of the two components did not agree with the subject.*

other grammar errors, for example, in number or case, but they could have occurred accidentally). To exclude mishearings during transcription, both authors of this paper and two other native speakers of Russian listened to all responses to target stimuli. The number of errors in each category is given in **Table 2**. In case of self-corrections, only the first variant was counted, both when participants changed an answer with an error to a correct one and when they did the opposite (this happened in three cases).

At the following stage of analysis, we collapsed all agreement errors together. The distribution of errors by experimental conditions is given in **Table 3**. In total, there were 77 agreement errors (5.4% from all responses). Only 13 out of them were not due to attraction (they are discussed in more detail below). The difference between the number of agreement errors with and without attraction is statistically significant according to the chi-square test[9] [$\chi^2_{(1,\,N=77)} = 18.97$, $p < 0.01$], so our results show that gender agreement in Russian is subject to attraction.

As **Table 3** shows, agreement errors were more frequent in predicate mismatch conditions, but were not limited to them. Out of 13 errors without attraction, in eight cases, a mismatched predicate was not changed, but there were also five cases where participants produced a neuter predicate with an MF subject, a masculine predicate with an NN subject etc., although they were provided with other forms, matched or mismatched with the subject. Out of 64 attraction errors, 11 errors occurred in predicate match conditions, i.e., participants changed the correct gender of the predicate they were provided with to an incorrect one due to attraction.

Conditions with matched and mismatched predicates are collapsed in **Table 4** showing that the number of agreement attraction errors differs depending on the combination of genders of the head and attractor nouns. To test whether these differences are statistically significant, we modeled the data with a mixed-effects logistic regression in the statistical software program R (R Core Team, 2014) using the *glmer* function from the *lme4* package (Bates et al., 2015).

Firstly, we compared MF and FM conditions. The logistic regression evaluated the likelihood of an agreement attraction error (coded as 1) vs. a correct response (coded as 0). The combination of genders was treated as a fixed effect. For the predictors we used contrast coding: MF was coded as 0.5, FM was coded as −0.5. Random intercepts by participant and by item were also included in the model. The results of the analysis are reported in **Table 5**. The coefficient for the intercept was significant, reflecting that most responses were correct. There was also a significant main effect of Gender Combination indicating that F attractors trigger significantly more errors than M attractors.

Secondly, we compared MN and NM conditions in the same way. MN was coded as 0.5, NM was coded as −0.5. The coefficient for the intercept was again significant because most responses were correct. But the main effect of gender combination did not reach significance. We also compared MF and MN conditions and FM and NM conditions, as well as the number of non-agreement ("other") errors in different conditions, but did not find any significant differences.

TABLE 3 | The distribution of responses by condition in Experiment 1.

	Correct response	Agr. error (attraction)	Agr. error (no attraction)	Other errors
Condition 1 (MM + M)[a]	69	0	0	16
Condition 2 (MM + F)	69	0	0	16
Condition 3 (MF + M)	66	3	1	20
Condition 4 (MF + F)	53	19	1	17
Condition 5 (FF + F)	65	0	0	25
Condition 6 (FF + M)	57	0	3	30
Condition 7 (FM + F)	66	1	0	23
Condition 8 (FM + M)	50	10	1	29
Condition 9 (MM + M)	74	0	0	16
Condition 10 (MM + N)	64	0	3	23
Condition 11 (MN + M)	69	1	0	20
Condition 12 (MN + N)	59	11	0	20
Condition 13 (NN + N)	64	0	1	25
Condition 14 (NN + M)	68	0	2	20
Condition 15 (NM + N)	62	6	1	21
Condition 16 (NM + M)	63	13	0	14

[a]*Due to our mistake, there are 85 responses in conditions 1 and 2 rather than 90.*

TABLE 4 | The Number of gender agreement attraction errors by condition in Experiment 1.

Head/attractor gender	Correct responses	Attraction errors	Other errors
MF	119	22	39
FM	116	11	53
MN	136	12	42
NM	125	19	36

[9]In half of the conditions, where the genders of the head and the attractor coincided, no agreement errors with attraction were possible, while in the other half of the conditions, these errors prevailed, but there were also agreement errors without attraction, as **Table 3** shows. This is why we chose the chi-square test.

TABLE 5 | Results of the analysis for Experiment 1.

Conditions	Predictor	Coefficient	Std. error	Wald Z	*p*
MF vs. FM	(Intercept)	−3.04	0.43	−7.01	<0.01
	GenComb	−0.95	0.48	−1.96	0.05
MN vs. NM	(Intercept)	−2.68	0.30	−8.82	<0.01
	GenComb	0.62	0.39	1.59	0.11

2.5. Discussion

The results of Experiment 1 are similar to the results of B&K's first experiment, which can be explained by the fact that the two languages have similar gender systems, as we demonstrated in the introduction. In both studies, F attractors triggered more errors than M attractors. N attractors triggered fewer errors than M attractors, but this difference was statistically significant only in B&K's study. As we mentioned in the introduction, other authors studying gender attraction in French and Spanish (which have two genders and where M is grammatical default), observed a different pattern: there were more errors with M attractors than with F attractors. We postpone further discussion until the general discussion section.

3. EXPERIMENT 2A

Experiment 2a was designed to find out whether gender agreement attraction can also be detected in comprehension. For the sake of comparison with Experiment 1, we used the same combinations of head and attractor noun genders.

3.1. Participants

Forty-eight native Russian speakers (19 female and 29 male) took part in the experiment. Ages ranged from 19 to 26 (mean age 20.9, *SD* 1.9).

3.2. Materials

The materials consisted of target and filler sentences. All target sentences were 9–10 words long and followed the schema: NP_1–preposition–NP_2–copula (*byt'*) - adjective/participle - four-five words modifying the predicate. We had the same 16 conditions as in Experiment 1 (see **Table 1** above). Almost all subject NPs and predicates were based on the materials from Experiment 1 and followed the same constraints. In half of the conditions, the predicate did not agree with the subject. Given existing findings on number agreement attraction, we expected parallel results in production and comprehension. In particular, we expected to find grammaticality illusions in conditions MFF, FMM, MNN, and NMM (this would mean that they would be read significantly faster than the other four ungrammatical conditions: MMF, FFM, MMN, NNM).

As in Experiment 1, conditions were grouped in sets, each set containing four conditions with the same head nouns. An example of a stimuli set is given in (10)[10]. For each condition set we constructed 12 sentences, 48 target sentences in total.

[10]The translation for all sentences is identical, so we only give it for the first one.

(10) a. Recept$_{M.NOM.SG}$ na porošok$_{M.ACC.SG}$ byl$_{M.SG}$ pom'atym$_{M.SG}$ iz-za sil'nogo volnenija$_{GEN.SG}$ pacienta.$_{GEN.SG}$
"The recipe for the powder was crumpled due to the patient's extreme nervousness."

b. Recept$_{M.NOM.SG}$ na maz'$_{F.ACC.SG}$ byl$_{M.SG}$ pom'atym$_{M.SG}$ iz-za sil'nogo volnenija$_{GEN.SG}$ pacienta.$_{GEN.SG}$

c. Recept$_{M.NOM.SG}$ na porošok$_{M.ACC.SG}$ byla$_{F.SG}$ pom'atoj$_{F.SG}$ iz-za sil'nogo volnenija$_{GEN.SG}$ pacienta.$_{GEN.SG}$

d. Recept$_{M.NOM.SG}$ na maz'$_{F.ACC.SG}$ byla$_{F.SG}$ pom'atoj$_{F.SG}$ iz-za sil'nogo volnenija$_{GEN.SG}$ pacienta.$_{GEN.SG}$

Additionally, we constructed 120 fillers, which had roughly the same structure as experimental sentences. Subject NPs in fillers consisted of a single noun modified by an adjective, or of a complex NP, where the embedded noun was not in accusative. All fillers were grammatical. Thus, we had 24 ungrammatical and 144 grammatical sentences, making the grammatical-to-ungrammatical ratio 6:1. Experimental sentences and fillers were distributed in four counterbalanced experimental lists. Every list started with ten fillers; then stimuli and fillers were presented in pseudo-random order with the constraint that a maximum of two stimuli could occur consecutively.

3.3. Procedure

The sentences were presented on a PC using Presentation software (http://www.neurobs.com). We used the word-by-word self-paced reading methodology (Just et al., 1982). Each trial began with a sentence in which all words were masked with dashes while spaces and punctuation marks remained intact. Participants were pressing the space bar to reveal a word and re-mask the previous one. One third of the sentences was followed by forced choice comprehension questions to ensure that the participants were reading properly. Two answer variants were presented on the left and on the right of the screen. Participants pressed "f" to choose the answer on the left, and "j" to choose the answer on the right. Participants were instructed to read at a natural pace and answer questions as accurately as possible. They were not informed in advance that sentences would contain errors. An experimental session lasted around 14 min.

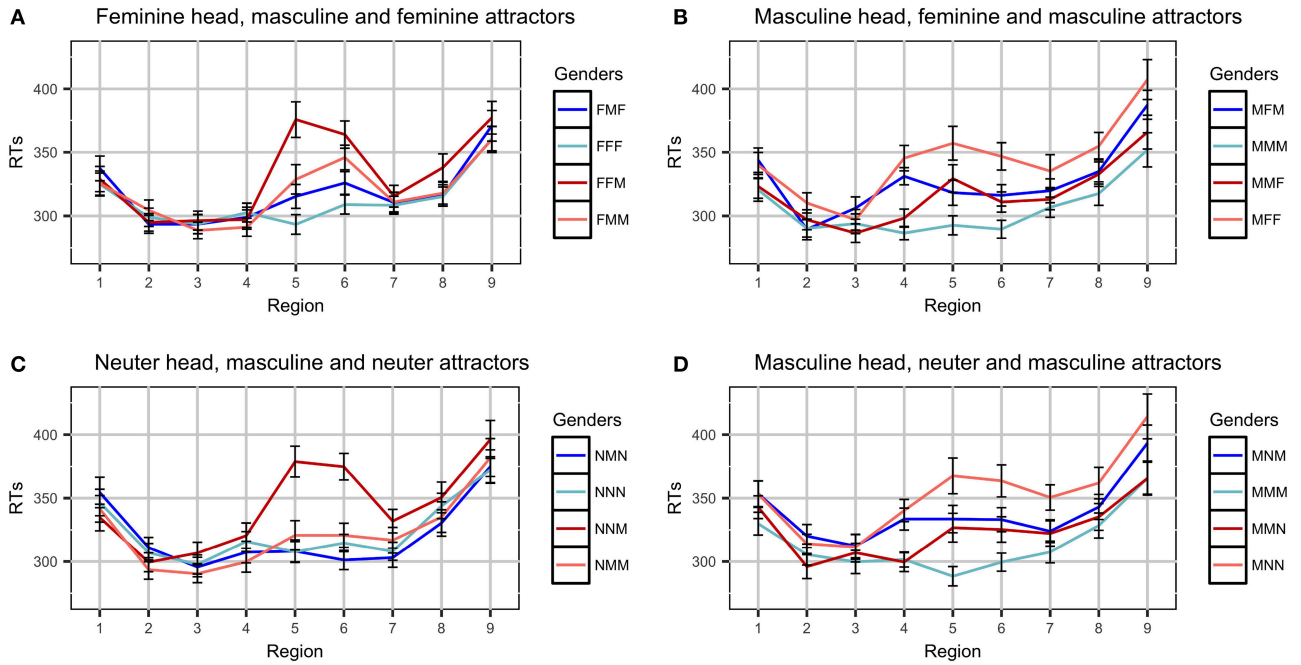


FIGURE 1 | Plots of mean RTs (in ms) by conditions in Experiment 2a. Error bars represent standard errors of the means. Regions: NP_1 (1) - preposition (2) - NP_2 (3) - copula *byt'* (4) - Adj/Part (5) - spillover (6–9). Ungrammatical conditions are red, grammatical ones are blue. Conditions where the gender of the attractor and the predicate coincide (for example, FFF and FMM) have dark colors, conditions where they do not (for example, FMF and FFM) have light colors. **(A)** Feminine head, masculine and feminine attractors, **(B)** Masculine head, feminine and masculine attractors, **(C)** Neuter head, masculine and neuter attractors, and **(D)** Masculine head, neuter and masculine attractors.

3.4. Results

We analyzed participants' question-answering accuracy and reading times. Two participants answered more than 20% questions incorrectly, so their data were discarded. Otherwise no participant made more than two mistakes when answering questions to target sentences (i.e., 10% at most). Reading times that exceeded a threshold of 2.5 standard deviations, by region and condition, were excluded (Ratcliff, 1993). For two participants, this led to the exclusion of more than 15% responses, so we did not include their data in further analysis.

After four participants were excluded, we had 44 participants (11 in each experimental list). In total, 2.3% of the data were excluded as outliers (never more than 3.6% per region and condition). Average RTs per region in different conditions are presented in **Figure 1**.

The data for each set of conditions (e.g., MMM - MFM - MMF - MFF) were entered in a 2 × 2 Repeated Measures ANOVA with grammaticality and gender match between the attractor and the head nouns as factors. We used IBM SPSS software (www.ibm.com/software/analytics/spss/). Analyses by items and by participants were performed. Data from all regions were tested, but there were significant results only in regions 4–6 in the conditions with M heads and in regions 5–6 in the conditions with F and N heads. Region 4 is the copula, region 5 is an adjective or participle, regions 6–10 contain several words modifying the predicate. The results of the tests for the relevant regions are given in **Table 6**.

3.4.1. Feminine Head, Masculine Attractor

The main effect of Grammaticality is significant in analysis by subjects and by items in regions 5–6, reflecting the fact that ungrammatical sentences were read slower than grammatical ones. The main effect of Gender Match is not significant in any region. The interaction of Grammaticality and Gender Match is significant in analysis by subjects and by items in region 5 and only in analysis by subjects in region 6. Ungrammatical sentences were read faster if the head and the attractor were mismatched in gender (i.e., in the FMM condition compared to the FFM condition). This is the classical attraction pattern.

3.4.2. Neuter Head, Masculine Attractor

The main effect of Grammaticality is significant in regions 5–6, reflecting longer RTs in ungrammatical conditions. The main effect of Gender Match is significant only in analysis by subjects in regions 5–6. The interaction of Grammaticality and Gender Match is significant in regions 5–6, which is again a reflection of the classical attraction pattern: NMM condition was read faster than NNM and, in fact, almost as fast as grammatical conditions.

TABLE 6 | Results of the analysis for Experiment 2a.

Conditions	Region	Factor	df	MS_{effect}	F_1	*p*	df	MS_{effect}	F_2	*p*
FF vs. FM	5	Gram	**1.43**	**103775.64**	**18.29**	**<0.01**	**1.11**	**35056.83**	**17.86**	**<0.01**
		GenMatch	1.43	5355.95	1.56	0.22	1.11	3326.67	2.05	0.18
		Gram * GenMatch	**1.43**	**50162.64**	**20.54**	**<0.01**	**1.11**	**21717.52**	**5.01**	**0.05**
	6	Gram	**1.43**	**62551.84**	**21.23**	**<0.01**	**1.11**	**16335.63**	**18.42**	**<0.01**
		GenMatch	1.43	0.05	<0.01	1.00	1.11	0.05	<0.01	1.00
		Gram * GenMatch	**1.43**	**14823.13**	**4.65**	**0.04**	1.11	3398.65	2.76	0.13
NN vs. NM	5	Gram	**1.43**	**78213.55**	**28.70**	**<0.01**	**1.11**	**20398.13**	**8.87**	**0.01**
		GenMatch	**1.43**	**33363.06**	**10.93**	**<0.01**	1.11	9996.53	1.92	0.19
		Gram * GenMatch	**1.43**	**36720.35**	**16.67**	**<0.01**	**1.11**	**11405.25**	**29.57**	**<0.01**
	6	Gram	**1.43**	**71017.35**	**32.99**	**<0.01**	**1.11**	**18794.17**	**27.45**	**<0.01**
		GenMatch	**1.43**	**51758.70**	**26.98**	**<0.01**	1.11	12558.27	1.99	0.19
		Gram * GenMatch	**1.43**	**20026.31**	**14.92**	**<0.01**	**1.11**	**4945.08**	**7.40**	**0.02**
MM vs. MF	4	Gram	**1.43**	**8423.58**	**5.21**	**0.03**	1.11	2100.13	2.12	0.17
		GenMatch	**1.43**	**93656.82**	**58.58**	**<0.01**	**1.11**	**25002.51**	**8.46**	**0.01**
		Gram * GenMatch	1.43	70.01	0.05	0.83	1.11	0.05	<0.01	0.99
	5	Gram	**1.43**	**63205.83**	**16.87**	**<0.01**	**1.11**	**19008.48**	**16.83**	**< 0.01**
		GenMatch	**1.43**	**34672.63**	**10.56**	**<0.01**	1.11	9163.21	1.11	0.32
		Gram * GenMatch	1.43	114.41	0.02	0.88	1.11	19.25	0.02	0.89
	6	Gram	**1.43**	**32730.00**	**13.37**	**<0.01**	**1.11**	**8554.68**	**9.06**	**0.01**
		GenMatch	**1.43**	**47491.25**	**27.90**	**< 0.01**	1.11	12185.81	2.24	0.16
		Gram * GenMatch	1.43	2128.79	0.88	0.35	1.11	401.36	1.17	0.30
MM vs. MN	4	Gram	1.43	1264.21	0.75	0.39	1.11	231.00	0.38	0.55
		GenMatch	**1.43**	**66406.31**	**24.34**	**<0.01**	1.11	15699.95	2.97	0.11
		Gram * GenMatch	1.43	2116.29	1.88	0.18	1.11	321.89	1.34	0.27
	5	Gram	**1.43**	**63247.53**	**20.26**	**<0.01**	**1.11**	**16965.12**	**23.11**	**<0.01**
		GenMatch	**1.43**	**86314.12**	**18.61**	**<0.01**	1.11	22733.11	2.23	0.16
		Gram * GenMatch	1.43	414.21	0.13	0.72	1.11	1.02	< 0.01	0.96
	6	Gram	**1.43**	**36279.68**	**9.04**	**<0.01**	**1.11**	**9509.07**	**7.94**	**0.02**
		GenMatch	**1.43**	**52540.19**	**14.24**	**<0.01**	1.11	15123.00	1.76	0.21
		Gram * GenMatch	1.43	29.05	0.01	0.93	1.11	40.33	0.02	0.90

Analyses with $p \leq 0.05$ are shown in bold.

3.4.3. Masculine Head, Feminine Attractor

The main effect of Grammaticality is significant in analysis by subjects in region 4 and in analysis by subjects and by items in regions 5–6. This reflects the fact that RTs were longer in ungrammatical conditions. The main effect of Gender Match is significant in analysis by subjects and by items in region 4, and only in analysis by subjects in regions 5–6. This corresponds to longer RTs in conditions where the genders on the nouns were mismatched. The interaction of Grammaticality and Gender Match did not reach significance in any regions, which points to the absence of agreement attraction.

3.4.4. Masculine Head, Neuter Attractor

The main effect of Grammaticality is significant in analysis by subject and by items in regions 5–6: the ungrammatical conditions are read slower than grammatical. The main effect of Gender Match is significant only in analysis by subjects in regions 4–6. The interaction of Grammaticality and Gender Match is not significant in any region, so these conditions also show no agreement attraction.

3.5. Discussion

As can be seen from the analyses, the results fall into two groups. In the conditions with F or N heads and M attractors there is clear evidence for gender agreement attraction. RTs exhibit the classical attraction profile with grammaticality illusions: ungrammatical sentences where the attractor and the predicate have the same gender (FMM and NMM) are read faster than other ungrammatical sentences (FFM and NNM). Discussing comprehension studies of number agreement attraction in the introduction, we outlined different approaches to this phenomenon, but will opt for one of them ourselves only in

TABLE 7 | Frequencies of the attractors used in Experiments 2a and 2b (in ipm, or instances per million).

Experiment	Head gender	Attractor gender	Mean attractor freq (ipm)
Experiment 2a	F	F	138.1
		M	120.0
	N	N	105.9
		M	81.7
	M	M	91.8
		F	41.1
	M	M	134.9[a]
		N	78.9
Experiment 2b	M	M	61.4
		F	61.9
	M	M	69.2
		N	68.1

[a]*It should be noted that one really frequent M noun influences this number a lot. If we get rid of it and of the corresponding N attractor, the frequencies become very close: 73.9 for M attractors and 84.6 for N attractors.*

the general discussion section once all experimental findings are presented. Let us also note that ungrammaticality illusions are absent: in the sentences with N heads there are virtually no differences between grammatical conditions; in the sentences with F heads, they are insignificantly small.

On the other hand, the conditions with M heads and F or N attractors do not show any evidence of attraction. Both grammatical and ungrammatical sentences where the head and the attractor match in features (MMM, MMF, and MMN) are read faster than the sentences where they are mismatched (MFM, MNM, MFF, and MNN). In case of ungrammatical sentences, this pattern is *the reverse* of what we usually see in attraction cases.

Looking for an explanation of such pattern, we discovered that we need to rule out an important confound first. Unfortunately, we made a mistake during the preparation of experimental materials, and the frequencies of attractors in conditions with M heads were not well balanced. Since this could influence the results in some unexpected way, we conducted an additional experiment where the frequencies were carefully controlled. Conditions with F and N heads did not have this problem, and the results reported for them hold.

4. EXPERIMENT 2B

In this experiment we follow up on potential frequency effects in the conditions with M heads from Experiment 2a.

4.1. Participants

Thirty-five native Russian speakers (17 female, 18 male) took part in the experiment. Ages ranged from 21 to 47 (mean age 31.3, *SD* 6.2).

4.2. Materials

We constructed 32 sets of stimuli according to the same schema as in Experiment 2a and observing the same constraints. Head nouns were always masculine. In 16 sets, the attractors were masculine and neuter; in the other 16 sets, the attractors were masculine and feminine. Most of the head nouns were re-used from the Experiment 2a, but we replaced attractors so that their frequencies were closely matched inside the two groups of conditions. We used *The Frequency Dictionary of Modern Russian Language* (Lyashevskaya and Sharoff, 2009). Average frequencies of head and attractor nouns in Experiments 2a and 2b are shown in **Table 7**. As in Experiment 2a, half of the predicates did not agree with the subject in gender. Additionally, we used 80 fillers from Experiment 2a. Experimental sentences were distributed into four experimental lists, with factors counterbalanced. As a result, we had 112 sentences per list (16 ungrammatical and 96 grammatical), making the grammatical-to-ungrammatical ratio 6:1.

4.3. Procedure

The procedure was the same as in Experiment 2a. An experimental session lasted around 9 min.

4.4. Results

Like in Experiment 2a, we analyzed participants' question-answering accuracy and reading times. At the first stages of analysis, the data from three participants were discarded: one of them had <75% accuracy in comprehension questions; the other two read too slowly compared with the others, so more than 15% of their RTs would have to be excluded as outliers (exceeding the threshold of 2.5 standard deviations). As a result, we had 32 participants, eight for each experimental list.

After three participants were excluded, on average, 1.5% RTs were excluded as outliers (never more than 3.1% per region and condition). Average RTs per region in different conditions are presented in **Figure 2**.

2 x 2 Repeated Measures ANOVAs with grammaticality and gender match as factors were used to analyze RTs, as in Experiment 2a. Significant results were found only in regions 5 (adjective/participle) and 6–7 (spillover regions). They are presented in **Table 8**.

4.4.1. Masculine Head, Feminine Attractor

The main effect of Grammaticality was significant in analysis by subjects and by items in regions 5–6, and only in analysis by subjects in region 7. This reflects the fact that ungrammatical sentences were read slower than grammatical ones. The main effect of Gender Match was significant only in analysis by subjects in regions 5–7. The interaction between Grammaticality and Gender Match was not significant in any region.

4.4.2. Masculine Head, Neuter Attractor

The results were almost the same as in the other set of conditions. The main effect of Grammaticality was significant in regions 5–7. The main effect of Gender Match was significant only in analysis by subjects in regions 5–7. The interaction between the factors never reached significance.

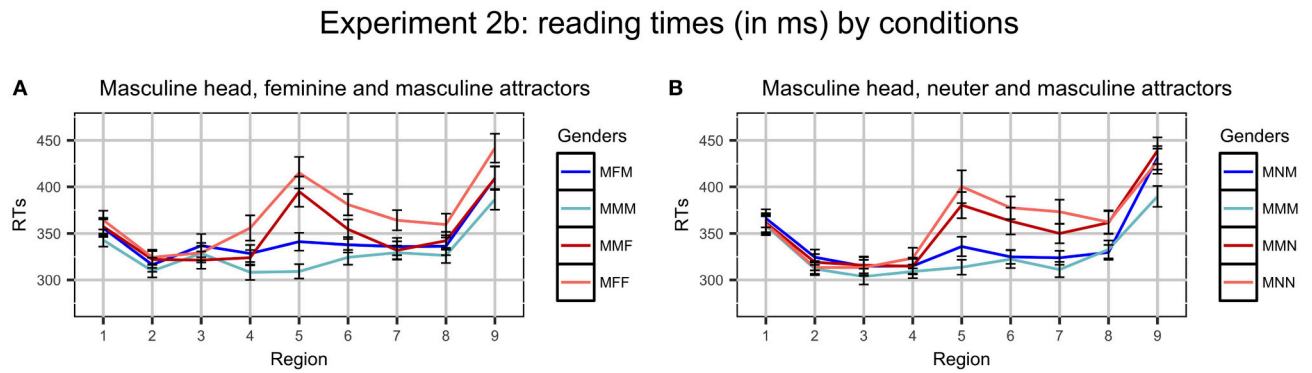


FIGURE 2 | Plots of mean RTs (in ms) by conditions in Experiment 2b. Error bars represent standard errors of the means. Regions: NP_1 (1) - preposition (2) - NP_2 (3) - copula *byt'* (4) - Adj/Part (5) - spillover (6–9). Ungrammatical conditions are red, grammatical ones are blue. Conditions where the gender of the attractor and the predicate coincide (for example, MMM and MFF) have dark colors, conditions where they do not (for example, MFM and MMF) have light colors. **(A)** Masculine head, feminine and masculine attractors, **(B)** Masculine head, neuter and masculine attractors.

TABLE 8 | Results of the analysis for Experiment 2b.

Conditions	Region	Factors	df	MS_{effect}	F_1	*p*	df	MS_{effect}	F_2	*p*
MM vs. MF	5	Gram	**1.31**	**231038.53**	**17.55**	**<0.01**	**1.15**	**104660.34**	**15.92**	**<0.01**
		GenMatch	**1.31**	**19014.94**	**16.92**	**<0.01**	1.15	9218.4	0.93	0.35
		Gram * GenMatch	1.31	1974.28	0.66	0.42	1.15	887.3	0.13	0.72
	6	Gram	**1.31**	**42715.99**	**39.18**	**<0.01**	**1.15**	**23558.41**	**10.52**	**0.01**
		GenMatch	**1.31**	**11614.79**	**9.32**	**0.01**	1.15	6878.63	1.88	0.19
		Gram * GenMatch	1.31	1445.20	1.07	0.31	1.15	688.41	0.22	0.65
	7	Gram	**1.31**	**9604.98**	**12.30**	**<0.01**	1.15	4761	1.46	0.25
		GenMatch	**1.31**	**11881.11**	**20.57**	**<0.01**	1.15	7267.56	2.65	0.12
		Gram * GenMatch	1.31	4900.50	2.89	0.10	1.15	3825.42	1.19	0.29
MM vs. MN	5	Gram	**1.31**	**137333.85**	**33.02**	**<0.01**	**1.15**	**79129.69**	**10.12**	**0.01**
		GenMatch	1.31	14897.54	3.58	0.07	1.15	10686.39	1.66	0.22
		Gram * GenMatch	1.31	10.64	<0.01	0.96	1.15	129.96	0.02	0.89
	6	Gram	**1.31**	**75044.22**	**14.43**	**<0.01**	**1.15**	**36864.00**	**8.80**	**0.01**
		GenMatch	1.31	3404.16	1.34	0.26	1.15	1161.11	0.37	0.56
		Gram * GenMatch	1.31	2087.39	0.88	0.36	1.15	704.9	0.12	0.73
	7	Gram	**1.31**	**57868.02**	**23.86**	**<0.01**	**1.15**	**32761.00**	**8.15**	**0.01**
		GenMatch	**1.31**	**10235.23**	**7.27**	**0.01**	1.15	5016.18	1.25	0.28
		Gram * GenMatch	1.31	1362.42	0.53	0.47	1.15	419.23	0.1	0.76

Analyses with $p \leq 0.05$ are shown in bold.

4.5. Discussion

The results of this experiment show that the basic finding from Experiment 2a holds: there is no evidence for agreement attraction in the sentences with M heads. The plots of the data also suggest that the unbalanced frequencies in Experiment 2a had some influence on reading times. In Experiment 2b, where this confounding factor was excluded, two ungrammatical and two grammatical conditions pattern more closely together within each condition set. Still, the conditions where the genders of heads and attractors are mismatched have longer RTs.

Notably, this difference in RTs is not an instance of ungrammaticality illusion, since it is observed in both grammatical and ungrammatical conditions. In case of illusions, a different pattern would be expected: gender mismatch between the head and the attractor should increase RTs in grammatical conditions and decrease RTs in ungrammatical ones. Rather, it can be suggested that gender mismatch carries some processing cost in the sentences with M heads. In any case, our data do not allow for strong claims: the main effect Gender Match is significant in by subjects analysis in regions 5–7, but never reaches significance in by items analysis.

Since the outcome of comprehension experiments was not parallel to the results of Experiment 1 and earlier experiments on Slovak (Badecker and Kuminiak, 2007), we decided to look

at the remaining combinations of head and attractor genders in Experiment 3 before suggesting an explanation.

5. EXPERIMENT 3

In this experiment, we studied sentences with N heads and N, F, and M attractors and sentences with F heads and F, N, and M attractors in comprehension. NF and FN combinations have not been examined before, and we added M attractors to be able to compare sentences with all possible attractors.

5.1. Participants

Thirty-nine native Russian speakers (22 female, 17 male) took part in the experiment. Ages ranged from 19 to 40 (mean age 25.4, *SD* 6.4).

5.2. Materials

We constructed 36 sets of stimuli according to the same schema as in Experiments 2a and 2b and observing the same constraints. Half of the sets had F head nouns and the other half had N head nouns. In all sets, we used M, N, and F attractors. Their frequency was closely matched inside the three groups of conditions, as **Table 9** shows. Half of the predicates were grammatical, and half were not. As a result, every target sentence appeared in six conditions: NNN, NNF, NMN, NMM, NFN, NFF for the sentences with N heads and FFF, FFN, FMF, FMM, FNF, FNN for the sentences with F heads. Thus, out of all possible combinations of head, attractor and predicate genders, we did not use NNM and FFM. We decided to do so to keep the number of grammatical and ungrammatical conditions equal and sacrificed two conditions without any potential for agreement attraction that we have already looked at in Experiment 2a. Additionally, we used 100 fillers from Experiment 2a. Experimental sentences were distributed into six experimental lists, with factors counterbalanced. As a result, we had 136 sentences per list (18 ungrammatical and 118 grammatical), making the grammatical-to-ungrammatical ratio 6.6:1.

5.3. Procedure

The procedure was the same as in Experiments 2a and 2b. An experimental session lasted around 11 min.

5.4. Results

We analyzed participants' question-answering accuracy and reading times. The data from three participants were discarded because they had <75% accuracy in comprehension questions. As a result, we had 36 participants, six for each experimental list. None of them made more than two mistakes when answering questions to target sentences (i.e., 12.5% at most).

As in the previous experiments, reading times that exceeded a threshold of 2.5 standard deviations, by region and condition, were excluded. In total, 1.8% of the data were excluded (never more than 3.7% per region and condition). Average RTs per region in different conditions are presented in **Figure 3** (notice that coloring conventions are different from the previous plots).

TABLE 9 | Frequencies of the attractors used in Experiment 3 (in ipm, or instances per million).

Head gender	Attractor gender	Mean attractor freq (ipm)
F	F	83.2
	M	76.6
	N	79.8
N	F	92.1
	M	92.8
	N	95.7

In Experiments 2a and 2b, we observed agreement attraction for some combinations of head and attractor genders (FM and NM), but not for the others (MF and MN). So the first question we asked in this experiment was whether there would be agreement attraction in NF and FN combinations. If the answer was yes, we were going to compare N and F attractors to M attractors. To answer the first question, we took two groups of conditions: FFF, FFN, FNF, FNN and NNN, NNF, NFN, NFF, and analyzed RTs using 2 x 2 Repeated Measures ANOVAs with grammaticality and gender match as factors, as in the previous experiments. Significant results were found only in regions 5 (adjective or participle) and 6–7 (a spillover region). They are presented in **Table 10**.

5.4.1. Feminine Head, Neuter Attractor

The main effect of Grammaticality was significant in analysis by subjects and by items in regions 5–6. This reflects the fact that ungrammatical sentences were read slower than grammatical ones. The main effect of Gender Match was significant only in analysis by subjects in regions 5–6. The interaction of Grammaticality and Gender Match was significant in analysis by subjects and by items in region 6 and only in analysis by subjects in region 5. Ungrammatical sentences were read faster if the head and the attractor were mismatched in gender (i.e., in the FNN condition compared to the FFN condition). This is the classical attraction pattern, also known as a grammaticality illusion. At the same time, there are no differences between grammatical conditions, i.e., no evidence of ungrammaticality illusions was found.

5.4.2. Neuter Head, Feminine Attractor

The results were the same as in the other set of conditions. Thus, the answer to our first experimental question was positive, so we proceeded to compare the size of the attraction effect for attractors of different genders. We compared two groups of conditions: FNF, FNN, FMF, FMM and NFN, NFF, NMN, NMM. We used 2 x 2 Repeated Measures ANOVAs with grammaticality and attractor gender as factors. Only the main effect of Grammaticality in region 6 was statistically significant [for conditions with *F* heads, $F_1(1, 35) = 19.31$, $p < 0.01$, $MS_{effect} = 86064.00$; $F_2(1, 17) = 10.17$, $p = 0.01$, $MS_{effect} = 24457.35$; for conditions with N heads, $F_1(1, 35) = 55.80$, $p < 0.01$, $MS_{effect} = 126973.44$; $F_2(1, 17) = 7.32$, $p = 0.02$, $MS_{effect} = 52915.47$]. The main effect of Attractor Gender or the interaction between the factors were not significant in any region.

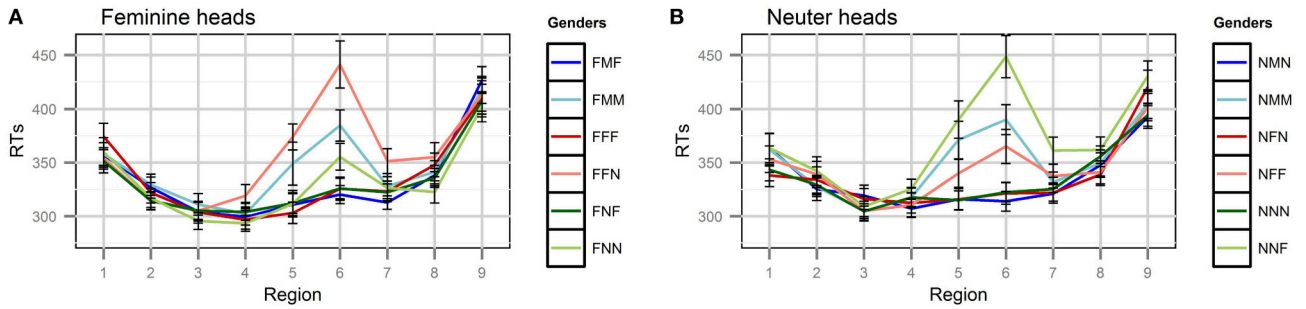


FIGURE 3 | Plots of mean RTs (in ms) by conditions in Experiment 3. Error bars represent standard errors of the means. Regions: NP_1 (1) - preposition (2) - NP_2 (3) - copula *byt'* (4) - Adj/Part (5) - spillover (6–9). The conditions with M attractors are blue, with F attractors - red, with N attractors - green. Dark colors indicate grammatical conditions, light colors - ungrammatical conditions. **(A)** Feminine heads, **(B)** Neuter heads.

TABLE 10 | Results of the analysis for Experiment 3.

Conditions	Region	Factor	df	MS_{effect}	F_1	*p*	df	MS_{effect}	F_2	*p*
FF vs. FN	5	Gram	**1.35**	**42202.85**	**13.54**	**<0.01**	**1.17**	**27667.36**	**14.05**	**<0.01**
		GenMatch	**1.35**	**24964.00**	**8.73**	**0.01**	1.17	14489.69	3.89	0.07
		GenMatch * Gram	**1.35**	**45411.61**	**12.98**	**<0.01**	1.17	13689.61	3.50	0.08
	6	Gram	**1.35**	**187272.56**	**17.49**	**<0.01**	**1.17**	**89662.01**	**12.62**	**<0.01**
		GenMatch	**1.35**	**67487.38**	**20.50**	**<0.01**	1.17	24857.07	2.33	0.15
		GenMatch * Gram	**1.35**	**65732.41**	**11.47**	**<0.01**	**1.17**	**40945.68**	**7.26**	**0.02**
NN vs. NF	5	Gram	**1.35**	**85176.42**	**23.48**	**<0.01**	**1.17**	**41103.22**	**7.33**	**0.02**
		GenMatch	**1,35**	**20168.73**	**5.43**	**0.03**	1.17	6258.94	1.05	0.32
		GenMatch * Gram	**1.35**	**25434.93**	**5.58**	**0.02**	1.17	7525.60	2.48	0.13
	6	Gram	**1.35**	**284248.92**	**32.88**	**<0.01**	**1.17**	**136808.09**	**41.17**	**<0.01**
		GenMatch	**1.35**	**79120.31**	**12.49**	**<0.01**	1.17	25233.81	3.98	0.06
		GenMatch * Gram	**1.35**	**76130.01**	**7.65**	**0.01**	**1.17**	**29658.30**	**4.78**	**0.04**

Analyses with $p \leq 0.05$ are shown in bold.

5.5. Discussion

Let us summarize the results of Experiments 2a, 2b, and 3. Gender agreement attraction was observed with F heads and M or N attractors and with N heads and M or F attractors, but not with M heads and F or N attractors. This leads us to the conclusion that attraction depends primarily on the features of the head rather than on the features of the attractor. If the features of the attractor played an additional role, ungrammatical sentences with M attractors would be read *faster* than ungrammatical sentences with other attractors. However, when we compared sentences with F heads and N or M attractors and sentences with N heads and F or M attractors, the Attractor Gender or the interaction between this factor and Grammaticality never reached significance, and average RTs even showed the opposite pattern: they were *longer* in the ungrammatical sentences with M attractors. This goes against the assumptions entertained in the absolute majority of previous agreement attraction studies, so a detailed analysis of this result will be presented in the General Discussion Section.

6. GENERAL DISCUSSION

In this paper we reported four experiments on gender agreement attraction in Russian. We observed attraction effects both in production and in comprehension. Badecker and Kuminiak (2007) is the only previous production study where gender agreement attraction was examined in a language with three genders (Lorimor et al., 2008 elicited very few gender errors in their experiments on Russian). In this paper, we replicated one of Badecker and Kuminiak's experiments and conducted the first comprehension experiments analyzing attraction with non-binary features.

Two outcomes of our experiments can be identified as the most important. Firstly, our results suggest that gender attraction works differently in production and comprehension. This does not agree with previous studies of number agreement attraction, in which production and comprehension results were largely parallel: only the combination of a singular head and a plural attractor triggered attraction. Secondly, our reading experiments

suggest that the features of the head, rather than the features of the attractor are crucial to determine the pattern of agreement attraction, while the absolute majority of previous agreement attraction studies rely on the opposite assumption.

6.1. Overview of Experimental Findings

In our comprehension experiments, attraction was observed in some combinations of head and attractor genders, but not in the others, while in the production experiment, all combinations exhibited attraction, only to a different extent. We will first consider production results, and then comprehension findings. The outcome of the production study was similar to the results of the first experiment conducted by Badecker and Kuminiak (2007): there were more errors with MF subjects than with FM subjects and with NM subjects than with MN subjects. Both differences were statistically significant in the Slovak study, while in our experiment, only the first one was.

Badecker and Kuminiak ran an additional experiment comparing NF and NM preambles and found that the error rates in these conditions were roughly the same. They claim that this pattern can be explained only in an optimality-theoretic framework where markedness effects are by definition relational. We believe that this is not the case. Given the impressive body of literature on number and gender features, we do not think that we can select a particular approach based on experimental data without a detailed consideration of other arguments. So we chose two models that have been applied to Russian to demonstrate that they are also compatible with the pattern described by Badecker and Kuminiak and may be better suited to explain other findings we reported.

In Kramer (2015), F is encoded as [+FEM], M is [−FEM] and N corresponds to no gender features. When zero and non-zero feature values are compared, the latter are marked, and it can be argued that for this comparison, it is not important whether non-zero values are plus or minus. Therefore, the same error rates are observed with NF and NM preambles. When non-zero values are compared, plus values are more marked. In Nevins (2011), F is [+FEM], [−MASC], M is [−FEM], [+MASC] and N is [−FEM], [−MASC]. N is less marked than M and F because it contains only minus values, while M and F both contain one plus value. But when we compare F and M directly, it can be argued that feature hierarchy becomes important. [FEM] is standardly assumed to be lower than [MASC], so F is more marked than M.

Now let us focus on another property of production findings from Slovak and Russian that is not discussed by Badecker and Kuminiak (2007), but seems crucial to us. In case of gender agreement, attraction errors are produced with all preambles in which the genders of the head and the attractor are mismatched, while in case of number agreement, errors are virtually absent with plural heads and singular attractors. One way to capture this would be to assume that all genders are marked by some feature combinations, as Nevins (2011) suggests, while singular corresponds to no number features.

Another important problem is the difference between experimental findings from Slovak and Russian on the one hand and Romance languages on the other. In Russian and Slovak, more errors are produced with MF preambles than with FM preambles, while in Spanish and French the situation is the opposite. Badecker and Kuminiak (2007) do not comment on this discrepancy, and we cannot offer any explanation for it so far. We can only note that the pattern observed in Slovak and Russian is similar to what we see with number: more errors when the head is less marked than the attractor.

Now let us turn to comprehension experiments. Attraction was observed in NMM, NFF, FMM, and FNN conditions, but not in MFF and MNN conditions. As we already noted, this indicates that features of the heads rather than features of the attractors play a crucial role for attraction. Before discussing this finding in the next section, we want to make two important observations. Firstly, the M gender exhibits a different pattern from the F and N genders. This can hardly be attributed to feature markedness: N is the grammatical default in Russian, and the psycholinguistic relevance of this fact is confirmed by the production data discussed above. We will explore alternative explanations below. Secondly, no ungrammaticality illusions (differences between grammatical conditions depending on whether the head and the attractor have matched or mismatched gender features) were detected in our experiments, which lends further support to the retrieval approach to agreement attraction.

6.2. The Role of Head and Attractor Features in Attraction

In the literature on agreement attraction, the presence or absence of the effect is traditionally associated with the features of the attractor. There are at least two reasons for this. Firstly, experimental findings suggest that *some* properties of attractors do influence attraction effects [e.g., as we discussed in the introduction, Hartsuiker et al. (2003) showed that the incidence of agreement errors was much higher when attractors were formally similar to nominative plural forms]. The second reason is tradition: the first proposed account of agreement attraction relied on feature percolation, which means focusing exclusively on the attractor whose features can erroneously spread upwards.

The assumption that the features of the attractor are crucial has been maintained in the more recent retrieval account. However, it is important to realize that in this account the properties of the head can influence the agreement process as well. For example, to explain the plural markedness effect, it is traditionally assumed that singular nouns are not marked for number, and "the system is biased to return explicitly number marked constituents" (Wagers et al., 2009, p. 233), therefore plural attractors can easily be retrieved, while singular ones almost never are. But another interpretation is possible: the plural feature makes the heads easier to retrieve and thus more stable, less prone to attraction errors. This is why attraction in the plural-singular configurations is virtually non-existent. On the other hand, the retrieval of singular heads is prone to error, hence the abundance of errors in singular-plural configurations[11].

While we look at binary features or at the cases where attraction is observed in all feature combinations (as in production experiments on Slovak and Russian), we can only

[11]Let us note that under this scenario an attractor can also be retrieved in a singular-singular configuration, but this will not provoke any agreement errors.

use indirect evidence to estimate the contribution of head and attractor features to the agreement process. Our reading experiments allow for the first direct comparison and show that at least in comprehension, the features of heads, not attractors play the crucial role. We observed attraction with attractors of all three genders, but only with N and F heads. The gender of the attractor did not even influence the size of the effect. These results suggest that the gender of the attractor has very little or no influence on its chances to be retrieved (it should only match the gender of the incorrect verb form).

Notably, Julie Franck expressed similar ideas in a recent talk (Franck, 2015). The first part of the talk was dedicated to summarizing existing data on agreement attraction. Franck adopted the retrieval approach for production and comprehension and identified the following groups of factors that can lead to attraction: semantic factors (primarily related to the conceptual numerosity of the subject NP), stability of the head's features, accessibility of the attractor (defined by its structural position) and similarity between the head and the attractor. Discussing stability of the head's features Franck examined asymmetries between feature values, morphophonological and semantic influences.

Franck's reexamination of attraction phenomena was driven by the findings on morphophonology (other data she considered could be accounted for in the old models). As we noted in the introduction, studies on several languages demonstrated that number and gender agreement attraction errors are less frequent when heads have regular inflections, but this plays no role for attractors (e.g., Bock and Eberhard, 1993; Vigliocco et al., 1995; Vigliocco and Zilli, 1999; Franck et al., 2008). For attractors, only morphological ambiguity making them more similar to a subject is important (e.g., Hartsuiker et al., 2003; Badecker and Kuminiak, 2007)[12]. This led Franck to conclude that the features of the head are crucial, and she reanalyzed existing data according to this idea. She argued that features that have a semantic correlate are more resistant to attraction (for example, Vigliocco and Franck, 1999 observed lower error rates when heads had conceptual rather than purely grammatical gender) and that the same is true for marked feature values. The latter conclusion was based on number agreement attraction findings and on the results of Badecker and Kuminiak's and our production experiments.

Thus, the findings summarized by Franck and the outcome of our reading experiments point into the same direction, but we still have to explain the difference between our comprehension and production results. Of course, to make definitive conclusions, it would be great to have data from several languages (for example, comprehension data from Slovak), but let us suggest several hypotheses based on existing findings. Our reading experiments strongly indicate that M heads are resistant to attraction, while N and F heads are not. The data from production experiments on Russian and Slovak are open to several interpretations because attraction was observed in all head-attractor combinations with mismatched genders. Therefore, we assume that M heads in general are the most stable ones and the least prone to attraction, and production data need an independent explanation. This assumption is supported by independent evidence: several production experiments on number agreement attraction in Russian reported by Nicol and Wilson (1999) and Yanovich and Fedorova (2006) demonstrated that the incidence of number errors depends on the gender of the head noun. Errors arise most often with N heads and least often with M ones.

If our assumption is on the right track, M heads and plural heads exhibit similar properties in comprehension. But why should they do so, given that M features are neither the most marked nor the least marked in Russian? Let us come back to the idea expressed in the previous subsection: number is privatively marked (i.e., singular nouns have no number features), while gender is not (all nouns have some gender features with plus and minus values). We hypothesize that with privative features, the non-zero value is the most stable, while with non-privative features, where all values are non-zero, other considerations come into the picture. We are reluctant to appeal to frequency, but maybe it plays a role that M gender vastly outnumbers F and N in Russian. In any case, our data indicate that that there is no straightforward relation between feature markedness and stability. The next subsection considers some differences between comprehension and production and how these differences could explain our results.

6.3. Differences between Production and Comprehension

Based on parallel results from number agreement attraction experiments most authors assume that the same mechanisms underlie attraction in production and comprehension. The opposite view has been recently advocated by Tanner et al. (2014). They claim that the mechanisms responsible for attraction in comprehension are a subset of those involved in production. In particular, they argue that attraction in comprehension is due to retrieval interference, while attraction in production is best described by the representational account, namely, by the Marking and Morphing model (Eberhard et al., 2005), although retrieval interference is also present.

As we noted above, the Marking and Morphing model is incompatible with gender agreement attraction. We believe that the core mechanism underlying number and gender agreement attraction in production is the same, so we opt for the retrieval approach. Evidently, in case of number, semantic factors influence agreement, and it is expected that their influence is much more readily detected in production than in comprehension: in production, we start with the conceptual structure, while in comprehension, it is our goal. Vigliocco and Franck (1999) demonstrated that gender agreement attraction errors are less frequent when head nouns have conceptual, rather than purely grammatical gender. So semantic factors also play a role here, but, given the relevant distinctions[13] between number

[12]Let us add that Badecker and Kuminiak (2007) demonstrated that ambiguity is important not only for attractors, but also for heads: if the form is ambiguous between nominative and accusative, the chances of the head to be retrieved are lower.

[13]In case of number, we can find many words that are formally singular, but denote plural entities, for example, nouns like *crowd* or heads of the phrases like *the label on the bottles* that have a distributive interpretation. Gender is usually semantically

and gender, this role is different: they mainly reduce the size of the effect. It would be very interesting to assess their influence on gender agreement attraction in comprehension: we expect that it should be much smaller, as in case of number agreement. Thus, the differences between production and comprehension noted by Tanner et al. (2014) may also be relevant for gender agreement, but the picture revealed by our experiments cannot be explained by them.

In the previous subsection we argued that agreement attraction patterns in comprehension are due to the fact that heads with plural features and M features are resistant to attraction, i.e., that during the retrieval process, they tend to be identified correctly, while the retrieval of heads with other features can be disturbed by attractors. Findings summarized by Franck (2015) show that the stability of head's features should also be relevant for agreement attraction in production. This is further confirmed by the results from Nicol and Wilson (1999) to Yanovich and Fedorova (2006) indicating that heads with M features are indeed more stable when we look at number agreement production in Russian. Based on these data, we would expect to see no errors in MF and MN conditions in production experiments on gender agreement, but this is not what we found.

To address this problem, we should specify in more detail how retrieval may work in comprehension and production. Wagers et al. (2009) who analyze comprehension show that the retrieval account has two versions that may be difficult to tease apart based on the current experimental data. On the one hand, cue-based retrieval may be initiated every time we deal with an agreeing verb. On the other hand, we may predict the features of the upcoming verb relying on the subject NP and initiate retrieval only when our predictions are not met. Both versions give roughly the same results if we assume that when the true subject matches all the cues, it is successfully retrieved in the absolute majority of cases. Then in both scenarios, problems are expected only when we encounter an incorrect verb form and the sentence contains an attractor a non-subject NP that matches the incorrectly specified feature of the verb.

We believe that two similar scenarios can also be distinguished for production: we can decide which features we need on an agreeing predicate while processing the subject or once we get to the predicate. Accordingly, retrieval might be initiated every time we deal with an agreeing predicate or only when a wrong verb form that does not match our predictions is spuriously generated. The models proposed by Solomon and Pearlmutter (2004) or by Badecker and Kuminiak (2007) instantiate the first scenario. For example, Solomon and Pearlmutter argue that attraction in production arises because two nouns, the head of the subject NP and the attractor, are simultaneously active in the syntactic structure, and a wrong agreement controller may be selected. However, we argue for the second scenario below.

To summarize, in comprehension, we construct the set of retrieval cues based on the verb form that is provided to us. As we demonstrated above, different versions of the account share this basic observation. If the first scenario is adopted for production (the features of the upcoming verb are predicted, and retrieval is initiated only when we spuriously generate a wrong verb form), the picture should be quite similar: the set of cues will be based on this form.

However, we do not believe that this scenario is the most plausible. In particular, it implies that we generate the subject NP with all its feature specifications before we turn to the verb. In reality, the process should be much more complicated. On the one hand, we cannot determine the case of an NP before we select the predicate (for example, experiencers may receive nominative, accusative, or dative case in Russian, depending on the verb, so it is impossible to plan a nominative NP having only some abstract V in mind). On the other hand, we cannot select some features of the verb form without looking at the subject.

This leads us to adopt the second scenario, in which the relevant features are retrieved at some point during the derivation, rather than predicted and then rechecked. Then we do expect certain differences between production and comprehension. Namely, under the second production scenario it is not the case that we look for an NP with a particular number or gender feature. Rather, we look for the values of number and gender features inside the subject NP. These features should belong to the head of this NP, but sometimes we spuriously pay attention to the features of other nouns. We hypothesize that feature markedness plays a role in this process, and this is what causes different outcomes in our production and comprehension experiments.

To explain how markedness effects may arise, let us summarize different factors that have been shown to play a role for retrieval. More stable head nouns have more chances to be retrieved than less stable ones. Structurally accessible attractors looking like subjects have more chances to be retrieved than the attractors without these characteristics. This is true both for production and for comprehension. And, independently of these factors, marked features have more chances to be retrieved. In comprehension, when we encounter a particular verb form and construct a set of retrieval cues based on it, different number or gender features do not compete with each other: we always look for a particular value. In production, we need to find the value of the gender feature of the subject NP, there is no value that is provided in advance, thus different values may enter the competition[14]. Thus, production involves competition and comprehension does not, therefore we can observe feature markedness effects in production, but not in comprehension. This is why production and comprehension results for gender agreement are different. We do not observe any differences in case of number agreement because plural is at the same time a more stable feature and a marked one. This is a very tentative hypothesis, so further experiments are necessary to test it or to suggest an alternative explanation for the observed asymmetry between production and comprehension findings.

empty, and when it is not, the conceptual and formal gender typically coincide (and thus the former reinforces the latter). If they do not coincide, it never depends on the properties of modifiers, only on the noun itself (for example, *vrač* "doctor$_M$" can refer both to a man and to a woman in Russian).

[14]In our production experiment, participants were provided with predicates in a particular form. Still, we also expect competition here because participants had to produce a correct form if the provided form was wrong, and to do so, they had to retrieve the subject NP and determine its gender.

AUTHOR CONTRIBUTIONS

All authors listed have made substantial, direct and intellectual contribution to the work, and approved it for publication.

ACKNOWLEDGMENTS

The study was partially supported by the grant #16-18-02071 from the Russian Science Foundation. We are grateful to many colleagues for their valuable comments and would especially like to thank Colin Phillips. We are also very grateful to the reviewers.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpsyg.2016.01651/full#supplementary-material

REFERENCES

Acuña-Fariña, J. C., Meseguer, E., and Carreiras, M. (2014). Gender and number agreement in comprehension in Spanish. *Lingua* 143, 108–128. doi: 10.1016/j.lingua.2014.01.013

Anton-Mendez, M. I., Nicol, J., and Garrett, M. F. (2002). The relation between gender and number agreement processing. *Syntax* 5, 1–25. doi: 10.1111/1467-9612.00045

Badecker, W., and Kuminiak, F. (2007). Morphology, agreement and working memory retrieval in sentence production: evidence from gender and case in Slovak. *J. Mem. Lang.* 56, 65–85. doi: 10.1016/j.jml.2006.08.004

Badecker, W., and Lewis, R. (2007). "A new theory and computational model of working memory in sentence production: agreement errors as failures of cue-based retrieval," in *Talk Given at 20th Annual CUNY Sentence Processing Conference*. San Diego, CA: University of California. Available online at: http://crl.ucsd.edu/cuny2007/program/235_Abstract.pdf

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01

Bock, K., and Cutting, J. C. (1992). Regulating mental energy: performance units in language production. *J. Mem. Lang.* 31, 99–127.

Bock, K., and Eberhard, K. M. (1993). Meaning, sound and syntax in english number agreement. *Lang. Cogn. Process.* 8, 57–99.

Bock, K., and Miller, C. (1991). Broken agreement. *Cogn. Psychol.* 23, 45–93.

Brehm, L., and Bock, K. (2013). What counts in grammatical number agreement? *Cognition* 128, 149–169. doi: 10.1016/j.cognition.2013.03.009

Cohen, J. D., MacWhinney, B., Flatt, M. R., and Provost, J. (1993). Psyscope: a new graphic interactive environment for designing psychology experiments. *Behav. Res. Methods Instrum. Comput.* 25, 257–271.

Cowan, N. (2001). The magical number 4 in short-term memory: a reconsideration of mental storage capacity. *Behav. Brain Sci.* 24, 87–185. doi: 10.1017/S0140525X01003922

Dillon, B., Mishler, A., Sloggett, S., and Phillips, C. (2013). Contrasting intrusion profiles for agreement and anaphora: experimental and modeling evidence. *J. Mem. Lang.* 69, 85–103. doi: 10.1016/j.jml.2013.04.003

Eberhard, K. M. (1999). The accessibility of conceptual number to the processes of subject–verb agreement in English. *J. Mem. Lang.* 41, 560–578.

Eberhard, K. M., Cutting, J. C., and Bock, K. (2005). Making syntax of sense: number agreement in sentence production. *Psychol. Rev.* 112, 531–559. doi: 10.1037/0033-295X.112.3.531

Farkas, D., and de Swart, H. (2010). The semantics and pragmatics of plurals. *Semant. Pragmat.* 3, 1–54. doi: 10.3765/sp.3.6

Francis, W. N. (1986). Proximity concord in English. *J. Eng. Linguist.* 19, 309–317.

Franck, J. (2015). "Attraction errors: novel insights into the relationship between grammar and processing," in *Talk Given at the "Agreement across Boards" Conference*, (Zadar, Croatia).

Franck, J., Vigliocco, G., Antón-Méndez, I., Collina, S., and Frauenfelder, U. H. (2008). The interplay of syntax and form in sentence production: a cross-linguistic study of form effects on agreement. *Lang. Cogn. Process.* 23, 329–374. doi: 10.1080/01690960701467993

Franck, J., Vigliocco, G., and Nicol, J. (2002). Subject-verb agreement errors in French and English: the role of syntactic hierarchy. *Lang. Cogn. Process.* 17, 371–404. doi: 10.1080/01690960143000254

Hartsuiker, R. J., Schriefers, H., Bock, K., and Kikstra, G. M. (2003). Morphophonological influences on the construction of subject-verb agreement. *Mem. Cogn.* 31, 1316–1326. doi: 10.3758/BF03195814

Haskell, T. R., and MacDonald, M. C. (2005). Constituent structure and linear order in language production: evidence from subject-verb agreement. *J. Exp. Psychol. Learn. Mem. Cogn.* 31, 891–904. doi: 10.1037/0278-7393.31.5.891

Haspelmath, M. (2006). Against markedness (and what to replace it with). *J. Linguist.* 42, 25. doi: 10.1017/S0022226705003683

Jespersen, O. (1924). *The Philosophy Of Grammar*. London: Allen and Unwin.

Just, M. A., Carpenter, P. A., and Woolley, J. D. (1982). Paradigms and processes in reading comprehension. *J. Exp. Psychol.* 111, 228.

Kramer, R. (2015). *The Morphosyntax of Gender*. Oxford: Oxford University Press.

Lewis, R. L., and Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cogn. Sci.* 29, 375–419. doi: 10.1207/s15516709cog0000_25

Lorimor, H., Bock, K., Zalkind, E., Sheyman, A., and Beard, R. (2008). Agreement and attraction in Russian. *Lang. Cogn. Process* 23, 769–799. doi: 10.1080/01690960701774182

Lyashevskaya, O. N., and Sharoff, S. A. (2009). *Chastotnyj Slovar' Sovremennogo Russkogo Jazyka (Frequency Dictionary of Modern Russian Laguage)*. Moscow: Azbukovnik.

Martin, A. E., Nieuwland, M. S., and Carreiras, M. (2014). Agreement attraction during comprehension of grammatical sentences: ERP evidence from ellipsis. *Brain Lang.* 135, 42–51. doi: 10.1016/j.bandl.2014.05.001

McElree, B. (2006). Accessing recent events. *Psychol. Learn. Motiv.* 46, 155–200. doi: 10.1016/S0079-7421(06)46005-9

Mirkovic, J., and MacDonald, M. C. (2013). When singular and plural are both grammatical: semantic and morphophonological effects in agreement. *J. Mem. Lang.* 69, 277–298. doi: 10.1016/j.jml.2013.05.001

Nevins, A. (2011). Marked triggers vs. marked targets and impoverishment of the dual. *Linguist. Inq.* 42, 413–444. doi: 10.1162/LING_a_00052

Nicol, J., Foster, K., and Veres, C. (1997). Subject-verb agreement processes in comprehension. *J. Mem. Lang.* 36, 569–587.

Nicol, J., and Wilson, R. (1999). "Agreement and case-marking in Russian: a psycholinguistic investigation of agreement errors in production," in *The Eight Annual Workshop on Formal Approaches to Slavic Languages. The Philadelphia Meeting* (Ann Arbor, MI: Michigan Slavic Publications), 314–327.

Pearlmutter, N. J., Garnsey, S. M., and Bock, K. (1999). Agreement processes in sentence comprehension. *J. Mem. Lang.* 41, 427–456.

Quirk, R., Greenbaum, S., Leech, G. N., and Svartvik, J. (1972). *A Grammar Of Contemporary English*. Oxford: Oxford University Press.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychol. Bull.* 114, 510.

Sauerland, U., Anderssen, J., and Yatsushiro, K. (2005). "The plural is semantically unmarked," in *Linguistic Evidence*, eds S. Kepser and M. Reis (Berlin: Mouton de Gruyter), 413–434.

Slioussar, N., and Samoilova, M. (2014). "A database to estimate frequencies of different grammatical features and inflectional affixes in Russian nouns," in *The 9th International Conference on the Mental Lexicon* (Niagara-on-the-Lake: Brock University and McMaster University), 104–105.

Solomon, E. S., and Pearlmutter, N. J. (2004). Semantic integration and syntactic planning in language production. *Cogn. Psychol.* 49, 1–46. doi: 10.1016/j.cogpsych.2003.10.001

Staub, A. (2009). On the interpretation of the number attraction effect: response time evidence. *J. Mem. Lang.* 60, 308–327. doi: 10.1016/j.jml.2008.11.002

Staub, A. (2010). Response time distributional evidence for distinct varieties of number attraction. *Cognition* 114, 447–454. doi: 10.1016/j.cognition.2009.11.003

Tanner, D., Nicol, J., and Brehm, L. (2014). The time-course of feature interference in agreement comprehension: multiple mechanisms and asymmetrical attraction. *J. Mem. Lang.* 76, 195–215. doi: 10.1016/j.jml.2014.07.003

Van Dyke, J. A., and Johns, C. L. (2012). Memory interference as a determinant of language comprehension. *Lang. Linguist. Compass* 6, 193–211. doi: 10.1002/lnc3.330

Vigliocco, G., Butterworth, B., and Garrett, M. F. (1996). Subject-verb agreement in Spanish and English: differences in the role of conceptual constraint. *Cognition* 61, 261–298.

Vigliocco, G., Butterworth, B., and Semenza, C. (1995). Constructing subject-verb agreement in speech: the role of semantic and morphological factors. *J. Mem. Lang.* 34, 186–215.

Vigliocco, G., and Franck, J. (1999). When sex and syntax go hand in hand: Gender agreement in language production. *J. Mem. Lang.* 40, 455–478.

Vigliocco, G., and Nicol, J. (1998). Separating hierarchical relations and word order in language production: is proximity concord syntactic or linear? *Cognition* 68, B13–B29.

Vigliocco, G., and Zilli, T. (1999). Syntactic accuracy in sentence production: the case of gender disagreement in Italian language-impaired and unimpaired speakers. *J. Psycholinguist. Res.* 28, 623–648.

Wagers, M. W., Lau, E. F., and Phillips, C. (2009). Agreement attraction in comprehension: representations and processes. *J. Mem. Lang.* 61, 206–237. doi: 10.1016/j.jml.2009.04.002

Yanovich, I. (2012). "What can Russian gender tell about the semantics of phi-features?," in *Formal Approaches to Slavic Linguistics 21. Indiana University, Bloomington. Talk given at FASL 21, Indiana University*. Available online at: http://www.sfs.uni-tuebingen.de/~yanovich/papers/Yanovich_FASL_2012_handout.pdf

Yanovich, I., and Fedorova, O. (2006). "Subject-verb agreement errors in Russian: head noun gender effect," in *Proceedings of "Dialog 2006"*. Available online at: http://www.dialog-21.ru/digests/dialog2006/materials/html/Yanovich2.htm

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2016 Slioussar and Malko. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 02 May 2016
doi: 10.3389/fpsyg.2016.00548

Minimal Interference from Possessor Phrases in the Production of Subject-Verb Agreement

Janet L. Nicol[1,2]*, *Andrew Barss*[2] *and Jason E. Barker*[3]

[1] *Department of Linguistics, Program in Cognitive Science, University of Arizona, Tucson, AZ, USA,* [2] *Department of Psychology, Program in Cognitive Science, University of Arizona, Tucson, AZ, USA,* [3] *Formerly Affiliated with Department of Psychology, Program in Cognitive Science, University of Arizona, Tucson, AZ, USA*

We explore the language production process by eliciting subject-verb agreement errors. Participants were asked to create complete sentences from sentence beginnings such as *The elf's/elves' house with the tiny window/windows* and *The statue in the elf's/elves' gardens.* These are subject noun phrases containing a head noun and controller of agreement (*statue*), and two nonheads, a "local noun" (*window(s)/garden(s)*), and a possessor noun (*elf's/elves'*). Past research has shown that a plural nonhead noun (an "attractor") within a subject noun phrase triggers the production of verb agreement errors, and further, that the nearer the attractor to the head noun, the greater the interference. This effect can be interpreted in terms of relative hierarchical distance from the head noun, or via a processing window account, which claims that during production, there is a window in which the head and modifying material may be co-active, and an attractor must be active at the same time as the head to give rise to errors. Using possessors attached at different heights within the same window, we are able to empirically distinguish these accounts. Possessors also allow us to explore two additional issues. First, case marking of local nouns has been shown to reduce agreement errors in languages with "rich" inflectional systems, and we explore whether English speakers attend to case. Secondly, formal syntactic analyses differ regarding the structural position of the possessive marker, and we distinguish them empirically with the relative magnitude of errors produced by possessors and local nouns. Our results show that, across the board, plural possessors are significantly less disruptive to the agreement process than plural local nouns. Proximity to the head noun matters: a possessor directly modifying the head noun induce a significant number of errors, but a possessor within a modifying prepositional phrase did not, though the local noun did. These findings suggest that proximity to a head noun is independent of a "processing window" effect. They also support a noun phrase-internal, case-like analysis of the structural position of the possessive ending and show that even speakers of inflectionally impoverished languages like English are sensitive to morphophonological case-like marking.

Keywords: subject-verb agreement, possessive, possessor, genitive, production error, attraction error, case marking, semantic integration

OPEN ACCESS

Edited by:
Colin Phillips,
University of Maryland, USA

Reviewed by:
Laurel Brehm,
Northwestern University, USA
Sol Lago,
University of Potsdam, Germany

****Correspondence:***
Janet L. Nicol
nicol@email.arizona.edu

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: *21 November 2015*
Accepted: *01 April 2016*
Published: *02 May 2016*

Citation:
Nicol JL, Barss A and Barker JE (2016) Minimal Interference from Possessor Phrases in the Production of Subject-Verb Agreement. Front. Psychol. 7:548. doi: 10.3389/fpsyg.2016.00548

INTRODUCTION

When speakers produce language, they need to map the elements of a to-be-conveyed proposition onto an appropriate sentence structure, and keep track of these assignments as the utterance is produced. This process is relatively straight-forward for simple sentences like *The key is shiny*, but becomes more challenging when additional information needs to be encoded. For example, in *The key to the cabinets is shiny*, the subject of the verb is the phrase *the key to the cabinets*; however within this phrase, *key* is the head, and *cabinets* is part of a modifying phrase, and the head noun must be selected as the controller of verb agreement. Speakers need to keep this distinction in mind if they are to produce sensible sentences: for the most part, the head noun is the thing the sentence is about, the main element of the predicate's argument (the key, not the cabinets, is what is shiny), and the element a verb may need to agree with. Occasionally, the process goes awry, and a subject-verb agreement error is the result. Studying the variables that affect the incidence of such errors illuminates the language production process.

Subject-verb agreement errors occur with some regularity in both spoken and written language (Jespersen, 1913/1961; Visser, 1963; Quirk et al., 1985; Bock and Miller, 1991). In a seminal paper, Bock and Miller (1991) elicited errors in the laboratory by presenting participants with sentence beginnings, or *preambles*, and asking them to repeat these and create a sentence ending that included a verb. The results showed that agreement errors arise when a singular head is modified by a prepositional phrase containing a plural noun (typically called the *local noun*, or when it is plural, the *attractor*); e.g., *The key to the cabinets were shiny*. The error is not simply due to participants' forgetting the head and implementing agreement "locally" between the attractor and the verb because *The keys to the cabinet* does not elicit errors at the same rate. One explanation for the difference is that the singular is seen as the default: a plural is derived from the singular by the addition of a marked feature, and this plural feature has an autonomy that allows it to intrude on the number specification of a verb. Since that initial study, a great deal of research has explored the kinds of variables that influence the production of agreement errors, and these have led to a refinement and elaboration of syntactic encoding operations in language production.

The focus of this paper is the production of subject-verb agreement errors in English sentences containing a complex subject noun-phrase that includes a singular head noun and a local noun[1], but also a possessor phrase bearing the possessive marker ["s/"], as in (1) and (2). Our experiments examine possessors in two positions: modifying the head noun, as in (1), and modifying a local noun, as in (2).

(1) The women's subscription to the newsletter...
(2) The subscription to the women's newsletter...

The effects of possessors have, until now, been unexplored. In addition to expanding the range of constructions examined in this experimental paradigm, they permit us to explore two issues: (a) the nature of the structural effects that have been argued to influence the presence and magnitude of errors; and (b) the role of the possessive ending as a potential cue to non-subjecthood in potentially reducing errors, akin to the role that overt case marking has been found to play in several languages.

These two issues, and predictions for our experiments, are detailed below.

Proximity Effects

Previous research has shown that an attractor that appears within a modifier that is adjacent to the head noun triggers more errors than one that is located more distantly from the head noun. Several accounts have been offered for this difference, which make contrasting predictions with respect to the behavior of the two types of possessors in (1)–(2).

First, consider some of the empirical findings. Bock and Cutting (1992) found that an attractor within a prepositional phrase (PP) modifier (e.g., 3 and 5 below) elicits significantly more agreement errors than a plural attractor within a clausal modifier (e.g., 4 and 6), and further, that a plural attractor within a relative clause modifier (e.g., 4) elicits more errors than one in a complement clause (e.g., 6).

(3) The editor of the history books
(4) The editor who rejected the books
(5) The dream about the castles
(6) The dream that Anne inherited the castles

Further proximity effects are presented by Franck et al. (2002), who examined contrasts like the following, in which a plural attractor appears inside one of two PP modifiers with different syntactic attachment heights (see **Figure 1**):

(7) The helicopter for the flights over the canyon
(8) The helicopter for the flight over the canyons

A comparison of error rates associated with these sentence types revealed a substantial difference, with the latter eliciting very few errors.

Bock and Cutting (1992) argued that the difference in error rates between (6) and (3–5) was due to the extent to which the head noun and attractor are co-active, and that because a complement clause contains its own subject and predicate, its contents are insulated, in a sense, from the head, making an attractor less likely to be co-active with the head noun. An attractor within a PP or relative clause modifier is not insulated in this way. A variant of the co-activation view is that of Nicol (1995), who proposes that the verb-valuing operation must occur

[1] All noun phrases have head nouns, but we use the term "head noun" in this article, following common practice in the literature, to refer to the noun that is the structural head of the subject noun phrase, and "local noun" to refer to the head noun of a modifier of the subject head noun. Thus in *the treasure in the cave*, *treasure* is the head noun and *cave* is the local noun.

Throughout the paper, we use the term "noun phrase" to refer to phrases like *the key to the cabinets*, although they are analyzed in some current work as maximal projections of a determiner, i.e., a DP (see **Figures 1–4**). The DP may be thought of as the "extended projection" of the noun (Grimshaw, 1990). For most of the discussion the distinction between an NP and a DP is not important, and we use "noun phrase" due to its familiarity.

We refer to examples like *the elf's house* as a possessive construction, and to *the elf* as the possessor.

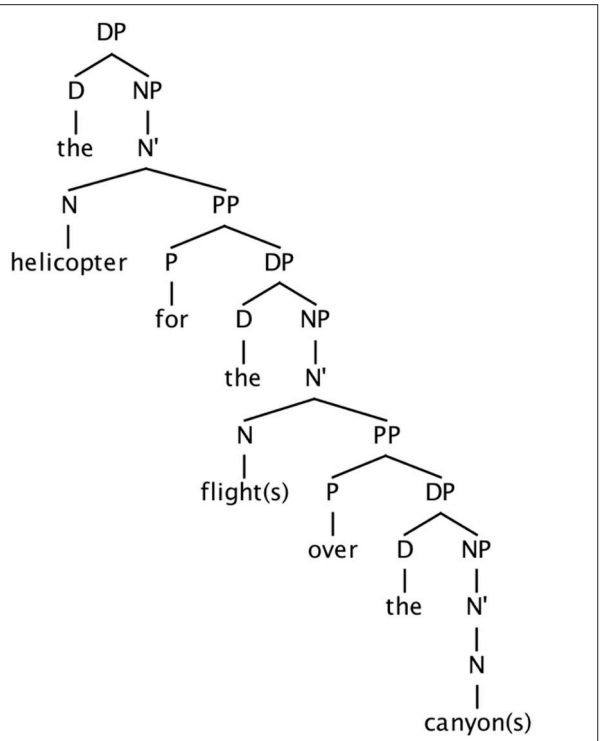


FIGURE 1 | Approximate structure for examples (7) and (8), showing relative embedding of two PP modifiers.

within the limited timeframe in which the subject noun phrase is active. A limited processing window exists, and only noun phrases that are co-active with the head noun (i.e., within the same processing window) will produce errors. This account extends to the complement-clause modifier vs. relative-clause/PP modifier contrast in (3–6), as well as the stacked PP modifiers in (7) and (8). The first modifier noun phrase down [e.g., *flights* in (7)] will be more likely than the deepest modifier (e.g., *canyons*) to be within this processing window, and therefore will be more likely to cause an agreement error. (A similar argument was reiterated by Gillespie and Pearlmutter, 2011).

Note that the processing window hypothesis treats the much lower rate of errors in (6) and (8) as a type of threshold effect: the attractors in those cases lie outside a processing window which includes the head noun and the more local plural attractor in (5) and (7). An attractor is either within the window, in which case it will potentially cause errors, or outside it, in which case error rates will be low. And the experimental results suggest that the processing window including the head noun extends rightward to encompass the first noun phrase, and not the second.

An alternative view is presented by Vigliocco and Nicol (1998). They attributed the difference to *relative* structural proximity of the head noun and attractor: the closer the attractor is to the head noun, the more likely it is to produce an error. In (5) and (7), the plural attractor is closer to the head noun (in terms of hierarchical distance, i.e., nodes separating the two) than the plural attractor is in (6) and (8), and thus more likely to cause errors.

Returning to our possessor phrases in (1) and (2), the two accounts make differing predictions with respect to expected error rates. Note that the possessors are embedded to different extents within the structure of the subject noun phrase. The head-noun-modifying possessor in (1) is more shallowly embedded than the local-noun-modifying possessor in (2), and closer to the head noun. On the relative structural proximity account, head-noun-modifying possessors should produce more errors than the local-noun-modifying possessors. The processing window analysis makes a different prediction: both possessors should lie within the processing window that includes the first PP modifier, and so both types of possessor should produce an equal number of errors. Our experiments compare error rates for the two possessor positions, allowing us to empirically distinguish the two accounts.

The Possessor Ending and Case Marking

A number of studies have examined the effect of case marking of a local noun on attraction errors, and reported that overt case—case marking that is phonologically realized—acts to dampen errors. Case is variation in the form of a noun or determiner that depends on its grammatical function, e.g., subject, object, indirect object, oblique, etc., and is largely redundant with structural information. Yet the additional phonological marking appears to help speakers keep straight which noun is the agreement controller.

Local nouns inflected for case are less likely to produce errors when that case is unambiguously non-nominative (i.e., incompatible with the local noun being the head of a subject noun phrase) relative to local nouns that are either unmarked for case or bear case that could be nominative (i.e., a case marker that is ambiguous between nominative and non-nominative). The logic of this is clear: subject head nouns are typically nominative (either explicitly or covertly marked as such), and a local noun is less likely to become confused with this controller when its morpho-phonology is incompatible with subjecthood. Studies showing this effect include Nicol and Wilson (1999) and Lorimor et al. (2008) for Russian, Hartsuiker et al. (2001) for Dutch, Badecker and Kuminiak (2007) for Slovak, and Nicol and Antón-Méndez (2009) for English.

In the one study conducted on English, Nicol and Antón-Méndez (2009) created English preambles containing as the local noun either a non-casemarked full noun phrase or a case-inflected pronoun. Comparing e.g., *The bill from the accountants...* and *The bill from them...*, they found a significant reduction in the number of agreement errors associated with the case-marked condition. (The rate of agreement errors following the plural pronouns was about 6%, compared with 15% following full noun phrases which were not explicitly case-marked; this reduction in error rate by more than half mirrors that in several of the aforementioned studies).

We note that although case can be manifest somewhat differently in different languages, it always appears internal to the noun phrase[2], a point relevant to our next prediction.

[2]It is encoded synthetically on the English pronouns (Nicol and Antón-Méndez, 2009), as a noun affix in Russian (Nicol and Wilson, 1999; Lorimor et al., 2008), and in Slovak (Badecker and Kuminiak, 2007), and on noun-adjacent determiners

There is a puzzle presented by the prepositional phrase modifier cases that have been extensively studied in English, e.g., *the key to the cabinet(s),* and which robustly produce attraction errors. Although neither the noun nor the determiner in *the cabinets* is casemarked, the presence of the preposition to the left of the determiner has a similar non-subjecthood signaling function: a noun phrase immediately preceded by a preposition is never the subject. Cross-linguistically, there is a close affiliation of case and preposition use. It is something of a puzzle, then, that the presence of case marking dampens agreement errors in a way that the occurrence of a preposition does not. A formal way to resolve this puzzle is to distinguish case marking from prepositions by observing that only the prepositions lie outside the noun phrase, and to conjecture that *only information internal to the noun phrase itself* is capable of acting as a non-subject cue strong enough to significantly reduce agreement errors.

This conjecture is relevant to formal syntactic treatments of the possessive. In the next section we review syntactic analyses of the possessive ending, and note that there is controversy as to whether the ending is part of the possessor phrase structurally (which would put it on a par with case inflection), or a structurally separate phrase-structure head, which would make it more like the noun-phrase-external prepositions which do not dampen agreement errors to the extent that casemarkers do. Exploring errors triggered by possessors offers a way to experimentally distinguish these formal analyses: an overall lower rate of errors with possessors would provide support for the noun phrase-internal view, and an error rate comparable to that with prepositional phrase modifiers would provide support for the noun phrase-external analysis of the possessor ending.

The Phrase Structure of English Possessives

The possessive ending ["s/"] and the possessor phrase that it attaches to have received two distinct types of analysis in formal syntax. The possessor phrase itself occurs as the Specifier of a determiner phrase (DP). On the first type of analysis (Abney, 1987; Zwicky, 1987; Barker, 1991) the possessive ending is analyzed as a phrase-final affix, attached at the right edge of the possessor, as in **Figure 2**. We refer to this analysis as the *noun phrase-internal view* of the possessive ending, since it is both syntactically and morpho-phonologically part of the possessor noun phrase. On this account, the determiner of the overall possessive construction is null. On the second account, the possessive ending is analyzed as a syntactically autonomous determiner, as in **Figure 3**, which then phonologically encliticizes onto the possessor phrase (Abney, 1987; Delsing, 1998; Carnie, 2013). We refer to this as the *noun phrase-external view* of the possessor ending.

What is important for our discussion below about the latter analysis is that the ending is external to the noun phrase syntactically, occurring in a different region of the phrase structure. In this regard it is much like a preposition—occurring adjacent to, but not as a part of, the noun phrase itself. Since

in German (Hartsuiker et al., 2003). In all these cases the morphological expression of case is structurally within the noun phrase.

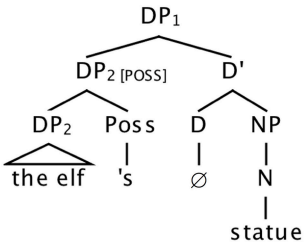


FIGURE 2 | Structure for *the elf's statue*, with possessive ending analyzed as noun phrase-final syntactic clitic.

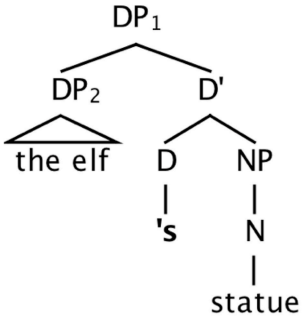


FIGURE 3 | Structure for *the elf's statue*, with possessive ending as a syntactic determiner head, external to the noun phrase/DP.

attractors in prepositional-phrase modifiers robustly elicit errors, the preposition—perhaps due to this structural separation from the noun phrase—apparently does not act as a strong cue for non-subjecthood in the way that noun phrase-internal case marking does in the studies discussed in the previous section showing a dampening effect of case marking.

This pair of contrasting syntactic analyses of the possessor ending leads to differing predictions about the effect that ending will have on agreement errors. If possessors are robust attractors, this will be consistent with the *noun phrase-external* syntactic analysis (**Figure 3**) of the possessor ending, which treats the ending as a syntactically autonomous head, much like a preposition (see **Figure 4**). On the other hand, if possessors are weak attractors, this would be consistent with the analysis of the possessor ending that assimilates it to the class of *noun phrase-internal* case morphology (**Figure 2**).

In our materials, possessors either occur to the left of the head noun, as (9a) (Experiments 1 and 2), or to the right of the head noun (and to the left of the local nouns) when modifying the local noun, as in (9b) (Experiment 3).

(9) a. The elves' statue in the garden
b. The statue in the elves' garden

Before turning to the experiments, we summarize the two sets of predictions we have presented. As noted above, the two possessor positions contrast in proximity to the head noun. Possessors of type (9a) are expected to have a higher error rate due to their greater proximity to the head noun, all other factors being

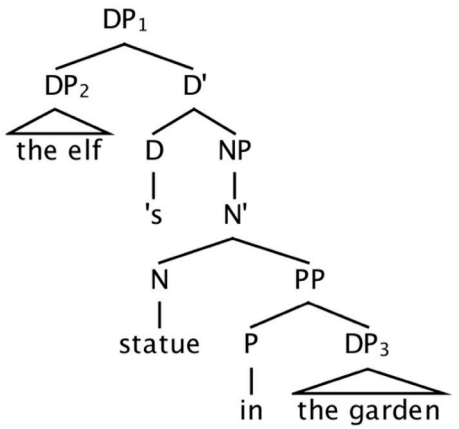


FIGURE 4 | Structure for *the elf's statue in the garden,* showing possessive ending and preposition as DP-adjacent and DP-external heads.

equal, if the relative hierarchical proximity hypothesis (Vigliocco and Nicol, 1998) is the correct account of the locality effects seen in attraction errors. By contrast, the two possessor types should induce equal numbers of errors if the processing window hypothesis (Bock and Cutting, 1992; Nicol, 1995) is correct. And overall error rate for both types of possessors (compared to PP-contained local nouns) will reveal whether the possessor ending functions like noun phrase-internal case information in dampening errors, or a noun phrase-external preposition in not having such an effect.

EXPERIMENT 1—AUDITORILY-PRESENTED PREAMBLES, POSSESSOR MODIFIES HEAD

The purpose of the first experiment was to examine whether a plural possessor phrase modifying the head noun (type 9a) causes interference in the agreement process. For this first experiment, auditory presentation of preambles was chosen, in line with the majority of experiments using the usual paradigm of providing subjects with preambles to turn into complete sentences.

Method

Participants

Forty-four native English-speakers participated in this experiment. Here, and in the studies described below: All were undergraduates at the University of Arizona who received course credit for their participation. They were native English speakers 18 years of age or older. All provided written consent to participate in these experiments, which had received prior approval by the University of Arizona Human Subjects Protection Program.

Materials

Because the stimuli were to be presented auditorily, we needed to ensure that the possessor was unambiguously singular or plural. This meant that we could not use possessors such as *girl's* and *girls'*, because these are homophonous. Therefore, the possessor in our experimental preambles was always a noun with an irregular plural form. The set of possessors included items such as *woman, child, person, housewife, midwife, wolf, thief, elf.*

Twenty quadruplets were such as those in (10) were created. (Here and throughout, preamble types are coded as follows: "s" = singular, "p" = plural, uppercase = head noun. Within the preamble examples, the head noun is underlined and plural nouns are boldfaced).

(10) a. sSs The elf's house with the tiny window...
b. sSp The elf's house with the tiny **windows**...
c. pSs The **elves'** house with the tiny window...
d. pSp The **elves'** house with the tiny **windows**...

Each member of the quadruplet contained a singular head that was preceded by a possessor noun that was singular or plural, and followed by a prepositional phrase modifier containing a singular or plural noun. These were counterbalanced across four presentation lists such that a given participant was presented with only one member of a quadruplet (the full set of experimental items for this and subsequent experiments appear in the Supplementary Material). Each list also included 16 plural-head filler preambles which contained a singular possessor that modified either the head or the (singular or plural) noun within a PP modifier. In addition, there were 64 preambles that were the focus of a separate experiment. These contained a head noun followed by a PP and relative clause modifier; each of the three nouns was singular in half the items and plural in the other half. Finally, there were eight fillers that contained a head noun followed by a PP modifier; each of the two nouns was singular in half of the items and plural in the other half. The preambles were arranged in a fixed pseudorandom order (the same order for each list), and preceded by four practice items. The preambles were recorded by a female speaker.

Procedure

Participants were tested individually in a small test room. Preambles were presented auditorily over headphones. Participants were instructed to repeat each preamble and form a sensible completion. All utterances were recorded for transcription purposes.

Scoring

Transcribed sentences were scored using the following response categories: (a) Correct Inflected (the preamble was repeated correctly and the correct form of an inflected verb was used); (b) Correct Uninflected (the preamble was repeated correctly and an uninflected verb was used); Agreement Error (the preamble was repeated correctly and an incorrectly inflected verb was used); Other Error (the preamble was incorrectly repeated, and/or the verb was missing, or there was no response).

Analyses

Here, and in the following experiments, analyses of variance were performed on the error data, one with subjects (F1) and one with items (F2) as the random variable.

TABLE 1 | Results of Experiment 1 (Auditory Preambles).

Type	Examples	Agreement error (%)	Correct inflect-ED (%)	Correct uninflected (%)	Other errors (%)
sSs	The elf's house with the tiny window	1	67	22	10
sSp	The elf's house with the tiny **windows**	8	49	19	24
pSs	The **elves**' house with the tiny window	2	57	20	21
pSp	The **elves**' house with the tiny **windows**	12	43	17	28

Percentage of responses in each response category for each preamble type. Preamble examples are coded by type (s = singular, p = plural, uppercase = head). Within preamble examples, the head is underlined, and plurals are boldfaced. (Note: due to rounding error, rows sum to 100% only approximately).

In addition, statistical analyses were performed by fitting a linear mixed-effects model to error scores using logit mixed-effects models (Jaeger, 2008). We used the lme4 and lmerTest packages in R (version 3.2.3; CRAN project; R Development Core Team, 2008). Included in each analysis were by-subject and by-item random intercepts, and, if warranted (i.e., if the random intercepts analysis showed a significant effect), also random slopes. The models contained as fixed and random effects the same factors as in the analyses of variance.

We provide the results of the ANOVAs for each experiment, with a brief reference to the results of the mixed-effects modeling with further details of these results offered in the endnotes.

Results and Discussion

The results are shown in **Table 1**.

More errors were associated with plural local nouns than singular ones. Analyses of variance revealed this to be significant. $F1_{(1,\ 43)} = 29.900$, $p < 0.001$; $F2_{(1,\ 19)} = 19.106$, $p < 0.001$. The effect of possessor number was not significant [$F1_{(1,\ 43)} = 2.342$, $p = 0.113$; $F2_{(1,\ 19)} = 2.021$, $p = 0.171$, and the interaction of the two variables was not significant (p's > 0.33]. Mixed effects modeling showed the same pattern: only the main effect of local noun number was significant[3].

A pairwise test of the two conditions containing only one plural [plural possessor (pSs) and plural local noun (sSp)] revealed a significant difference [$F1_{(1,\ 43)} = 9.38$, $p = 0.002$; $F2_{(1,\ 19)} = 9.73$, $p = 0.002$]. The mixed effects analysis also showed a significant difference[4].

Data for the other response conditions were not analyzed statistically; they are displayed in order to show that for the two preamble types in which a single element is plural (pSs vs. sSp), the "opportunity" for an agreement error (derived by summing agreement errors and correctly inflected verbs) is similar, and that the Correct Uninflected and Other Errors are similar in magnitude.

[3]Linear mixed effects modeling showed a main effect of Local Noun Number (Estimate = −2.2065, $SE = 0.4521$, $z = -4.880$, $p = 1.06e\text{-}06$); the effect of Possessor Number was not significant ($p = 0.125$). The Local Noun Number effect was still significant when by-participant and by-item Local Noun Number random slopes were included (Estimate = 0.08809, $SE = 0.02499$, $t = -3.525$, $p = 0.001431$). A comparison of the models (using X2 tests) showed that inclusion of local noun number random slope provided a better fit to the data.

A model that included the interaction of Possessor Number and Local Noun Number showed the interaction to be nonsignificant ($p = 0.771$).

[4]Mixed effects modeling showed a significant effect of Attractor Type (Possessor vs. Local Noun): Estimate = 1.5956, $SE = 0.5652$. $z = 2.823$, $p = 0.00476$. However, with the inclusion of random slopes, the model failed to converge.

Although possessor number had no statistically significant effect on error production, we note that, numerically, more errors were associated with the plural possessor items than the singular possessor items. This difference could become statistically significant with greater power and a more challenging task. This is the motivation for Experiment 2.

EXPERIMENT 2—VISUALLY-PRESENTED PREAMBLES, POSSESSOR MODIFIES HEAD

In this experiment, we used a visual mode of presentation of stimuli in order to increase the overall error rate. Past experiments from our lab have indicated that visual presentation typically results in more errors than auditory presentation. Visual presentation also allowed us to use orthographically distinct cases such as *girl's* vs. *girls'* so that we could increase the number of preambles. Finally, in order to further increase the production of usable data, we included an adjectival ending to promote the use of the copula, which is inflected for number.

Method

Participants

There were 40 participants in this experiment, drawn from the same population as Experiment 1.

Materials and Procedure

Each preamble was paired with an adjective that participants would be asked to use in their sentence completions. We used the 20 quadruplets used in Experiment 1 and created 20 additional quadruplets, for a total of 32. These were counterbalanced across four presentation lists, as described above. Each list also contained 56 filler preambles. Twenty-four of these contained a plural head modified by a singular or plural possessor and by PP containing a singular or plural head. There were also 32 preambles containing a head and PP. Of these, 20 contained a plural head and singular or plural local noun and 12 contained a singular head and singular or plural local noun. Across the set of 88 items, half contained a singular head and half contained a plural head. The preambles were presented in a different random order to each subject, but always preceded by 8 practice trials.

During the experiment, the preamble appeared along with an adjective, as follows: *The elf's house with the tiny window... cute*. Each preamble appeared for approximately 2 s. Participants were asked to silently read each preamble and adjective and then say a

TABLE 2 | Results of Experiment 2 (Visual Preambles).

Type	Examples	Agreement error (%)	Correct inflect-ED (%)	Correct uninflected (%)	Other errors (%)
sSs	The elf's house with the tiny window...cute	4	95	0	2.0
sSp	The elf's house with the tiny **windows**...cute	28	69	0.6	1.0
pSs	The **elves**' house with the tiny window...cute	11	88	0	0.2
pSp	The **elves**' house with the tiny **windows**...cute	35	63	0.3	0.3

Percentages of responses in each category, for each preamble type.

complete sentence out loud. They pressed a foot-pedal to advance to the next item.

Scoring and Analyses

The same response categories and statistical analyses described previously were used here.

Results and Discussion

The results appear in **Table 2**. As the table shows, the error rates were indeed higher than in Experiment 1.

Analyses of variance revealed a significant effect of possessor number [$F1_{(1, 39)} = 10.75$; $p = 0.002$; $F2_{(1, 31)} = 28.63$, $p < 0.001$], and a robust effect of local noun number [$F1_{(1, 39)} = 66.62$, $p < 0.001$; $F2_{(1, 31)} = 88.07$, $p < 0.001$]. The two factors did not significantly interact (p's = 1.0). Mixed effects modeling showed significant main effects, but also a marginal interaction[5].

A comparison of the sSp condition (e.g., *The elf's house with the tiny windows...*) with the pSs condition (e.g., *The elves' house with the tiny window...*) reveals a significant difference between the two, shown by ANOVAs [$F1_{(1, 39)} = 32.94$; $p < 0.001$; $F2_{(1, 31)} = 41.59$; $p < 0.001$] and mixed effect modeling[6].

Data from the other categories were not analyzed statistically. The percentages of correctly inflected verbs complement the Agreement Error results. Very few uninflected verbs were used, and there were very few errors in the Other category.

These results show that, within this more challenging task, plural possessors can induce attraction errors, though significantly fewer than plural local nouns. Further, the presence of a plural possessor and plural local noun appear to have additive effects, resulting in a relatively high rate of errors in the pSp condition.

The next study investigates whether increasing the distance between a potential attractor and the head reduces the potency of the attractor.

EXPERIMENT 3—VISUALLY-PRESENTED PREAMBLES, POSSESSOR MODIFIES LOCAL NOUN

This experiment was conducted in order to explore the effect of a plural possessor when it appeared with the local noun. Just as in the previous experiments, local noun number was also manipulated. Here, it is especially important to show that local nouns induce attraction effects; if the local noun induces errors, it must be co-active with the head, and if it is, then the possessor must also be co-active with the head.

Method

Participants

There were 40 participants in this experiment; again drawn from the same pool as in the previous experiments.

Materials and Procedure

The materials from Experiment 2 were revised to create sensible preambles containing local nouns modified by possessor phrases. Within each quadruplet, the head noun was always singular, the possessor was either singular or plural, and the local noun was either singular or plural [see the examples in (11)]. The filler items were identical to those used in Experiment 2, except that in the 24 fillers containing possessors, the possessor now appeared with the local noun.

(11) a. Sss The statue in the elf's garden ... amusing.
b. Ssp The statue in the elf's **gardens** ... amusing.
c. Sps The statue in the **elves'** garden ... amusing.
d. Spp The statue in the **elves' gardens** ... amusing.

The procedure was identical to that of Experiment 2.

Scoring and Analyses

These were identical to those used in Experiments 1 and 2.

Results

The percentages of agreement errors across conditions appear in **Table 3**.

[5]The mixed effects analysis of the data from Experiment 2 showed a significant effect of Possessor Number (Estimate = 0.5577, $SE = 0.1593$, $z = -3.502$, $p = 0.000462$); with the inclusion of Possessor Number random slope, this was still significant (Estimate = −0.06875, $SE = 0.02200$, $z = -3.126$, $p = 0.0035$. (This model was no better than the original model).

Local Noun Number was also significant (Estimate = −1.9620, $SE = 0.1842$, $z = -10.652$, $p < 2e\text{-}16$); and still significant when the model included Local Noun Number random slope (Estimate = −0.24063, $SE = 0.03573$, $t = -6.734$, $p = 2.31e\text{-}08$). This model was significantly better than the original.

A model that included the interaction of the two variables showed a just significant effect Estimate = −0.7559, $SE = 0.3845$, $z = -1.966$, $p = 0.049328$). Analyses that also included Local Noun Number random slopes showed only a marginal interaction effect (Estimate = −0.7484, $SE = 0.3953$, $z = -1.893$, $p = 0.058295$). The model that included Possessor Number random slopes failed to converge. A comparison of the first two models showed the second to be a better fit.

[6]The pairwise comparison of the plural possessor vs. the plural local noun conditions showed a significant effect (Estimate = 1.2290, $SE = 0.2244$, $z = 5.478$, $p = 4.31e\text{-}08$) that held up when random slopes were included (Estimate = 0.17188, $SE = 0.04011$, $t = 4.285$, $p = 0.000103$). A comparison of the two models reveals the second to be better than the first.

TABLE 3 | Results of Experiment 3 (Visual Preambles).

Type	Examples	Agreement error (%)	Correct Inflect-ED (%)	Correct uninflected (%)	Other errors (%)
Sss	The statue in the elf's garden...amusing	4	92	2	3
Ssp	The statue in the elf's **gardens**...amusing	23	66	2	10
Sps	The statue in the **elves'** garden...amusing	6	89	1	4
Spp	The statue in the **elves' gardens**...amusing	24	64	1	10

Percentages of responses in each category, for each preamble type.

Statistical analyses of the agreement errors revealed a robust effect of local noun number [$F1_{(1, 39)} = 79.77$, $p < 0.001$; $F2_{(1, 31)} = 104.52$, $p < 0.001$], and a non-significant effect of possessor number [$F1_{(1, 39)} = .92$, $p = 0.343$; $F2_{(1, 31)} = 1.05$, $p = 0.307$]. The interaction of the two variables was not significant (p's = 1.0). A comparison of the conditions in which only one element was plural—Sps vs. Ssp—revealed a significant difference [$F1_{(1, 39)} = 43.45$, $p < 0.001$; $F2_{(1, 31)} = 40.76$, $p < 0.001$]. The effects appear not to be additive. Results of mixed effects modeling showed exactly the same effects[7].

In contrast to Experiment 2, the appearance of a plural possessor downstream from the head has virtually no effect on the rate of agreement errors. A comparison of error rates across Experiments 2 and 3 for the conditions in which a plural possessor appeared with a singular head and local noun (pSs vs. Sps) revealed a significant difference by both ANOVA [$F1_{(1, 78)} = 5.6$, $p = 0.02$; $F2_{(1, 62)} = 6.02$, 0.017], and mixed-effect analyses[8].

Discussion

Overall, our results have shown the following: (a) The closer the possessor attractor to the head noun, the greater the likelihood of verb agreement errors. This cannot be reducible to a processing window effect because the local noun attractor in downstream position does produce errors (showing that it is within the same processing window as the head), providing support for the relative proximity hypothesis. (b) Plural possessors in general induce few errors, suggesting that the cue to non-headedness provided by the possessive ending is robust in the same way overt case-marking is, and quite distinct from the cue that is specified by a preposition, lending support for the noun phrase-internal syntactic analysis of the possessive ending.

This latter result is consistent with findings for case-marking languages like Russian (Nicol and Wilson, 1999; Lorimor et al., 2008), which show low rates of error. But note that some of the research on case-marking languages has shown that phonological distinctiveness also plays a role. For example, Hartsuiker et al. (2003) found that an attractor with unambiguous non-nominative case marking induced fewer errors than a case-ambiguous attractor. We observe that the two variants of the possessive ending, ['s] and ['], differ in salience (both phonological and orthographic), and question whether salience plays a role in the effectiveness with which the possessor ending dampens errors.

In order to assess whether this kind of form-related distinctiveness played a role in our studies, we conducted a *post-hoc* analysis of the data from Experiment 2, the only experiment in which possessor number had a significant effect. We divided the items into two groups: plural possessors which marked the possessive with the morpheme –s (e.g., *policewomen's, children's, councilmen's*, etc...) vs. those which marked the possessive only with an apostrophe (e.g., *companies', families', elves'*). The former set of materials contained 15 items; the latter set 17 items. The mean percentages of agreement errors are displayed in **Table 4**.

As **Table 4** shows, there were more errors when case-marking was less salient (orthographically and phonologically).

ANOVAs showed a main effect of case-marking type [$F1_{(1, 39)} = 12.75$, $p = 0.001$; $F2_{(1, 30)} = 6.73$, $p = 0.015$]. Type of case-marking did not interact with possessor number. Linear mixed effects analyses showed the same effects[9].

Overall, then, we have seen that both structural and morphophonological variables affect the rate of agreement errors. But we can flesh out the picture even further by investigating semantic effects.

Research by Pearlmutter and his colleagues (e.g., Solomon and Pearlmutter, 2004; and by Brehm and Bock, 2013) has shown that the extent to which a head and local noun are integrated—in a semantic sense—affects whether the ensuing verb is singular or plural. For example, the component elements *drawing* and *flowers* are more tightly integrated in *the drawing of the flowers* than in *the drawing with the flowers*. Interestingly, although Solomon and Pearlmutter (2004) found more agreement errors associated with preambles of the former type (the "of" preambles) than the latter, Brehm and Bock (2013) found the opposite. Brehm and Bock posit that highly integrated

[7]Analysis of Experiment 3 data showed a significant effect of only Local Noun Number (Estimate = −1.9738, $SE = 0.2152$, $z = -9.172$, $p = <$ 2e-16). Inclusion of Local Noun Number random slopes: Estimate = −0.18750, $SE = 0.02765$, $t = -6.780$, $p =$ 5.43e-08). Chi-square analysis shows the latter to better fit the data than the former.

The effect of Possessor Number was not significant ($p = 0.261$).

Mixed effects modeling that included the interaction of the two variables was not significant ($p = 0.430$).

The comparison of the sPs and ssP conditions showed a significant effect (Estimate = 1.7694, $SE = 0.2994$, $z = 5.911$, $p =$ 3.41e-09) that was maintained with the addition of random slopes (Estimate = 0.17188, $SE = 0.04011$, $t = 4.285$, $p = 0.000103$). Comparison of the two models indicated the latter to be superior.

[8]The effect of the plural possessor in different positions within the complex noun phrase subject was analyzed. Linear mixed effect modeling revealed a significant effect (Estimate = −0.7751, $SE = 0.3630$, $z = -2.135$, $p = 0.0327$) that remained significant with the inclusion of random slopes (Estimate = −0.05304, $SE = 0.02526$, $t = -2.100$, $p = 0.0398$). A chi-square test showed the latter model to better fit the data.

[9]Mixed modeling analysis of Case-marking Type: Estimate = 0.6197, $SE = 0.2331$, $z = 2.659$, $p = 0.00784$. With Case-marking Type random slopes: Estimate = 0.07588, $SE = 0.03055$, $t = 2.484$, $p = 0.018964$).

TABLE 4 | Experiment 2 data, grouped by type of case-marking.

Case-marker on plural	Type	Examples	Agreement error (%)
	sSs	The woman's position on the issue	2
's	sSp	The woman's position on the **issues**	23
	pSs	The **women's** position on the issue	8
(*N* = 15)	pSp	The **women's** position on the **issues**	28
	sSs	The country's response to the attack	5
'	sSp	The country's response to the attacks	33
	pSs	The **countries'** response to the attack	15
(*N* = 17)	pSp	The **countries'** response to the **attacks**	43

Shown here are percentages of responses in each category, for each preamble type and case-marking type.

TABLE 5 | Analysis of Semantic Differences associated with the Possessor (data from Experiment 2).

	Type	Agreement error (%)
Possessive = "Of" (*N* = 17)	sSs	3
	sSp	24
	pSs	8
	pSp	29
Possessive = Other (*N* = 15)	sSs	6
	sSp	34
	pSs	13
	pSp	43

phrases such as *the drawing of the flowers* are simply more likely to be treated as a unitary conceptual object (at what is called the "message level" representation, the conceptual representation that feeds the production system). If such phrases are construed as singular, they will be treated as the unmarked singular in the linguistic representation. In contrast, *the drawing with the flowers* is more likely to be treated at the message level as referring to several objects, and thus would be more likely to be marked linguistically with a plural feature.

In phrases containing PP modifiers, the relationship between the head and local noun is signaled by the preposition. But with possessors, the relationship must be computed based on real-world knowledge. Possessors can serve sometimes as arguments to the head (e.g., bearing the agent role in *the salesman's promise* to *the customers*) but need not. Possessors have a very broad and essentially unlimited range of possible connections to the head noun[10]: *the elf's house* can be the house owned by the elf, occupied by the elf, designed by the elf, in which the elf is kept as a prisoner, where the elf bakes cookies, defended by the elf as a matter of duty, etc...

Do speakers compute these various relationships? To address this question, we divided our materials based on which preposition would be used if the possessor-head relationship were recast as a head-PP relationship, choosing the most appropriate preposition in each case. For example, the *women's position* would be recast as *the position of the women* and the *spokeswomen's announcement* would be recast as *the announcement by the spokeswomen*. The semantic integration/referential subordination notion aligns with the preposition choice in our recasting of our materials. In the cases with high referential subordination, the preposition in the converted materials is *of*, unique among prepositions in having no lexical-semantic meaning (it is, for example, the default preposition used with objects of deverbal nouns: *announce the award, announcement of the award*, where the complement of the verb has no accompanying preposition, and the same thematic role between verb or noun and object is understood). The less integrated, less referentially subordinate possessors tend to be converted with prepositions with lexical meaning: *from, by*, and *to*.

We grouped the "of" versions together (seventeen items), and the other conditions together (fifteen items). Results appear in **Table 5**.

Analyses of variance show a main effect of Encoded-Preposition Type (*of* vs. other): $F1_{(1,\ 39)} = 12.71$, $p = 0.001$, $F2_{(1,\ 30)} = 5.84$, $p = 0.022$). This variable did not interact significantly with the other variables (which is similar to the Brehm and Bock, 2013, findings). Results of linear mixed effects modeling were similar[11].

We found significantly fewer errors associated with the "of" versions, in line with Brehm and Bock's findings (2013).

GENERAL DISCUSSION

Our findings can be summarized as follows.

First, a possessor attractor that is closer to the head induces more agreement errors than one that is more distant from the head, even when the more distant attractor is co-active with the head[12]. The difference between the two possessor positions shows that relative structural proximity to the head noun is a key factor in determining the magnitude of errors, supporting the view of Vigliocco and Nicol (1998), and arguing against a processing window analysis as an alternative to a proximity account (Nicol, 1995).

[10]Some nouns have an inherent argument structure, including relational nouns like *sister*, *friend*, and *mother*, and deverbal nouns like *teacher, author*. With such nouns the dominant reading of the Possessor is that of one of the arguments of the noun, though other readings are available—*the teacher's mother* could be a mother assigned as helper to the teacher, for example. For nouns with no argument structure, the possible semantic connections between the Possessor and noun is unlimited. Partee and Borschev (2003) present the example *John's team*, and observe that it may be the team John owns, founded, works for, is a teammate on, covers as a reporter, is a fan of, runs in a fantasy league, etc. See Partee and Borschev (2003) and Barker (1991) for much discussion.

[11]Possessor semantics (preposition type): Linear mixed effects modeling showed a significant effect of possessor semantics (*of* vs. other): Estimate = 0.6113, *SE* = 0.2338, $z = 2.615$, $p = 0.008925$.
A model that included the interaction of the variables showed a just-significant interaction of possessor number and possessor semantics: Estimate = 0.618546, *SE* =0.309129, $z = 2.001$, $p = 0.0454$. No other interactions were significant.

[12]As shown by the robustness of errors with the local nouns that occur downstream from the possessors, showing them both to be within the same activation window as the subject head noun.

Second, compared to a local noun attractor within a modifying PP, plural possessors are much less robust as attractors. Averaging across Experiments 2 and 3, and using the all-singular condition as a baseline, net rates of attraction (subtracting out errors associated with the all-singular condition) were roughly 21.5% for plural local nouns and 4.5% for plural possessors. We have suggested that one reason possessors induce fewer errors is that they carry *within their form* information that they are nonheads. In contrast, when a local noun is the object of a preposition, information about its nonhead status derives from information that is not inherent in its noun phrase: its position within the complex subject noun phrase structure, and the fact that it is the object of a preposition. We conclude that the possessor ending is an element of the noun phrase itself, on a par with case markers, thus supporting the syntactic analysis shown in **Figure 2**, and arguing against the noun phrase-external analysis of the ending as a separate determiner head (**Figure 3**).

The fact that form information matters is supported by our first *post-hoc* analysis that showed that the more salient the orthographic/phonological cues about the possessor's role, the fewer errors there were, with possessors with more salient marking such as *women's* causing fewer errors than those with less salient marking like *countries'*.

It is interesting to note that the attraction effect elicited by a possessor as a satellite of the local noun induces fewer errors than that caused by a pronoun in local noun position. Recall the study by Nicol and Antón-Méndez (2009). They showed that a case-marked pronoun in local noun position elicited 6.5% verb agreement errors (5.8% if singular-pronoun errors are subtracted, the *net effect*). This is still substantially larger than the 2% net effect observed in Experiment 3. Obviously, cross-experiment comparison must be interpreted with caution. However, possessors and pronouns are both case-marked, and in the relevant experiments, both intervened between and head and the verb and are roughly the same distance from the root node. In addition, the contrast between singular vs. plural and nominative vs. accusative forms (e.g., *he/him* vs. *they/them*) is more salient than the contrasts in the experiments here. If salience reduces errors, it is even more surprising that pronouns are relatively more powerful attractors. We conjectured that this may be tied to the message level representations of pronouns vs. possessives, specifically with respect to the degree of semantic integration involved.

Our results suggest that the semantic integration between the head and its possessor also matters: when the possessor merely possesses (as in *the elves' house*), fewer errors result than when the possessor is a creator or recipient (e.g., *the congressmen's telegram*). One way in which integration can be understood is that in cases of high integration one entity (the referent of the head noun) is referentially dominant and foregrounded, with the other(s) subordinate to it; this will encourage a singular construal of cases like *the drawing of the flowers*. This referential subordination is reflected in one's intuitions about whether both entities are called to mind with more equal foregrounding. In the case of our possessives, *the elves' house,* plausibly gives rise to a house-dominant conceptual representation, while *the congressmen's telegram* could elicit a representation in which congressmen and telegram are both highlighted (perhaps reflecting the fact that the specifics of a telegram are dependent upon the type of author).

This set of results is consistent with the dominant theory of how verb agreement is computed during language production: the Marking and Morphing model proposed by Bock and colleagues (e.g., Bock et al., 2001; Eberhard et al., 2005). This model assumes a multi-staged architecture in which processing proceeds from top to bottom. First, a non-linguistic proposition (the *message*) leads to the selection of abstract (non-phonological) lexical representations that correspond to concepts within the message, and simultaneously to the computation of a predicate-argument structure. Within the message representation, the roles of the participants are identified, and this information is transmitted to, and coded within, the predicate-argument structure. This includes information about whether, for example, the subject as a whole is singular or plural, and whether the elements that comprise the subject (like modifier-contained noun phrases) are singular or plural. In addition, components of the predicate-argument structure are linked to the abstract lexical representations such that a given lexical item may be assigned to a theme/object role, etc.... At a second stage, a phrasal structure is computed; this structure inherits grammatical number features from the predicate-argument structure. (Other grammatical features are inherited as well, including definiteness, verb tense, and so forth). Verb number is specified via a copying operation that copies number marking from the subject phrase to the verb. Ultimately, form information associated with the selected lexical items is retrieved and slotted into position within the phrasal structure, and inflectional and other grammatical elements are also phonologically realized.

There are two ways for an agreement error to arise. One is during the *marking* process, in which a subject phrase is marked as singular or plural based on its conceptual representation within the message (see also, Vigliocco and Franck's, 1999 *Maximal Input Hypothesis*). Semantic integration of a complex subject exerts its influence here. Following our discussion above, a conceptual level representation corresponding to *The elves' house* will likely be determined to be singular (referring to a singular entity), and marked as such. By comparison, *The congressmen's telegram* will slightly more often receive plural marking, if the message-level representation highlights both congressmen and a telegram.

The other way an error arises within this model is during the later *morphing* process. Morphing involves a set of operations that include connecting lexical information to positions within a syntactic frame that is annotated for number (and other grammatical features), and copying the number feature from the subject noun phrase to the verb (or inflectional node). Part of this process also includes the possibility of percolation of the number feature from the head noun to the root node of the subject phrase. Percolation is a way for the number specification of a head noun to modify the number specification of the subject phrase at the root node (this is described as a "reconciliation" process). (This mechanism is necessary to accommodate cases in which notional number and grammatical number diverge, such

as *scissors*, a singular entity with plural marking. If *scissors* is the head, the plural feature percolates to the highest node, effectively turning the subject phrase plural, and triggering plural agreement with the verb). Occasionally, a plural feature from the wrong noun can percolate to the subject's root node, leading to a verb agreement error. The more deeply embedded the attractor, the less likely it is that percolation of a feature would be able to overwrite the phrasal feature. Our results are consistent with this: the greater the distance between a plural possessor and head noun, the smaller its impact.

Morphophonological effects to do with case marking also come about during the morphing process. Bock and Middleton (2011) describe the effect of case ambiguity as follows: "A plausible consequence of this ambiguity is a sparse or unstable feature set when such nouns serve as agreement controllers...this would induce competition between the (intended) nominative and (uninvited but consorting) accusative. In turn, competition increases the likelihood of attraction, which arises when the morphological specifications of an attractor occupy the feature set of the controller." (p. 1052). In our preambles, the head noun was always case-ambiguous, and therefore subject to competition from the other two nouns, the local noun and the possessor. The local noun was also case-ambiguous, offering greater competition with the head than the case-marked possessor. But in order for case-marking to be useful, it needs to be noticed; our *post-hoc* analysis show that within the set of case-marked possessors, more salient phonological/orthographic case-marking was associated with fewer errors.

CONCLUSION

The present results extend the empirical domain of studies of the production of verb agreement by examining possessors, previously unstudied. We have experimentally investigated the magnitude of errors induced by possessors in two positions differing in structural proximity to the head noun, both in comparison to one another and to local nouns in the canonical position investigated in much previous research. We have shown that the higher possessor produces errors at a greater magnitude than the lower, and that both types induce fewer errors than a local noun. These results show that proximity to the head noun matters, and further that some property of possessors significantly dampens errors with this type of phrase, a property we have identified as case marking.

These results bear on three theoretical issues in the account of agreement production. The first is the nature of the proximity effect, where we have argued from the asymmetry between head-modifying and local-noun modifying possessors that relative structural proximity to the head noun plays a key role. The second issue is the role that the possessor ending has in modulating errors, where we have argued it plays a role akin to case marking in richly inflected languages, thus showing English speakers attend to case in spite of the relative lack of case in that language. We have also noted that the salience of the two variants of this ending affects the magnitude of errors, as does the semantic integration of the possessor with the noun it modifies. Finally, we have argued that the psycholinguistic results bear upon the formal syntactic analysis of the possessor ending.

AUTHOR CONTRIBUTIONS

JN conceived of study, created materials, oversaw lab in which experiments were run, conducted statistical analyses and *post-hoc* analyses, cowrote paper. AB provided linguistic expertise and cowrote paper. JB set up experiments to run, supervised testing of subjects and coding of data, did preliminary statistical analyses.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpsyg.2016.00548

REFERENCES

Abney, S. (1987). *The English Noun Phrase in Its Sentential Aspect*. Doctoral Dissertation, Cambridge, MA: MIT.

Badecker, W., and Kuminiak, F. (2007). Morphology, agreement and working memory retrieval in sentence production: evidence from gender and case in Slovak. *J. Mem. Lang.* 56, 65–85. doi: 10.1016/j.jml.2006.08.004

Barker, C. (1991). *Possessive Descriptions*. Chicago, IL: CSLI Publications, distributed by University of Chicago Press.

Bock, K., and Cutting, J. (1992). Regulating mental energy: Performance units in language production. *J. Mem. Lang.* 31, 99–127. doi: 10.1016/0749-596X(92)90007-K

Bock, K., Eberhard, K. M., Cutting, J. C., Meyer, A. S., and Schriefers, H. (2001). Some attractions of verb agreement. *Cogn. Psychol.* 43, 83–128. doi: 10.1006/cogp.2001.0753

Bock, K., and Middleton, E. L. (2011). Reaching agreement. *Nat. Lang. Linguist. Theory* 29, 1033–1069. doi: 10.1007/s11049-011-9148-y

Bock, K., and Miller, C. (1991). Broken agreement. *Cogn. Psychol.* 23, 45–93. doi: 10.1016/0010-0285(91)90003-7

Brehm, L., and Bock, K. (2013). What counts in grammatical number agreement? *Cognition* 128, 149–169. doi: 10.1016/j.cognition.2013.03.009

Carnie, A. (2013). *Syntax: A Generative Introduction, 3rd Edn*. Malden, MA: Wiley-Blackwell.

Delsing, L.-O. (1998). "Possession in Germanic," in *Possessors, Predicates, and Movement in the Determiner Phrase,* eds by A. Alexiadou and C. Wilder (Amsterdam: Johns Benjamin Publishing), 87–108.

Eberhard, K. M., Cutting, J. C., and Bock, K. (2005). Making syntax of sense: number agreement in sentence production. *Psychol. Rev.* 112, 531–559. doi: 10.1037/0033-295X.112.3.531

Franck, J., Vigliocco, G., and Nicol, J. (2002). Subject–verb agreement errors in French and English: the role of syntactic hierarchy. *Lang. Cogn. Process.* 17, 371–404. doi: 10.1080/01690960143000254

Gillespie, M., and Pearlmutter, N. J. (2011). Hierarchy and scope of planning in subject–verb agreement production. *Cognition* 118, 377–397. doi: 10.1016/j.cognition.2010.10.008

Grimshaw, J. (1990). *Argument Structure*. Cambridge, MA: MIT Press.

Hartsuiker, R. J., Antón-Méndez, I., and van Zee, M. (2001). Object attraction in subject-verb agreement construction. *J. Mem. Lang.* 45, 546–572. doi: 10.1006/jmla.2000.2787

Hartsuiker, R. J., Schriefers, H. J., Bock, K., and Kikstra, G. M. (2003). Morphophonological influences on the construction of subject–verb agreement. *Mem. Cogn.* 31, 1316–1326. doi: 10.3758/BF03195814

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *J. Mem. Lang.* 59, 434–446. doi: 10.1016/j.jml.2007.11.007

Jespersen, O. (1913/1961). *A Modern English Grammar on Historical Principles*. Copenhagen: Munksgaard/London: George Allen & Unwin.

Lorimor, H., Bock, K., Zalkind, K., Sheyman, A., and Beard, R. (2008). Agreement and attraction in Russian. *Lang. Cogn. Process.* 23, 769–799. doi: 10.1080/01690960701774182

Nicol, J., and Antón-Méndez, I. (2009). "The effect of case marking on subject–verb agreement errors in English," in *Time and Again: Theoretical Perspectives on Formal Linguistics in Honor of D. Terence Langendoen,* eds W. D. Lewis, S. Karimi, H. Harley, and S. O. Farrar (Amsterdam: John Benjamins Publishing), 135–150.

Nicol, J. L. (1995). Effects of clausal structure on subject-verb agreement errors. *J. Psychol. Res.* 24, 507–516. doi: 10.1007/BF02143164

Nicol, J., and Wilson, R. (1999). "Agreement and case-marking in Russian: a psycholinguistic investigation of agreement errors in production," in *The Eight Annual Workshop on Formal Approaches to Slavic Languages. The Philadelphia Meeting* (Ann Arbor, MI: Michigan Slavic Publications), 314–327.

Partee, B. H., and Borschev, V. (2003). "Genitives, relational nouns, and argument-modifier ambiguity," in *Modifying Adjuncts,* eds E. Lang, C. Maienborn, and C. Fabricius-Hansen (Berlin: Walter de Gruyter), 67–112.

Quirk, R., Greenbaum, S., Leech, G., and Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. London: Longman.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available online at: http://www.R-project.org

Solomon, E. S., and Pearlmutter N. J. (2004). Semantic integration and syntactic planning in language production. *Cogn. Psychol.* 49, 1–46. doi: 10.1016/j.cogpsych.2003.10.001

Vigliocco, G., and Franck, J. (1999). When sex and syntax go hand in hand: gender agreement and language production. *J. Mem. Lang.* 40, 455–478. doi: 10.1006/jmla.1998.2624

Vigliocco, G., and Nicol, J. (1998). Separating hierarchical relations and word order in language production: is proximity concord syntactic or linear? *Cognition* 68, B13–B29. doi: 10.1016/S0010-0277(98)00041-9

Visser, F. T. (1963). *An Historical Syntax of the English Language*. Leiden: E. J. Brill.

Zwicky, A. (1987). Suppressing the Zs. *J. Linguist.* 23, 133–148. doi: 10.1017/S0022226700011063

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2016 Nicol, Barss and Barker. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

HYPOTHESIS AND THEORY
published: 16 February 2016
doi: 10.3389/fpsyg.2016.00120

Language Processing as Cue Integration: Grounding the Psychology of Language in Perception and Neurophysiology

Andrea E. Martin *

Department of Psychology, School of Philosophy, Psychology and Language Sciences, University of Edinburgh, Edinburgh, UK

I argue that cue integration, a psychophysiological mechanism from vision and multisensory perception, offers a computational linking hypothesis between psycholinguistic theory and neurobiological models of language. I propose that this mechanism, which incorporates probabilistic estimates of a cue's reliability, might function in language processing from the perception of a phoneme to the comprehension of a phrase structure. I briefly consider the implications of the cue integration hypothesis for an integrated theory of language that includes acquisition, production, dialogue and bilingualism, while grounding the hypothesis in canonical neural computation.

Keywords: language comprehension, sentence processing, cue-based retrieval, cue integration, neurobiology of language

OPEN ACCESS

Edited by:
Matthew Wagers, University of California, Santa Cruz, USA

Reviewed by:
John E. Drury, Stony Brook University, USA
Darren Tanner, University of Illinois at Urbana-Champaign, USA

***Correspondence:**
Andrea E. Martin
andrea.martin@ed.ac.uk

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: *20 August 2015*
Accepted: *22 January 2016*
Published: *16 February 2016*

Citation:
Martin AE (2016) Language Processing as Cue Integration: Grounding the Psychology of Language in Perception and Neurophysiology. Front. Psychol. 7:120. doi: 10.3389/fpsyg.2016.00120

INTRODUCTION

Despite major advances in the last decades of language research, the linking hypothesis between ever-more plausible neurobiological models of language and ever-better empirically supported psycholinguistic models is weak, if not absent. Moreover, we are struggling to answer, and even to ask well, questions like *why is language behavior the way it is? How is language processed? What is "processing difficulty?" What is the source of difficulty in psychological and neurobiological terms? What can it tell us about the computational architecture of the language system?* These questions, however frustratingly difficult, speak to our persistent awe at the fact that we humans flap our articulators, we move the air, and in doing so, stimulate formally-describable complex meaning in the heads of other people. And then those people usually do it to us back. So how do we, or rather, our brains, do it?

There must be a good reason for the weak link between psycho- and neurobiological theories of language—namely that it is really hard to find a concept that would be explanatory on multiple levels of analysis in cognitive science (see Marr, 1982). Questions like *what makes language the way it is* probe the computational level of Marr's tri-level hypothesis, asking what the system's goal is, what computation is being performed and to what end. Questions like *how does the system do it* occur at the algorithmic level, asking what the nature of the mechanism that carries out the computation is. Recent debates in cognitive science have cast these two kinds of questions in opposition, or at least, in opposing theoretical camps. Bayesian modelers of perception and cognition form the statistical *what* camp, and non-Bayesians the mechanistic *how* camp (Jones and Love, 2011; Bowers and Davis, 2012). The *what* camp is purportedly less interested in how the mind "does it," but is focused on reverse engineering how

the natural world (or the statistics that describe it) makes cognition the way it is. The *how* camp purportedly wants to uncover the mechanism that the mind/brain uses, instead of a statistical approximation (Jones and Love, 2011; Bowers and Davis, 2012). I will argue that any model of language computation must answer both *how* and *what* questions, and the best model will most likely include both mechanistic and probabilistic elements. The model articulated here asserts a mechanistic psychological operation over representations derived via Bayesian inference (or an approximation there of), which are represented by neural population codes that are flexibly combined using two simple canonical neural computations: summation and normalization.

Rather than trying derive novel psychological mechanisms specific to language, I will ask whether insights from perception and psychophysiology can inform process models of psycholinguistic theory to try to explain why language behavior is the way it is *and* how formal linguistic representations might be extracted from sensory input and represented by the brain. First, I will briefly consider two recent advances in psycholinguistic theory, the Cue-based Retrieval framework (CBR) and Expectation-based parsing (EBP), which have shaped the field in the last decade. Then I will briefly explore the implications of sensory processing models in order to argue that the main insights of these frameworks can be transferred to psycholinguistics as a single mechanism derived from neurobiological principles. Then I will attempt to apply this principle to sentence comprehension, and briefly explore its implications for production, dialogue, language acquisition, and bilingualism. Finally, I will try to deliver predictions that could falsify this approach.

Two Influential Theories: Cue-Based Retrieval and Expectation-Based Parsing

The cue-based retrieval framework offers an account for processing difficulty in language comprehension that is based on architectures and mechanisms from human memory, specifically recognition memory (McElree, 2000; McElree et al., 2003; Lewis et al., 2006). It originates from the classic insight that retrieval from memory might be needed to form grammatical interpretations, especially for syntactic structures where words that form a linguistic dependency are separated from each other by other words (Miller and Chomsky, 1963). Quite naturally then, CBR has focused on non-adjacent dependencies of different kinds, mostly subject-verb dependencies (McElree, 2000; McElree et al., 2003; Lewis et al., 2006; Van Dyke, 2007; Wagers et al., 2009; Van Dyke and McElree, 2011; Tanner et al., 2014) but also pronouns, ellipsis and other situations with referential or anaphoric consequences (Foraker and McElree, 2007; Martin and McElree, 2008, 2009, 2011; Xiang et al., 2009; Martin et al., 2012, 2014; Dillon et al., 2013; Jäger et al., 2015).

The appeal of the cue-based framework is the parsimony of explanation—language behavior is the way it is because of the architecture of human memory. Memory is *content-addressable*[1], or organized by content, and therefore is highly susceptible to interference (see McElree, 2000; McElree et al., 2003; Lewis and Vasishth, 2005; Lewis et al., 2006; and see McElree, 2006; Van Dyke and Johns, 2012, for reviews). Interference occurs when the link between the cues used at retrieval and the intended target representation is not diagnostic (McElree, 2000, 2006; McElree et al., 2003; Martin et al., 2012). Therefore, according to CBR, processing difficulty in language comprehension is due to interference[2], or more specifically, *cue overload*, the term for the situation when the cues at retrieval are insufficient to elicit the needed representation (McElree, 2006; Van Dyke and McElree, 2011; Van Dyke and Johns, 2012). Whether cue overload arises purely due to similarity between representations and cues, or whether distinctive items in memory are somehow disruptive during retrieval, is an on-going challenging question (see Jäger et al., 2015, for an overview on effect reversals for pronouns). Another important architectural assumption of CBR is that retrieval speed is constant, so effects on performance (either accuracy or reaction time) arise from differences in representation, namely cue-target match vs. the match of the cue to other items in memory (McElree, 2006; see Nairne, 2002 for more on diagnostic cues). Additionally, representations appear to be retrieved without a serial or parallel search (see Townsend and Ashby, 1983; McElree and Dosher, 1989, 1993; Martin and McElree, 2009, for details on how parallel search is falsified). CBR has been well-implemented: Lewis and Vasishth (2005) and Lewis et al. (2006) describe compelling symbolic models of parsing implemented with only one additional parameter than the standard ACT-R model (Anderson, 1983).

Expectation-based parsing has focused on modeling classic sentence processing phenomena (syntactic ambiguity resolution and relative clause processing asymmetries) in a Bayesian framework (Hale, 2001; Levy, 2008, 2013; Smith and Levy, 2013). The approach aim to predict which parts of a sentence will be more difficult to process as reflected in behavioral measures. It marks a renaissance for the role of expectation and its formalization in psycholinguistic theory (cf. MacDonald et al., 1994; Altmann and Kamide, 1999; DeLong et al., 2005; Van Berkum et al., 2005). In EBP, parsing decisions are based on probabilities built up from prior experience, and difficulty stems from the violation of word-by-word expectations of syntactic structure. In other words, the main claim is that surprisal, or the degree to which expectations are not met, is the best predictor of reading time slow down and therefore, of processing difficulty (Hale, 2001; Levy, 2008). This striking insight has a lot in common with ideal observer models of perception, which I will review

[1] In contrast to location-addressable systems, where data is stored irrespective of its content and a search must be executed to retrieve a particular target item.

[2] Cue-based retrieval interference, although some psycholinguistic work invokes the notion of *encoding* interference, whereby representations fail to be stably encoded when there are multiple similar items in memory (Hofmeister and Vasishth, 2014). However, the spirit of that notion is usually cached out as *proactiveinterference* in the recognition memory literature, whereby forgetting occurs due to information learned or encoded prior to the onset of the study item, but is still due to cue-overload (Anderson and Neely, 1996; Öztekin and McElree, 2007; Martin and McElree, 2009; Van Dyke and McElree, 2011; Van Dyke and Johns, 2012). Retroactive interference refers to forgetting due to information acquired after the onset of the study item (Gillund and Shiffrin, 1984; Anderson and Neely, 1996; McElree, 2006).

shortly, by virtue of the fact that both are rational and formalized with Bayes' rule. EBP continues the tradition of frequentist accounts of parsing (e.g., MacDonald et al., 1994) and statistical learning in psycholinguistics (e.g., Charniak, 1996; Saffran et al., 1996; Tabor et al., 1997; MacDonald and Christiansen, 2002). EBP's advantage over previous statistical learning accounts might be that it is formalized with a probabilistic grammar and can be highly predictive of which parse or *where* in a structure difficulty will be encountered (Levy, 2008).

Challenges for CBR and EBP

Each of these approaches is motivated by the central challenge of parsing: incorporating incoming, new information (phonemes, syllables, morphemes, or lexical items) into a continuously unfolding complex representation. Each approach brings an important insight from a related areas of cognitive science to bear on language processing: (1) for CBR, the parsimony of ACT-R principles and the explanatory concepts of cues and interference, and (2) for EBP, the vital importance of prior experience and expectations, and of formalizing uncertainty. Despite these important insights, the architectural claims that each approach makes are not fully articulated. CBR and EBP might tell us about how an aspect of language processing is carried out, but many questions remain about the nature of the representations and mechanistic processes that are at stake.

The beauty of CBR is that its principles are independently motivated by the architecture of human memory. But despite this, many issues still need to be resolved. First, the psychological mechanism that the additional CBR parameter might correspond to would need to be hypothesized about and tested. Larger architectural questions persist, such as whether retrieval is identical during lexical access and dependency resolution, and whether additional mechanisms besides retrieval might be needed for a fully specified model of parsing. More fundamentally, if grounding language processing in memory processes is what gives explanatory power, then difficult issues about memory processes, such as whether encoding and retrieval ever really separate, need also be addressed. Similarly, complex questions about cues remain: why some representations function as cues and other not, how cues are learned and represented, and how their weights are determined, and whether those weights are determined dynamically all need to be established. The *how* questions might be clearer in CBR, but the answer to *what* questions is offloaded onto memory research.

Similarly, though Levy and colleagues have exacting predictions as to where in a sentence reading slow down will occur, EBP's explanation for "processing difficulty" is not psychological or mechanistic in nature. It is computationally descriptive: re-ranking of probability distributions regarding expected input. Re-ranking of probability distributions actually has a neurophysiological appeal, but is not yet a psychological concept. Since EBP focuses on capturing extant behavioral data patterns and predicting patterns of reading slow down, rather than deriving representational states and processing mechanisms that are both neurobiologically and psychologically plausible, it is not clear how EBP would answer *how* questions. Simply put, EBP is not a process model. Architectural questions about representation also persist, especially as to which representations are being counted and why, and how are probabilistic estimates of being in a parse given the input are formed. The origin of these representations is also unclear, as is the mechanism that is acquiring the statistics and the mechanism that is re-ranking the distributions. If the claim of EBP is that ranking of probabilistic representations what parsing difficulty is, it begs questions as to how the system parses sequences that it has never encountered before, or how it can parse something that is highly unexpected at all, and moreover, what parsing is qua mechanism. If experience is the basis of obtaining probabilistic estimates of a given structural configuration, then it is unclear how parsing might occur without lots of or sufficient experience. Furthermore, how the system acquires experience about parsing, if experience is what is used to generate representations of the parse and probabilistic estimate regarding it, might lead to a circular explanation.

For these reasons, I see the core principles of interference and representing uncertainty as being valuable terms in a larger mechanistic process model, which, hopefully, can also be grounded in neurophysiological computation. By synthesizing mechanistic and Bayesian approaches, we can pose questions both about how language processing functions and why it is that way. But that does not mean that mapping hypotheses about representations and processes onto hypotheses about their priors is will be straightforward.

Ideal Observer Models in Perception

Ideal observer models have dominated research on perception because they lay bare the computational and statistical structure of the complex problems that the brain solves. They force the researcher to define the information available to the brain, and to construct a quantitative, predictive account of performance (Gibson, 1966; Marr, 1982). The *ideal observer* formally describes human behavior in terms of optimal performance on a given problem or task given uncertainty stemming from the environment or sensory system (Trommershauser et al., 2011). The main source of uncertainty in ideal observer models of visual perception is the probabilistic relationship between a given cue (e.g., contrast, color, shading) and a stimulus (e.g., an edge or object) in the environment. In other words, uncertainty stems from the probability of detection of the stimulus in the face of sensory or neuronal noise (Fetsch et al., 2013). Past experience weights the likelihood function of a cue. Thus Bayesian models that incorporate the right combination of cues and priors have become the best predictors of performance on motor control and visual or multisensory perception tasks (Griffiths et al., 2012; Ma, 2012), although some argue that they need not be Bayesian nor rational to achieve this (Maloney and Zhang, 2010; Rehder, 2011). The psychological mechanism by which the statistical relationship between the state of the environment and internal representation is achieved is not the primary focus of these models, rather finding the formal expression of the statistical relationship between cues, uncertainty, and stimulus such that human behavior is accurately predicted. Once the "right" statistical relationship is uncovered, conclusions can be

drawn about the algorithm that best reflects that relationship, and inference can be made as to whether that is indeed what the brain is doing (Griffiths et al., 2012). This approach implicitly assumes that performance or information is optimized, which, of course, does not have to be the case—in fact, a case can be made that energy efficiency or processing time, not information, are what cognitive systems optimize (Friston, 2010; Markman and Otto, 2011).

In any case, ideal observer models have not been prominent in comprehension and production apart from models of reading, speech perception, and rule learning in language (cf. Legge et al., 1997; Norris, 2006, 2013; Goldwater et al., 2009; Frank et al., 2010; Toscano and McMurray, 2010; McMurray and Jongman, 2011; Norris and Kinoshita, 2012). The paucity of ideal observer models in sentence parsing is particularly striking given that we arguably might know more about the formal descriptions of the representations being processed during language use (i.e., formal linguistic representations, perhaps especially during speech perception) than we do about the formal descriptions of levels of representations for visual objects and scenes, or multi-modal sensory representations. One reason ideal observer models might not have taken theoretical hold in parsing, apart of EBP, might be the difficulty in constraining or separating the likelihoods of language processing outcomes that are embedded in the perceptual tasks (button pressing, reading, and making overt linguistic judgments) that most psycholinguistic studies employ. Differences in the task demands of these paradigms may mask, or at least mix in non-straightforward ways, with reliably estimating "pure" language processing likelihoods. Moreover, the source of priors and how they are acquired and updated remains unknown. However, core principles from ideal observer models of perception, namely that including estimates of uncertainty can expose the nature of the problem the brain solves, may be suitable for the addressing the computational challenges that language processing presents.

Cue Combination and Integration

In both psychophysical and neurobiological models of perception, *cues* are any signal or piece of information that reflect the state of the environment (Fetsch et al., 2013). For example, when perceiving and localizing an object to act on, such as trying to catch a toddler who is screaming while running away from you, one cue is likely the visual contrast information created by the toddler moving across the visual scene, and another is the screaming, or more accurately the change in interaural time of the screams as the toddler moves in relation to your ears. And lastly, cues can come from any proprioceptive or tactile stimulation that is generated as you prepare to grab your toddler before s/he runs into traffic. Our brains combine and integrate these cues, often from different modalities, to form a stable percept upon which to act (see **Figure 1**, Ernst and Bülthoff, 2004). The key to stable and robust perception given sampling uncertainty is the integration of multiple sources of sensory information via two important psychophysiological operations, cue combination and cue integration. *Cue combination* is the process of combining cues via summation, and describes interactions between cues that are not redundant in the information they carry. Cues may be in different units during combination, and may signal complementary aspects of the same environmental property. For example, when knocking on a door, one perceives the knock as emanating from the location where one knocked. This percept is the result of the combination of sensory signals from vision, audition, and proprioception (see **Figure 1**). After cue combination, comes integration, or the weighting of the cues by estimates of their reliability as cue to the true stimulus. *Cue integration* describes an interaction between cues of the same units that may carry redundant signals, and that regard the same aspect of the environment. Evidence across different domains and species implicate cue integration as the mechanism from which stable percepts emerge (Deneve et al., 2001; Ernst and Bülthoff, 2004; Fetsch et al., 2013). Summation is the canonical neural computation, and Carandini and Heeger (2012) argue that normalization, the principle operation underlying cue integration, is also a canonical population-level neural computation for brains of all levels of complexity.

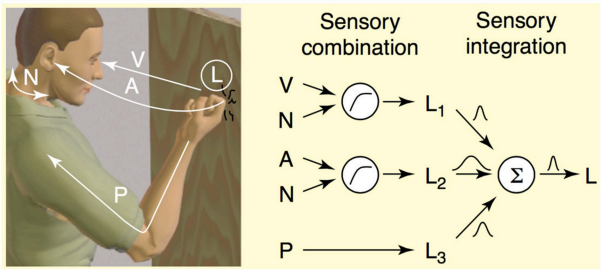


FIGURE 1 | Illustration of cue combination and integration of the perception during knocking on wood from Ernst and Bülthoff (2004). Sensory cue combination occurs between sensory signals that are not redundant, which can be represented in different units or coordinate systems, and which might reflect complementary aspects of the environment, for example visual or auditory information (Ernst and Bülthoff, 2004). This figure from Ernst and Bülthoff (2004) depicts how three sensory estimates about the location (L) of the knocking event are combined to form a stable percept. Information from visual (V), auditory (A), and proprioceptive (P) sensory percepts comprise three different signals about location. Before these signals can be integrated, V and A signals can be combined with the proprioceptive signals (N) to be transformed into body-centric coordinates with the same units. Following that, the three signals (L_1, L_2, L_3) are integrated with their reliabilities to form a coherent percept of the location of the knocking event. Sensory cue integration occurs between so-called redundant signals, or signals that are in the same units or coordinates and that reflect the status of the same aspect of the stimulus in the environment.

Cue integration is typically expressed in an estimate of the likelihood of the stimulus being present in the environment ($\hat{S}$) given the cues[3] ($c_1 ... c_n$) and scaled by the reliability of those cues ($\hat{r}_1 \ldots \hat{r}_n$):

$$\hat{S} = \sum_{i=1}^{n} c_i \hat{r}_i + \ldots c_n \hat{r}_n, \quad \hat{r} = \frac{1}{\sigma_c^2}$$

Equation (1) From Ernst and Bülthoff (2004) the equation above describes the processing moment at the onset of a stimulus. It

[3]Summation of the activation of the neural population tuned to a given stimulus or feature of the environment.

describes the activation state of a neural population that codes for a given sensory representation. This representation can be said to emerge from integrated sensory cues.

An estimate of *cue reliability* ($\hat{r}$) is the inverse variance of the distribution of inferences made based on a given cue (Bülthoff and Yuille, 1996; Jacobs, 2002). The smaller the variance in the relationship between cue and stimulus, the more reliable the cue is. Correlation between cues also affects their reliability: a cue is regarded as more reliable if the inferences based on it are consistent with the inferences based on other cues in the environment (Averbeck et al., 2006). If a cue is inconsistent with other cues, it is regarded unreliable. Studies on cue reliability have shown that cues that have not changed their value in the recent past are weighted more strongly (Jacobs, 2002). Thus, returning to our example of the screaming toddler, cue combination summates activation from the sensory populations associated with the visual, auditory, proprioceptive, and tactile stimuli that issue from chasing a screaming toddler. Upstream from these primary sensory cue population codes, other neural populations code for combined or composite representations of these cues. At each stage of representation, cue integration weights the representation by that cue's reliability. The reliability of combined cues is equal to the sum of the individual cue reliabilities, so the only neurophysiological operations required are summation and normalization (Fetsch et al., 2012). I will discuss the appeal of this point in Section A Neurophysiologically Inspired Mechanism for Neurobiological Models of Language. Whether cue reliability is best thought of as a prior in a Bayesian framework, or as a probabilistic variable in a Statistical Decision Theory framework is an open question (Maloney and Zhang, 2010; Rehder, 2011). In any case, even non-optimal weighting by cue reliability is probably a better estimate than an individual trial data or single sample measurements (Ernst and Bülthoff, 2004).

CAPTURING MULTIPLE DISTINCTIONS IN PARSING

A desideratum of psycholinguistic theory is a taxonomy of the mental representations and computational mechanisms that language use requires. A particularly satisfying theory would unify the mechanisms occurring during diverse computations such as speech perception, word recognition, parsing into phrase structures, establishing referential and agreement relations, forming long-distance dependencies, and forming discourse representations. Such a theory would have general principles derived from domain general canonical neural computations, and would hold for both for comprehension and production. Processing difficulty would be predictable from first principles, that is, from how the representations at stake are generated. Traditionally, mechanistic theories of language comprehension and production have proposed multiple language-specific mechanisms, often operating at distinct levels of linguistic representation. These have been as diverse as lexical access, reanalysis, binding, lemma selection, and unification (Frazier and Fodor, 1978; Marslen-Wilson and Welsh, 1978; Swinney, 1979; Clifton and Frazier, 1989; Ferreira and Henderson, 1991; Levelt, 1999; Hagoort, 2005), or have invoked heuristics like Minimal Attachment, Late Closure, the Active-filler Strategy, Attach Anyway (Frazier and Rayner, 1982; Frazier and Clifton, 1996; Fodor and Inoue, 1998). Other impactful approaches to parsing have focused on metrics to quantify the difficulty of certain structural configurations in terms of capacity limits on memory, or the number of dependencies to be resolved, or the number of parses to be considered, but not on mechanism *per se* (Just and Carpenter, 1992; Gibson, 2000; Vosse and Kempen, 2000). Yet other dynamical systems approaches to parsing derive empirical phenomena, such as *local coherence*, where local match between constituents' features can override the global parse from architectural aspects of the model (Tabor et al., 2004; Tabor and Hutchins, 2004). A notable antecedent psycholinguistic theory based on cues, albeit with a different goal and level of analysis, comes from Bates and MacWhinney (1987)'s Competition Model (CM), a lexicalist framework focused on the acquisition of grammar in the face of the challenge of cross-linguistic variation. As its name suggests, its main processing claim is that lexical representations compete with each another for case and thematic role assignment during comprehension, and that languages differ in how information is expressed via cues. The CM is an important antecedent for cue integration because it invokes both the notions of cues and cue reliability, but in different senses than in the perceptual literature and thus, than herein. It posits that languages vary in how their forms cue meaning, and in how linguistic form and function are related by cues, and is largely concerned with how different linguistic representation types cue argument relations in different languages and how cues and their reliability facilitate language acquisition. However, the framework I will outline draws strongly on the notion of cues and their reliabilities as internal representations, processed by a neurophysiologically plausible mechanism, rather than on cross-linguistic variation in how information is carved up to cue between form and meaning.

In some ways, mechanistic approaches are just as vulnerable to the criticism of falsifiability that Bayesian approaches are—just as you can change the priors to fit your data—you can, similarly, change the number of hypothesized mechanisms at stake, fail to generate falsifiable hypotheses or testable predictions, or arbitrarily change the architectural bottlenecks in your process model to account for your data (Bowers and Davis, 2012; Griffiths et al., 2012). How does one keep from "over fitting" a process model? Moreover, the frameworks that developed past hypothesized language-specific mechanisms were steeped in the modularity debate, which naturally focused on questions about what operations are language specific or not (Fodor, 1983), and whether processes operated in serial or in parallel (Frazier and Clifton, 1996). Though there is less worry now about sterility and modularity of linguistic representation, and more about incrementally in language processing, it remains a fact that the brain can be said to be modular in its organization (Carandini and Heeger, 2012; cf. Fedorenko et al., 2012) though likely with interesting and important overlap or redundancy in coding in diverse systems (e.g., Schneidman et al., 2003; Puchalla et al., 2005; Rothschild et al., 2010). This presents our desired linking hypothesis between psycholinguistic and neurobiological

theories with a conundrum wrapped in a mystery: capturing the incrementally of language processing within a modular system of neural populations, whose coding we do not yet know how to read. In other domains of cognition focused population codes, the relevant questions become: what factors determine the organization of neural populations, what are populations coding for, and how are those representations transformed from population to population (see Pouget et al., 2000; Averbeck et al., 2006)? Translating these questions to a psycholinguistic level of analysis, we then must ask whether signals in brain or behavior that reflect representation of linguistic units can be detected, whether such a modular neural architecture can indeed capture important distinctions for linguistic representation and processing, and whether cue combination and integration alone can account for language processing from speech and visual onset all the way to higher level meaning.

Language Comprehension as Cue Combination and Integration

Can a satisfying analogy can be made between language comprehension and perceiving a complex natural environment? Like object perception or localization, scene perception, or motor control, language processing is multimodal. In conversation, language comprehension minimally involves integration of auditory and visual information[4]. All this must occur while planning and producing language in return. Furthermore, language use is highly goal-directed and joint, an issue that is rapidly gaining theoretical importance (Pickering and Garrod, 2004; Gambi and Pickering, 2013; MacDonald, 2013). But aside from the issues of modality and joint-action, language may present a processing situation that fundamentally differs in the kind of representational relationships that the brain must form in order to explain linguistic taxonomy. Information from multiple, sometimes hierarchical, sources of formally discriminable representations must be perceived from the environment. Extracting linguistic representations from a speech or visual input may be, in some ways, analogous to the binding problem in vision and attention (cf. Treisman, 1999). In both situations, information that is distributed over time and space at different frequencies must be grouped or bound into higher-level representations for processing to occur. Cues, whatever they may be, from each sensory input level are combined and integrated with their reliability estimates, and emerge as a linguistic representation, e.g., a phoneme or phrase. Populations coding the reliability of a given representation as a cue to higher-level representations are activated and updated. Those reliabilities are integrated with the population code representation for a given representation, which in turn produces the next level of representation.

As in the psychophysical literature, most of the explanatory work would be carried out by cues, a notion that is difficult to define both in the positive (what cues are), and in the negative (what can't be a cue). In fact, often the term "cue" is treated as if should be implicitly understood, as in, as if it has no specialist or jargon meaning. In the perception literature, a cue is any sensory information that gives rise to an estimate of the state of the environment (Ernst and Bülthoff, 2004). Here I will augment that definition as follows: a psycholinguistic cue is any internal representation that signals, indicates, or is statistically related to the state of some property of the environment relevant for language processing. Thus, a cue to a given psycholinguistic representation is simply any representation that is reliably related to that given representation, in contrast with a representation that is not related to it. The only way for this simple definition of cue to become explanatory is if it can speak to how abstract linguistic representations might be formed from perceptual inputs, or more specifically, formed from an interaction or convolution of sensory percepts with extant knowledge (read: other representations) in the brain[5]. The problem of satisfactorily defining a cue for functional use in a process model bumps up against the even harder problem of defining mental representation, or defining what perceptual or cognitive features are. Both of these philosophical challenges are, luckily, beyond the scope of this model. However, the functional role of cues may be to simply to map out the structure, path, or links between representations as they are activated in moment-to-moment processing. In this sense, is it not so much what cues precisely are that matters (although that is no doubt an important, troubling question), but which representations cue which other representations to form a map of language processing, from percept to abstract representation that matters for a model. Thus, cues are representations of linguistic input *and* what links those representations in a "chain" for processing from sensory input to abstract representations. I will sketch how a cue integration model might handle processing from speech onset to phrase or sentence comprehension (see **Figure 2** for visual illustration). I simplify the representational levels at stake as: phonemes, syllables, morphemes, words, phrases, syntactic and event structures, and discourse context.

Sensory Resampling to Recover Hierarchical Representations

In the case of linguistic representations, aside from the first perceptual cues to enter the processing stream, further cues must come from the same sensory input: a sort of resampling of the sensory percept, or a form of perceptual inferencing (Ernst and Bülthoff, 2004). This resampling would recover hierarchical representations in memory that are activated by that percept, via the same cue integration mechanism that is hypothesized to work for exogenous cues. In other words, cue integration can take as its input an endogenously stimulated representation or set of sensory features (e.g., phonemes from acoustic features, or on a higher level, morphemes and lexical entries), and output another

[4]Though the highest levels of hierarchical representation are reached via an arguably single modality in phone conversations, sign language, and reading—it is an empirical question as to whether processing in these cases activates linked representations generated from other modalities.

[5]Such an assertion then attributes most of the burden (and magic) of online language processing onto language acquisition. In a system of representation where only cues and their reliabilities are computed to activate the next representation, it is this existing knowledge that parses input and links up the right representations properly. But how in the world are these all important, pre-existing, parse-making representations acquired? See Section Cue Integration in Language Acquisition and Bilingualism for more discussion but no definitive answers.

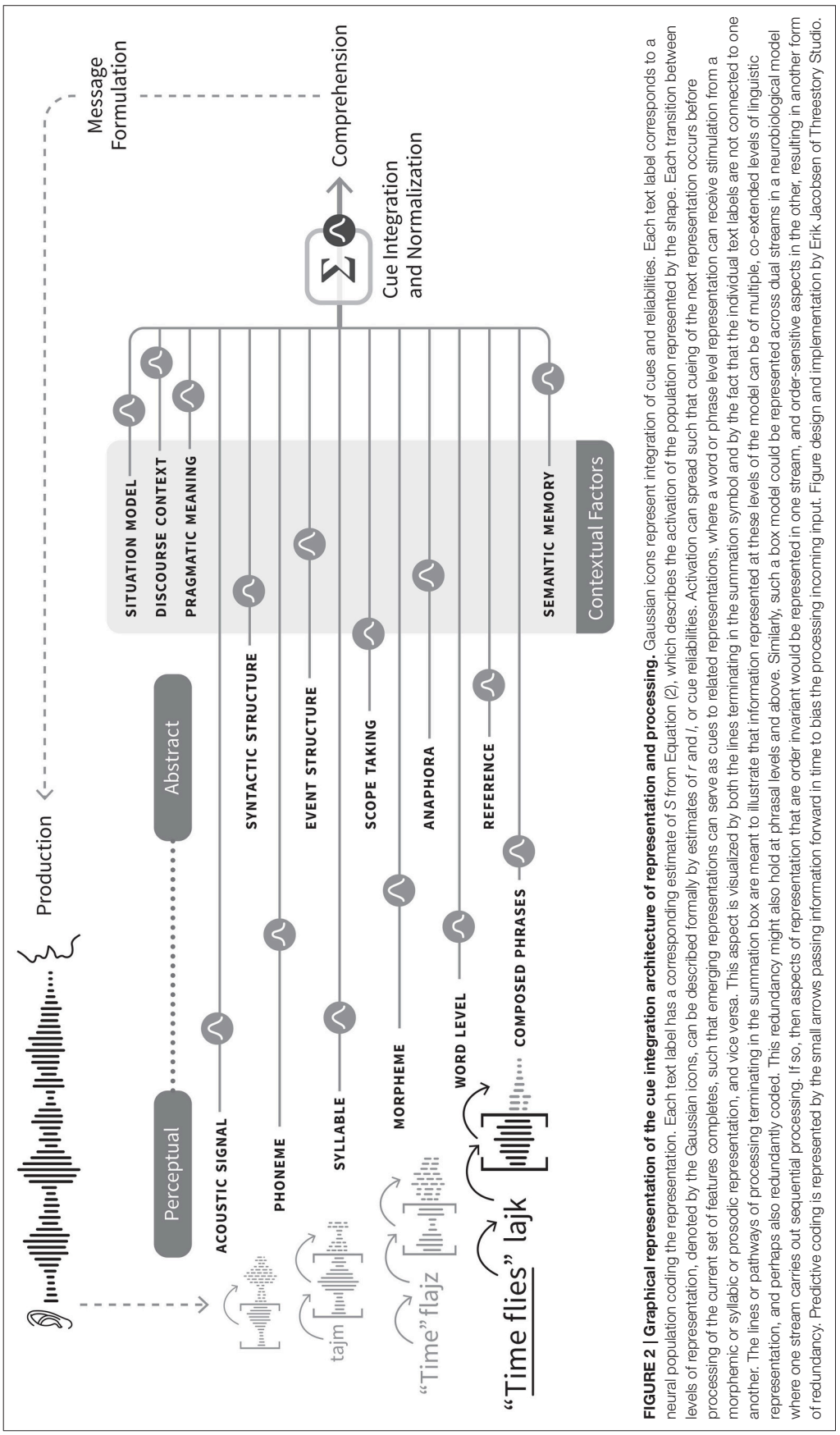


FIGURE 2 | Graphical representation of the cue integration architecture of representation and processing. Gaussian icons represent integration of cues and reliabilities. Each text label corresponds to a neural population coding the representation. Each text label has a corresponding estimate of S from Equation (2), which describes the activation of the population represented by the shape. Each transition between levels of representation, denoted by the Gaussian icons, can be described formally by estimates of *r* and *l*, or cue reliabilities. Activation can spread such that cueing of the next representation occurs before processing of the current set of features completes, such that emerging representations can serve as cues to related representations, where a word or phrase level representation can receive stimulation from a morphemic or syllabic or prosodic representation, and vice versa. This aspect is visualized by both the lines terminating in the summation symbol and by the fact that the individual text labels are not connected to one another. The lines or pathways of processing terminating in the summation box are meant to illustrate that information represented at these levels of the model can be of multiple, co-extended levels of linguistic representation, and perhaps also redundantly coded. This redundancy might also hold at phrasal levels and above. Similarly, such a box model could be represented across dual streams in a neurobiological model where one stream carries out sequential processing. If so, then aspects of representation that are order invariant would be represented in one stream, and order-sensitive aspects in the other, resulting in another form of redundancy. Predictive coding is represented by the small arrows passing information forward in time to bias the processing incoming input. Figure design and implementation by Erik Jacobsen of Threestory Studio.

pattern of activation or representational state (e.g., syllables from phonemes, or on a higher level, phrases). The representations from the last cycle of processing serve as cues to the next level of representation or cycle of processing.

The architectural hypothesis is that each level of representation is a cue to higher levels of representation, resulting in a cascaded architecture: phonetic features are cues to phonemes that are cues to syllabic and morphemic representations, which in turn are cues to lexical and phrasal representations, leading to phrase-based parsing and larger sentential or event structures. **Figure 2** illustrates how the phrase *Times flies like an arrow* would be processed using cue integration. Activation can spread such that cueing of the next representation occurs before processing of the current set of features completes, such that emerging representations can serve as cues to related representations, where a word or phrase level representation can receive stimulation from a morphemic or syllabic or prosodic representation, and vice versa[6] (see **Figure 2** for illustration). As the phonemes in *time* are parsed, they cue the morpheme and word representations of "time," which in turn activates syntactic or structural representations, and conceptual representations associated with *time* (e.g., phonotactically licensed syllables, verbs, phrases, related semantic knowledge).

Population coding parameters would constrain how information is represented in the model, but how can such a radically interactive and redundantly coded system be represented? An efficient way to represent a true multitude of representations without overcommitting neural "real estate" might be opponent channel processing. In color vision, a multitude of colors are perceived from photons interacting with photopsin proteins that are tuned to different frequency spectra in cone cells in the retina. The activation patterns of these cells together form opponent channels, where a given channel can be said to detect the difference in activation between cone cells (with different photopsin proteins) tuned to two opponent ends of a spectrum of light (e.g. red and green, blue and yellow), rather than representation via a series of cells or ensembles dedicated to each color or frequency band. Such an opponent system has also been implicated for spatial coding in auditory cortex, where, while most auditory neurons respond maximally to sounds located to the far left or right side, few appear to be tuned to the frontal midline (Stecker et al., 2005). Paradoxically, psychophysical performance reflected optimal acuity in the frontal midline, thus the existence of an opponent process system synthesized these apparently conflicting findings (Stecker et al., 2005). Opponent processing may be a possible architectural feature to represent a multitude, or even a discrete infinity, of linguistic representations via cue integration (e.g., of minimal pairs or other representations in complementary distribution), though it thus far observed has only been observed in much more primary or lower-level sensory processing stages. An opponent channel representational system, operated on by cue combination and integration, would likely be able to flexibly and efficiently code the number of representations needed for such a massively interactive architecture without taking up an implausible amount of neural real estate.

While cues determine which representation is activated, cue reliabilities determine the strength of the evidence for a particular representation and thus how good of a model of the world the system has. To create and maintain an accurate and robust set of representations reflecting the linguistic environment, reliabilities need to reflect local context as well as latent knowledge, or a global prior. Cue integration can account for processing variables in one of two ways: either by modulating the information expressed in the cue reliabilities, or by modulating the circuit of representations, the order or domain of cue computations. Information from memory might be expressed as both an immediate prior (*r)*, representing recent processing and the local environment, similar to the notion put forth by Jaeger and Snider (2013), and as a more stable, long-term set of global priors (*l*) that reflect information like discourse context and pragmatic meaning, and semantic and world knowledge. In the set of expressions below, I separate reliability into two terms (see Equation 2). Although both terms are subject to summation, I want to make it clear that they represent different sources of uncertainty, that are likely to be represented by different populations, or redundantly on different levels.

$$\hat{S} = \sum_{i=1}^{n} c_i \hat{r}_i \hat{l}_i + \ldots c_n \hat{r}_n \hat{l}_n, \qquad \hat{r} = \frac{1}{\sigma_{ci}^2}, \qquad \hat{l} = \sum_{i=1}^{n} \frac{1}{\sigma_{c}^2}_{i}$$

Equation (2) Ernst and Bülthoff (2004)'s expression of likelihood of activation adapted to parsing. It describes the activation state of a neural population that codes for a given representation. This estimate of activation is composed of cues (e.g., representational features or any representation), weighted by their reliability, or the likelihood that the stimulus is in the environment given the cue. Estimate of *S* is the likelihood a level of representation is activated by the cues or representational features denoted by *c,* weighted by their reliability *r*, the recent inverse variance of the link between that cue and its related or antecedent representation, and by its latent reliability *l*, the global reliability of that cue over a longer time scale. Estimates of *S* would describe the activation represented by any of the shapes denoted in **Figure 2**, while estimates of *r* and *l* would be denoted by the arrows feeding forward on back between each level of representation.

To unpack the cue integration process, we can take the example phrase "Times flies like an arrow…" from **Figure 2**, and examine how the first two words *time* and *flies* would be extracted from the phonemic stage to achieve the morphemic- lexical stage. I will outline how Equation (2) would describe this step in processing. The phonemic string /tajmflajz/ has been parsed from acoustic information[7], so the next step is for /tajmflajz/ to cue the morphemes/words [tajm|*time*] and [flajz|*flies*] into the phrase *time flies:*

Cartoon process: */tajmflajz/->[taj][m][flajz]-> time, flies -> Time flies*

[6]Also in feedback or top-down connections, which Singer (2013) claims are more numerous in neocortex than feed-forward connections.

[7]Out of fear, and for simplicity's sake, I am skipping how acoustic representations are transduced into phonetic and then phonemic representations.

$$S\widehat{[[tajm]]} = \sum_{i=1}^{n} c_{[taj]}\hat{r}_{[taj]}\hat{l}_{[taj]} + \ldots c_{[m]}\hat{r}_{[m]}\hat{l}_{[m]},$$

$$\hat{r} = \frac{1}{\sigma_{ci}^2}, \qquad \hat{l} = \sum_{i=1}^{n} \frac{1}{\sigma_{c\ i}^2}$$

$$S\widehat{[[time]]} = \sum_{i=1}^{n} c_{[tajm]}\hat{r}_{[tajm]}\hat{l}_{[tajm]},\ S\widehat{[[flies]]}$$

$$= \sum_{i=1}^{n} c_{[flajz]}\hat{r}_{[flajz]}\hat{l}_{[flajz]},$$

$$S\widehat{[[time\ flies]]} = \sum_{i=1}^{n} c_{[tajm]}\hat{r}_{[tajm]}\hat{l}_{[tajm]} + c_{[flajz]}\hat{r}_{[flajz]}\hat{l}_{[flajz]},$$

$$\hat{r} = \frac{1}{\sigma_{ci}^2}, \qquad \hat{l} = \sum_{i=1}^{n} \frac{1}{\sigma_{c\ i}^2}$$

Equation (3) Describing processing moments from the phonemic representation of /tajmflajz/ cueing the words *time* and *flies*, and finally the phrase *Time flies*

We can already see that the description of the activation of the model or system as described by Equation (3) is completely dependent upon the time step or processing moment that we choose to analyse or observe. The importance of time step may not be an issue for implementing a computational model based on cue integration that is dynamic in its activation, but it certainly is a theoretically troubling issue. Would processing moments or cycles be determined solely by the external stimulus, e.g., by speech envelope? Or would the current state of the system upon input instead structure processing time, for example, actually result convolving current activity with the incoming physical (and later, the abstract linguistic) properties of the input? I will explore this problem more in Section A Neurophysiologically Inspired Mechanism for Neurobiological Models of Language.

A second important consequence of cue integration is that it implies a hybridized notion of modularity: perceptual representations might still be encapsulated in the Fodorian sense, but once representations become either multi-modal, or are resampled as the cue other representations further up the processing stream, they are no longer so. In fact, as a reviewer pointed out, higher levels of representation in such a model would flatly deny Fodorian modularity (see Fodor, 1983). Another way of putting it is that, under cue integration, early representations, which tend to be perceptual, may be encapsulated until they are summated with other cues. This hybridized modularity would also play out in terms of deeming the pathways and networks that process the representations to be domain-specific or not.

Returning to the important psycholinguistic notions captured by EBP and CBR, how would a cue integration model cache out surprisal and interference? Surprisal might be cached out in terms of sub-optimal cue integration with reliability, poor trading off of global cue reliabilities for recent ones, such that global reliabilities are overweighting the current representation. Interference would amount to sub-optimal cue combination, where the cues for a competing parse or related representation activate an "attractor" representation, instead of the true stimulus. It would arise when sub-threshold activation is shared between representations that share features with the input, a form of cue overload, and may or may not fully activate the "attractor" representation. Cue overload in such a system would still depend upon how diagnostic a cue, or summated cue set, is to a unique representation in the system. Garden-path effects and other parsing ambiguities might be cached out in terms of poor estimates of recent reliabilities compared to global ones, such that summated cues point to ultimately ungrammatical representations. A cue integration process model would extend the notion of cue combinatorics during retrieval and formation of non-adjacent dependencies (Clark and Gronlund, 1996; Lewis and Vasishth, 2005; Van Dyke and McElree, 2011; Kush et al., 2015) to a general processing principle and makes a claim about how cues are combined with one another. The model would assert that processing difficulty is essentially always a form of cue overload, which stems from architectural first principles of how activation of representations occurs and how uncertainty flows through the system dynamically.

Even the first input step in sketching a processing stream is grossly oversimplifying and glossing over important and vibrant subareas, especially in the neurobiology of speech perception (Hickok and Poeppel, 2007; Poeppel, 2014). Recent compelling evidence suggests that neural populations entrain with an auditory stimulus using acoustic-phonetic "sharp edges" to latch onto the speech envelope (Luo and Poeppel, 2007; Doelling et al., 2014; see Poeppel, 2014 for discussion). Giraud and Poeppel (2012) show an emerging role for oscillatory activity as entrainment with speech envelope and syllable structure. This entrainment could be performing cue combination and integration of phonetic features into phonemes, but a clear experimental question is if cues and their reliabilities are coded in or recoverable from oscillatory activity. Such a simple process model must be able, at minimum, to capture the vagaries of speech perception, it being the stage of language processing most firmly grounded in perceptual processing (Samuel, 2001; Samuel and Kraljic, 2009).

Representations and Grammar

An issue that will clearly determine the success of a cue integration process model is the nature of the representations the model posits. The basic representational claim of a cue integration process model is that representational features make up a level of representation, and serve as cues to subsequent levels. They do so in a cascaded way and incorporate at least two error terms. This would mean that the system's organization comes from, or even just *is* the grammar of the language it was trained on. But probably any cue-based model also makes that claim that ungrammatical representations might be formed if the rest of the cues, i.e., non-structural ones, point toward a given representation. One way to avoid the "bag of words" problem (Harris, 1954), where semantic and other non-structural features dominate over structural relations would be to simply weight syntactic features more strongly in their reliability.

Without a traditional mechanistic structure that assumes multiple operations, one possible consequence is that

representations need to be similar to something like slash categories in a combinatorial constituent grammar, as in Combinatorial Categorical Grammar (CCG; Szabolcsi, 1989, 2003; Steedman, 1996, 2000; Jacobson, 1999). If they were, then the dependencies that are cached out as empty categories in other grammars, as well as other forms of dependency, could be carried forward during processing without the need for positing constructs like buffers or maintenance[8], because the dependency is represented as a grammatical feature that can "percolate[9]" to the highest tree, representation or population code. Separate operations for retrieval and interpretation may also become moot if grammatical features (of which dependency is now just one example of) can percolate up the path of population codes. By caching out problems like non-adjacent dependency as representational feature parsing, CCG, and perhaps cue integration, perform the classic programmer's trick of changing data structures to increase expressive power when of the processing architecture. However, this trick only means that the difficulty is merely transmogrified—now the cue integration process model is generating hypotheses about both psychological processing mechanisms *and* about the nature of representation. This is especially problematic because traditional dependent measures (e.g., performance on a task, brain responses, but especially reaction times) cannot discriminate between effects arising from differences in processing speed (a proxy for mechanism) and differences in representation strength or other aspect (Wickelgren, 1977; Davidson and Martin, 2013). This means that experimental designs will have to be careful not to conflate predictions about representation with predictions about mechanism itself. The speed-accuracy trade off procedure (SAT; Reed, 1973) offers a way to measure effects of processing speed orthogonally from representation-based differences, but it relies on an overt metalinguistic judgment. Given cue integration's grounding in perception, it is not unreasonable to think that SAT could be applied to study both the representations of cues and their reliabilities, especially because discriminability between signal and noise, or *d'*, is composed of hits (yes responses to trials from the signal distribution) and false alarms (yes responses to trials from the noise distribution). Nonetheless, deriving testable predictions about the natures of the representational architecture in a cue integration process model for behavioral data will be challenging.

A NEUROPHYSIOLOGICALLY INSPIRED MECHANISM FOR NEUROBIOLOGICAL MODELS OF LANGUAGE

How can we formulate a meaningful linking hypothesis between a psycholinguistic process model and current circuit-based neurobiological theories of language? First we must try to formulate it in term of mechanisms that are both grounded in canonical neurophysiological computation and psychologically meaningful. The class of neurobiological models exemplified by Hickok and Poeppel (2004, 2007) focus on sub-lexical processing and speech as the first information-processing hurdle. Such models tend to have more fine-grained, detailed claims about neurobiological architecture than models that focus on syntactic or semantic processing (Hagoort, 2005, 2013; Friederici, 2012), although some very recent phrase and sentence level models are becoming much more articulated in the complexity of the dual-stream circuitry and in claims about directionality and interaction of processing streams (Rauschecker, 2012; Hagoort and Indefrey, 2014; Bornkessel-schlesewsky et al., 2015; Friederici and Singer, 2015). In any case, trying to find a mechanistic foothold can be difficult. Cue combination and integration maps broadly onto the general concept of Unification from Hagoort (2005)'s Memory Unification and Control model, as a mechanism to combine processing units into larger, hierarchical structures. In MUC, unification is separated by modality or representational type, such that phonological, syntactic and semantic unification are separate, as are the processing streams that deal with them (Hagoort and Indefrey, 2014). A cue integration model would not stipulate encapsulation by formal representation class but, rather, by order of cue summation and thereby connectivity of the populations, which may or may not turn out not to be equivalent to representation class.

The cue integration model also differs from Unification in that it makes the claim that uncertainty, specifically cue reliability, is integrated with the population activation for a given cue or cue set. This would mean that cue reliabilities would need to be dynamically updated, and more broadly, that the representations carried by a given neural circuit would need some element of flexibility and would be robust due to redundant coding of features across certain populations. They would also need to be robust, and so redundantly represented in multiple populations. Friederici and Singer (2015) propose that the sparse, flexible, feature-based coding that is seen in other cognitive systems applies to linguistic representations in the brain. In such a system, there is both temporary coupling of populations coding cues or features of larger representations, as well as lasting couplings or "firmware" of anatomical assemblies, as outlined in Singer (2013). Careful experimental work would be needed to test this hypothesis and to determine if flexible sparse coding can handle formally complex linguistic representation, and furthermore, to determine which aspects of phonological, lexical, syntactic, semantic, discourse, or pragmatic representations are flexibly coded or "hard coded." Such an architecture would be highly suited to a cue integration process model but in combination with redundancy in coding to generate robust representations. Such an architecture may enable the system to represent discrete infinity.

To emphasize, the only computational mechanisms stipulated in a cue integration process model would be *summation*, the neurophysiological mechanism for cue combination, and *normalization*, the neurophysiological mechanism that integrates a cue with its reliability. If parsing and other language processing phenomena can be accounted for using only these two stalwart neurobiological mechanisms, it would be a step in the direction

[8]Along with the notion of search, both theories of grammar and processing often tacitly assume buffers and maintenance in the architectures they imply.

[9]By "percolate" I mean persist in being represented or coded in active neural populations as processing proceeds.

toward a unified theory of human information processing that includes language but is based on "brain-general" processes.

Cue Integration and Forward Models

Another powerful capacity that any process model would need to account for is the role of predictive processing in language behavior. Forward models from vision and motor control have already had some influence on theoretical work in cognition and language (Pickering and Garrod, 2007; Pickering and Clark, 2014), but have yet to be fully specified in models with clear predictions for language processing. In a classic computational model of vision, Rao and Ballard (1999) describe an architecture wherein top-down feedback connections carry predictions about bottom-up or lower-level population codes, and feed-forward connections carry residual error between those top-down predictions and the actual input. They illustrated that in this kind of forward model, architectural facts about the visual system, such as receptive field characteristics and surround suppression[10], emerge naturally. This seems to suggest that such architectural features occur as a result of cortico-cortical feedback, and that cortico-cortical feedback is a promising candidate mechanism for predictive coding (Rao and Ballard, 1999). Synthesizing predictive coding via cortico-cortical feedback with a cue integration process model, feed-forward connections would carry bottom-up activity corresponding to integrated cues and reliabilities. A subset of feed-forward cue reliability activity would be the error signal in response to predictive activation forecast via the top-down feedback circuit. Although predictive coding and forward models will no doubt play a larger role in psycholinguistic theory in the coming years, the fact that we can understand the unpredicted or unexpected utterances at all, or with reasonable ease, suggests that prediction is not the core language processing device (see also Jackendoff, 2002; Rabagliati and Bemis, 2013; Huettig, 2015; Huettig and Mani, 2016). But the fact remains that predictive coding plays a huge role in most sensory processing domains, so any model of language ought to have an architecture that can implement it using existing neural infrastructure.

Cue Integration in a Neurobiological Circuit

A cue integration process model could make contact with neurobiological models in two ways: (1) in terms of the claims being made about the cue-based computations being carried out in various neural circuits, and (2) in terms of the implied population codes or representations needed in a given circuit. The first issue returns to the question of how to falsify hypotheses about the number and kind of processing mechanisms. A way to circumvent the problem is to focus on the end-state computation or the transformation that a representation undergoes in a given processing stream.

In a similar spirit, Bornkessel-schlesewsky et al. (2015) derive a dual-route model for human language processing from speech to syntax that is rooted in primate audition (Rauschecker and Tian, 2000). The key differences between the antero-ventral and postero-dorsal pathways in Bornkessel-schlesewsky et al. (2015) is time invariance or order sensitivity: the antero-ventral stream processes or extracts increasingly complex hierarchical auditory representations with commutative properties whilst the postero-dorsal stream processes sequence information or is order sensitive. The postero-dorsal stream makes use of forward models via an efferent copy that carries predictions and detects error, enabling sequential order-sensitive processing (Bornkessel-schlesewsky et al., 2015). The cue integration process model does not make any claims about the location or make-up of language circuits, nor does it have fundamentally different assumptions about basic representation types (phonetic features, phonemes, lexical, phrasal, event, etc.) that many extant models posit. Rather, cue integration makes a specific claim about (1) the psychological and neurophysiological mechanism underlying formation of these representations (i.e., summation of population codes for cue combination and normalization of those codes for integration with reliability), and (2) the representational infrastructure (e.g., dynamic and redundant population-level encoding of feature-based representations and uncertainty about them).

The debate about the modularity of language from other cognitive systems has featured compelling arguments that theories of language evolution must shape or constrain theories of language and language processing (Hauser et al., 2002). The claim that language evolved too recently to derive a new domain-specific neural mechanism is linked to the notion that brain processes can be repurposed to suit timely organism-environment interaction needs (see Gervain and Mehler, 2010 for discussion; Knops et al., 2009). Cue integration is a good candidate for such a repurposed process. However, though the cue integration architecture can represent recursion in principle, that fact alone cannot explain why recursion is not more widely found in other representational systems in cognition (Jackendoff and Pinker, 2005). That is unsatisfying, especially if, in a hardline reductionist thought experiment, one really wants to claim that there is only one neurophysiological brain process relevant for cognition (or extraction of further representations from sensory input), and that process is cue integration. To entertain such a thought experiment further, or for such a reductionist position to be tenable, language also needs to be learnable using only cue integration over the cue-based architecture with reliabilities.

CUE INTEGRATION IN LANGUAGE ACQUISITION AND BILINGUALISM

A crucial aspect of any theory of language is that it must be learnable. How might representations be acquired under the assumptions of a cue integration process model?

The cue integration model does not radically differ from current thought on language development—it would hypothesize that linguistic representation develops in the infant as a function of perceptual cue decoding via statistical learning (Saffran et al., 1996), but that first hierarchical representations depend on

[10]Surround suppression is a characteristic of neurons in primary visual areas wherein a given neuron's activity is reduced in the presence of a stimulus outside its receptive field; lateral inhibition from neurons with different receptive fields is one possible mechanism through which surround suppression may arise (Xing and Heeger, 2000).

acquiring a cue-based architecture and cue reliabilities, which in turn shape the development of the assembly networks. Much of the same how-why camp tension exists in language acquisition between pure statistical learning-based accounts and nativist process models (Kuhl, 2004; Gervain and Mehler, 2010). Gervain and Mehler (2010) argue that the hard work for language acquisition theorists is discovering how the system combines statistical learning and rule acquisition or language-specific cues. Only from this combination can an account capture cross-linguistic variation and sensitivity to language-specific cues in infants and neonates (Kuhl, 2004; Gervain and Mehler, 2010). To this end, Gervain and Mehler (2010) synthesize nativist and statistical learning accounts of speech processing up to the acquisition of morphology, concluding that some types of linguistic representations may be more suited to statistical learning (e.g., consonants) than others (e.g., vowels). But the challenge lies in how acquisition occurs in learning situations where, for example, frequent monosyllabic speech that arises as in some infant directed speech and even in some languages, which renders statistics like transitional probability useless (Gervain and Mehler, 2010). Under their account, acquiring complex hierarchical representations must capitalize on both the statistical information from the linear of sequences, and on language-specific cues, or the formal representations of a particular language, but how that tradeoff or interaction occurs is of course unknown. The cue integration process model offers an architecture that may be able to capture both statistical learning aspects (via reliabilities) and rule-based aspects (through assemblies or cascaded cues networks). In order to avoid some of the same criticisms lodged earlier in this article, the cue integration model needs to be able to derive abstract hierarchical representations from noisy, sparse inputs with few priors. That seems dubious at the moment, mainly because the representations or the bias toward forming certain types of representations would have to be innate. This situation echoes the learning problem that statistical models usually face: how do you parse input without the representations to do so? In other words, how do you count anything if you don't know what it is you are trying to count? I turn to a model of concept learning for inspiration because learning by analogy seems to avoid many of the pitfalls of both nativist and statistical accounts (Doumas and Hummel, 2005), as well as having some striking computational overlap with current neurobiological models of language.

At least at the level of the sentence, the tension between statistical and nativist perspectives might be eased somewhat by well-articulated claims about acquisition of relational concepts like *above*, *bigger*, or *more*. The Discovery of Relations by Analogy (DORA) model of relational concept development by Doumas et al. (2008), uses associative learning to create symbolic, hierarchical relational concepts from linear input sequences. DORA learns multiple argument predicates using time or onset of activity in sub-nodes, or systematic synchrony or asynchrony of firing of the sub-nodes representing each argument[11]. In other words, DORA learns *bigger than* (X, Y) by predicating *larger* (X) and *smaller* (Y) and combining these single argument predicates by their occurrence in time, such that the model can discriminate between X is bigger than Y and Y is bigger than X (see Doumas et al., 2008 Figure 3 for illustration). Such a strategy would work well in a redundant, flexible architecture that is also self-organizing and associative in nature (cf., Singer, 2013; Friederici and Singer, 2015). Firing asymmetry offers an additional level of description or representational state for the model without positing another psychological mechanism or neurophysiological process. Modeling, in combination with empirical work, would of course be needed to substantiate any of these claims.

For the bilingual brain, the cue integration model has modest implications but casts several existing questions in relief. First, that reliabilities and cue architecture may or may not be shared between languages (Nieuwland et al., 2012). Second, that age of acquisition might determine how assemblies are formed (Nieuwland et al., 2012). Third, proficiency may be cached out as differences in network density, representational interconnectedness, or unstable reliabilities, all of which could underlie non-native performance for bilinguals. If assemblies are malleable until the critical period is over, at which point only reliabilities are in flux as a function of language experience, any subsequent language learning would require the system to use alternate circuits to form new language-related assemblies, resulting in differing neural infrastructure that can (but does not have to) affect the competence and performance of late bilinguals.

Cue Integration in Production and Dialogue

Regarding performance, the challenges facing an integrated theory of comprehension and production endure. Questions like whether the same representations are used in comprehension and production or whether analogs or "mirror image" representations are working in concert during production and comprehension are exciting but difficult to test. Brain imaging evidence suggests that similar areas are engaged during production and comprehension (Rauschecker and Scott, 2009; Menenti et al., 2011) but whether the representations at play are identical or analogous is not yet clear. Certainly an important interaction occurs that leads to suppression of activity in auditory cortex in response to one's own speech (Numminen et al., 1999). Cue integration would make a claim about the process through which representations are activated during production, and there is no principled reason why the cue integration process and cue-based architecture cannot be the same in both processes. However, reliabilities pertaining to the representations might need to be different for comprehension and production. Regarding the claim that prediction is based on production (Pickering and Garrod, 2007, 2013) and the claim that production difficulty is at the root of comprehension difficulty (MacDonald, 2013), cue integration forces an opposing view. Cue integration stipulates that the cue-based architecture for language arises from perceptual processing. There are several difficult challenges for the account to claim otherwise: first, if cue integration is a repurposed neurophysiological mechanism from perception, and it gives rise to linguistic representations from auditory percepts, then it is fundamentally based on comprehension, at least during

[11]Note the similarity in time-based mechanisms with Bornkessel-schlesewsky et al. (2015) and Giraud and Poeppel (2012), and similar to the notion of noise correlation in population coding put forth by Averbeck et al. (2006).

acquisition. Secondly, comprehension occurs before production during development, furthering support for basing at least the origins of linguistic representation in comprehension. Third, receptive vocabulary is larger and accrues faster in development, and is larger in bilinguals (Benedict, 1979; Laufer, 1998), so it is unclear how these facts fit into a model where comprehension and production draw on exactly the same representations. These arguments do not exclude the possibility that a significant portion of cue reliability during comprehension is uncertainty stemming from dynamic production-based experience in the adult, leading to a situation where comprehension difficulty is rooted in a production-based variable, as MacDonald (2013) argues.

Producing an utterance in the cue integration architecture would go as follows: activation for an event structure cascades down representational levels in a planning-cycle-sized chunk. The cue-architecture basically fires in reverse order, and reliabilities include uncertainty from articulatory planning and other production-based priors. Predictive coding would also have to operate in the opposite direction. The system would still be susceptible to cue overload whether or not production and comprehension representations are identical. Coupling between processing streams or analog representations during both production and comprehension could occur.

During dialogue, language production often based on comprehension of what was just by an interlocutor. If production reverses what is top-down and bottom-up and changes the predictive coding direction, then dialogue is a cascaded engagement of this stream coupled with the comprehension stream. In dialogue, these streams become coupled between two brains, forming a sort of ultimate cacophony of synchronous and asynchronous firing. The only new claim a cue integration model would make is that cue reliabilities would then have endogenous and exogenous sources, from the speaker and interlocutor, and would crucially have to contain predictions about the interlocutor's representational states. Alignment then might be cached out as how well-entrained dialogue partners' cue reliabilities for each other are. Cues in dialogue might also place more weight on non-linguistic percepts or cues, which may end up influencing the reliabilities of linguistic representations, for example, gaze, facial expression, gesture, and goal-directed or joint-action contexts and behavior. Turn-taking and other time-based behaviors between interlocutors would be entrained with or based on asynchronous firing across speakers (Stephens et al., 2010).

Predictions from Cue Integration and Persistent Challenges for any Cue-Based Model

The real work for this developing theory is generating testable predictions. What can a simple process model based on psychophysiological principles mean for brain data and for behavior?

Given the architectural nature of the claim, a starting point might be computational models of language that are based on primate and avian auditory processing (à la Doupe and Kuhl, 1999; Rauschecker and Tian, 2000; Bornkessel-schlesewsky et al., 2015) using associationist learning to acquired symbolic representations. If such a computational model can approximate human learning and processing of language, it would still be a form of confirmatory evidence rather than an attempt at falsification. But such an implemented computational model might be able to generate finer grained predictions for electrophysiology and behavior.

Another approach to falsification might be via the manipulation of the cue relationships between representations, and of cue reliabilities, in an artificial language. This approach would try to manipulate the reliability of a phoneme as a cue to a morpheme, or a morpheme as a cue to a phrase structure, to see if participants track reliabilities and if manipulating them affects reading time. Cue integration also predicts that a noise term for each level of representation should exist. An elegant point from Maloney and Zhang (2010) is that one way to falsify Bayesian accounts it to observe that estimates of priors transfer onto other trials or related tasks. Thus, estimates of priors might be expected to transfer onto other item sets, syntactic structures, lexical items, discourse or information structures. It is yet unknown how much of a role individual differences in language experience might contribute to both recent and global priors or cue reliabilities.

Another class of predictions the cue integration model might make regard neuroimaging data. Although the relationship between something like a population code and an electrophysiological frequency band or event-related component is highly speculative at best, I will try to generate predictions both on the population level (though they are not yet measureable in humans apart from intracranial electrocorticography), and try to predict an analog for a signal our existing psycholinguistic electrophysiological dependent measures can detect. First, formal linguistic distinctions in a particular language should determine population codes. Under an opponent processing system, the opponents in a channel would be determined by that language's minimal pairs at various levels of representation. Beyond the population level, such a language-specific population coding architecture's first fundamental prediction is, certainly for abstract constructs like event-related brain potentials (ERPs), for variety of indices (i.e., different ERP components elicited by strings with the same meaning across languages) showing sensitivity to different processing variables across languages (see Bornkessel-Schlesewsky et al., 2011).

Second, if firing asynchrony is important for perceptual grouping (both in processing and in learning), then a cue integration approach predicts a lack of phase in electrophysiological signal with stimulus onset. This "delay" should be true for population codes, oscillatory activity, and ERPs. But there should be some temporal relationship with onset as a function of the number or complexity of representations being extracted from the auditory percept (Luo and Poeppel, 2007; Giraud and Poeppel, 2012; Golumbic et al., 2013), though discovering what that relationship seems very challenging. Nonetheless, discovering the relationship may make contact with neurophysiological principles about oscillatory activity, namely regarding questions as to how oscillatory activity is driven both by the temporal properties of the incoming, exogenous stimulus *and* by the current endogenous processing moment,

and what the nature of the relationship between those two oscillation timescales is. Third, the cue integration model, which is built on cue reliabilities, or the representations of the probabilistic relationship between a given cue and an upcoming representation, predicts that there should be some neural signal that is related to the reliability of each level of representation as a cue to the next.

At least two fundamental problems seem to endure for a cue-based model. First, a persistent challenge is understanding why processing similar representations before or after the onset of a target representation is sometimes facilitatory (resulting in priming) and at other times inhibitory (resulting in interference). Is firing asynchrony somehow underlying the spectrum of priming and interference? Second, how might long-distance structural relationships, syntactic domains, and scope be encoded in a cue-based direct-access system (see Kush, 2013 for a discussion of c-command)? How does the parser "know where it is" to carry out these computations?

Summary

I have argued that any model of language computation must answer both *how* and *why* questions, and that the ideal model should be a fusion of mechanistic and probabilistic elements. I have sketched a framework for language processing based on the psychophysiological mechanism of cue integration. The cue integration framework asserts a mechanistic psychological operation over probabilistic representations, which are represented by neural population codes that are flexibly combined using two simple canonical neural computations: summation and normalization. Together these operations comprise the cue integration mechanism. By restricting computation to canonical neural mechanisms, cue integration may be able to form a linking hypothesis between psycholinguistic, computational, and neurobiological theories of language.

The heart of this mechanistic claim is that the relationship between a given level and the next level of representation (between cue and "target") is probabilistic, and that, in turn, this uncertainty forms a vital aspect of representation, incorporated via the cue integration mechanism of normalization. The main representational hypothesis of cue integration is that every level of representation is a cue to higher levels of representation, resulting in a cascaded architecture where activation can spread before processing of the current set of features completes. Crucially, cue integration can take as its input an endogenously stimulated representation and output another representational state, allowing all levels of linguistic input to be extracted from sensory input. While cues determine which representation is activated, cue reliabilities determine the strength of the evidence for a particular representation and thus how good of a model of the world the system has. Reliabilities reflect local processing context as well as global knowledge. Cue integration accounts for processing variables either by modulating the information expressed in the cue reliabilities, or by modulating the circuit of representations, in other words, changing the order of cue computations.

To close, the main criticism laid out in the first part of this article can of course be applied to the cue integration hypothesis: a central challenge for the cue integration model is to achieve a parsimony of cues, reliabilities, and population codes while preserving explanatory satisfaction.

AUTHOR CONTRIBUTIONS

The author confirms being the sole author and contributor to this work and has approved it for publication.

ACKNOWLEDGMENTS

I was supported by ESRC Future Research Leaders fellowship and grant ES/K009095/1. I thank Mante Nieuwland for comments on an earlier draft of this work. I thank Aine Ito for assistance in compiling the references for this article. All errors, inaccuracies and vagueness are my own.

REFERENCES

Altmann, G. T. M., and Kamide, Y. (1999). Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition* 73, 247–264. doi: 10.1016/S0010-0277(99)00059-1

Anderson, J. R. (1983). A spreading activation theory of memory. *J. Verbal Learn. Verbal Behav.* 22, 261–295. doi: 10.1016/S0022-5371(83)90201-3

Anderson, M. C., and Neely, J. H. (1996). Interference and inhibition in memory retrieval. *Memory* 22, 237–311.

Averbeck, B. B., Latham, P. E., and Pouget, A. (2006). Neural correlations, population coding and computation. *Nat. Rev. Neurosci.* 7, 358–366. doi: 10.1038/nrn1888

Bates, E., and MacWhinney, B. (1987). "Competition, variation, and language learning," in *Mechanisms of Language Acquisition*, ed B. MacWhinney (Hillsdale, NJ: Lawrence Erlbaum Associates), 157–193.

Benedict, H. (1979). Early lexical development: comprehension and production. *J. Child Lang.* 6, 183–200. doi: 10.1017/S0305000900002245

Bornkessel-Schlesewsky, I., Kretzschmar, F., Tune, S., Wang, L., Genç, S., Philipp, M., et al. (2011). Think globally: cross-linguistic variation in electrophysiological activity during sentence comprehension. *Brain Lang.* 117, 133–152. doi: 10.1016/j.bandl.2010.09.010

Bornkessel-schlesewsky, I., Schlesewsky, M., Small, S. L., and Rauschecker, J. P. (2015). Neurobiological roots of language in primate audition: common computational properties. *Trends Cogn. Sci.* 19, 1–9. doi: 10.1016/j.tics.2014.12.008

Bowers, J. S., and Davis, C. J. (2012). Bayesian just-so stories in psychology and neuroscience. *Psychol. Bull.* 138, 389–414. doi: 10.1037/a0026450

Bülthoff, H., and Yuille, A. (1996). "A Bayesian framework for the integration of visual modules," in *Attention Performance XVI: Information Integration in Perception and Communication,* eds T. Inui and J. L. McClelland (Cambridge, MA: MIT Press), 49–70.

Carandini, M., and Heeger, D. (2012). Normalization as a canonical neural computation. *Nat. Rev. Neurosci.* 13, 51–62. doi: 10.1038/nrn3136

Charniak, E. (1996). *Statistical Language Learning.* Cambridge: MIT Press.

Clark, S. E., and Gronlund, S. D. (1996). Global matching models of recognition memory: how the models match the data. *Psychon. Bull. Rev.* 3, 37–60. doi: 10.3758/BF03210740

Clifton C. Jr., and Frazier, L. (1989). "Comprehending sentences with long-distance dependencies," in *Linguistic Structure in Language Processing*, eds G. N. Carlson and M. K. Tanenhaus (Dordrecht: Springer), 273–317.

Davidson, D. J., and Martin, A. E. (2013). Modeling accuracy as a function of response time with the generalized linear mixed effects model. *Acta Psychol.* 144, 83–96. doi: 10.1016/j.actpsy.2013.04.016

DeLong, K. A., Urbach, T. P., and Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nat. Neurosci.* 8, 1117–1121. doi: 10.1038/nn1504

Deneve, S., Latham, P. E., and Pouget, A. (2001). Efficient computation and cue integration with noisy population codes. *Nat. Neurosci.* 4, 826–831. doi: 10.1038/90541

Dillon, B., Mishler, A., Sloggett, S., and Phillips, C. (2013). Contrasting intrusion profiles for agreement and anaphora: experimental and modeling evidence. *J. Mem. Lang.* 69, 85–103. doi: 10.1016/j.jml.2013.04.003

Doelling, K. B., Arnal, L. H., Ghitza, O., and Poeppel, D. (2014). Acoustic landmarks drive delta-theta oscillations to enable speech comprehension by facilitating perceptual parsing. *Neuroimage* 85, 761–768. doi: 10.1016/j.neuroimage.2013.06.035

Doumas, L. A. A., and Hummel, J. E. (2005). "Modeling human mental representations: what works, what doesn't, and why," in *The Cambridge Handbook of Thinking and Reasoning*, eds K. J. Holyoak and R. G. Morrison (Cambridge, UK: Cambridge University Press), 73–91.

Doumas, L. A. A., Hummel, J. E., and Sandhofer, C. M. (2008). A theory of the discovery and predication of relational concepts. *Psychol. Rev.* 115, 1–43. doi: 10.1037/0033-295X.115.1.1

Doupe, A. J., and Kuhl, P. K. (1999). Birdsong and human speech: common themes and mechanisms. *Annu. Rev. Neurosci.* 22, 567–631. doi: 10.1146/annurev.neuro.22.1.567

Ernst, M. O., and Bülthoff, H. H. (2004). Merging the senses into a robust percept. *Trends Cogn. Sci.* 8, 162–169. doi: 10.1016/j.tics.2004.02.002

Fedorenko, E., Duncan, J., and Kanwisher, N. (2012). Language-selective and domain-general regions lie side by side within Broca's area. *Curr. Biol.* 22, 2059–2062. doi: 10.1016/j.cub.2012.09.011

Ferreira, F., and Henderson, J. M. (1991). Recovery from misanalyses of garden-path sentences. *J. Mem. Lang.* 30, 725–745. doi: 10.1016/0749-596X(91)90034-H

Fetsch, C. R., DeAngelis, G. C., and Angelaki, D. E. (2013). Bridging the gap between theories of sensory cue integration and the physiology of multisensory neurons. *Nat. Rev. Neurosci.* 14, 429–442. doi: 10.1038/nrn3503

Fetsch, C. R., Pouget, A., DeAngelis, G. C., and Angelaki, D. E. (2012). Neural correlates of reliability-based cue weighting during multisensory integration. *Nat. Neurosci.* 15, 146–154. doi: 10.1038/nn.2983

Fodor, J. A. (1983). *The Modularity of Mind*. Cambridge: MIT Press.

Fodor, J. D., and Inoue, A. (1998). "Attach anyway," in *Reanalysis in Sentence Processing* (Netherlands: Springer), 101–141.

Foraker, S., and McElree, B. (2007). The role of prominence in pronoun resolution: active versus passive representations. *J. Mem. Lang.* 56, 357–383. doi: 10.1016/j.jml.2006.07.004

Frank, M. C., Goldwater, S., Griffiths, T. L., and Tenenbaum, J. B. (2010). Modeling human performance in statistical word segmentation. *Cognition* 117, 107–125. doi: 10.1016/j.cognition.2010.07.005

Frazier, L., and Clifton, C. (1996). *Construal*. Cambridge: MIT Press.

Frazier, L., and Fodor, J. D. (1978). The sausage machine: a new two-stage parsing model. *Cognition* 6, 291–325. doi: 10.1016/0010-0277(78)90002-1

Frazier, L., and Rayner, K. (1982). Making and correcting errors during sentence comprehension: eye movements in the analysis of structurally ambiguous sentences. *Cogn. Psychol.* 14, 178–210. doi: 10.1016/0010-0285(82)90008-1

Friederici, A. D. (2012). The cortical language circuit: from auditory perception to sentence comprehension. *Trends Cogn. Sci.* 16, 262–268. doi: 10.1016/j.tics.2012.04.001

Friederici, A. D., and Singer, W. (2015). Grounding language processing on basic neurophysiological principles. *Trends Cogn. Sci.* 19, 1–10. doi: 10.1016/j.tics.2015.03.012

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* 11, 127–138. doi: 10.1038/nrn2787

Gambi, C., and Pickering, M. J. (2013). Prediction and imitation in speech. *Front. Psychol.* 4:340. doi: 10.3389/fpsyg.2013.00340

Gervain, J., and Mehler, J. (2010). Speech perception and language acquisition in the first year of life. *Annu. Rev. Psychol.* 61, 191–218. doi: 10.1146/annurev.psych.093008.100408

Gibson, E. (2000). "The dependency locality theory: a distance-based theory of linguistic complexity," in *Image, Language, Brain,* eds Y. Miyashita, A. Marantz, and W. O'Neil (Cambridge, MA: MIT Press), 95–126.

Gibson, J. J. (1966). *The Senses Considered as Perceptual Systems*. Boston, MA: Houghton Mifflin.

Gillund, G., and Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychol. Rev.* 91, 1–67. doi: 10.1037/0033-295X.91.1.1

Giraud, A. L., and Poeppel, D. (2012). Cortical oscillations and speech processing: emerging computational principles and operations. *Nat. Neurosci.* 15, 511–517. doi: 10.1038/nn.3063

Goldwater, S., Griffiths, T. L., and Johnson, M. (2009). A Bayesian framework for word segmentation: exploring the effects of context. *Cognition* 112, 21–54. doi: 10.1016/j.cognition.2009.03.008

Golumbic, E. Z., Cogan, G. B., Schroeder, C. E., and Poeppel, D. (2013). Visual input enhances selective speech envelope tracking in auditory cortex at a "cocktail party." *J. Neurosci.* 33, 1417–1426. doi: 10.1523/JNEUROSCI.3675-12.2013

Griffiths, T. L., Chater, N., Norris, D., and Pouget, A. (2012). How the Bayesians got their beliefs (and what those beliefs actually are): comment on Bowers and Davis. *Psychol. Bull.* 138, 415–422. doi: 10.1037/a0026884

Hagoort, P. (2005). On Broca, brain, and binding: a new framework. *Trends Cogn. Sci.* 9, 416–423. doi: 10.1016/j.tics.2005.07.004

Hagoort, P. (2013). MUC (memory, unification, control) and beyond. *Front. Psychol.* 4:416. doi: 10.3389/fpsyg.2013.00416

Hagoort, P., and Indefrey, P. (2014). The neurobiology of language beyond single words. *Annu. Rev. Neurosci.* 37, 347–362. doi: 10.1146/annurev-neuro-071013-013847

Hale, J. (2001). "A probabilistic earley parser as a psycholinguistic model," in *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies* (Stroudsburg, PA: Association for Computational Linguistics), 1–8.

Harris, Z. S. (1954). Distributional structure. *Word* 10, 146–162. doi: 10.1080/00437956.1954.11659520

Hauser, M. D., Chomsky, N., and Fitch, W. T. (2002). The faculty of language: what is it, who has it, and how did it evolve? *Science* 298, 1569–1579. doi: 10.1126/science.298.5598.1569

Hickok, G., and Poeppel, D. (2004). Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. *Cognition* 92, 67–99. doi: 10.1016/j.cognition.2003.10.011

Hickok, G., and Poeppel, D. (2007). The cortical organization of speech processing. *Nat. Rev. Neurosci.* 8, 393–402. doi: 10.1038/nrn2113

Hofmeister, P., and Vasishth, S. (2014). Distinctiveness and encoding effects in online sentence comprehension. *Front. Psychol.* 5:1237. doi: 10.3389/fpsyg.2014.01237

Huettig, F. (2015). Four central questions about prediction in language processing. *Brain Res.* 1626, 118–135. doi: 10.1016/j.brainres.2015.02.014

Huettig, F., and Mani, N. (2016). Is prediction necessary to understand language? Probably not. *Lang. Cogn. Neurosci.* 31, 19–31. doi: 10.1080/23273798.2015.1072223

Jackendoff, R. (2002). "English particle constructions, the lexicon, and the autonomy of syntax," in *Verb-Particle Explorations,* ed N. Dehé (Berlin: Walter de Gruyter), 67–94.

Jackendoff, R., and Pinker, S. (2005). The nature of the language faculty and its implications for evolution of language (Reply to Fitch, Hauser, and Chomsky). *Cognition* 97, 211–225. doi: 10.1016/j.cognition.2005.04.006

Jacobs, R. A. (2002). What determines visual cue reliability? *Trends Cogn. Sci.* 6, 345–350. doi: 10.1016/S1364-6613(02)01948-4

Jacobson, P. (1999). Towards a variable-free semantics. *Linguist. Philos.* 22, 117–184. doi: 10.1023/A:1005464228727

Jaeger, T. F., and Snider, N. E. (2013). Alignment as a consequence of expectation adaptation: syntactic priming is affected by the prime's prediction error given both prior and recent experience. *Cognition* 127, 57–83. doi: 10.1016/j.cognition.2012.10.013

Jäger, L. A., Engelmann, F., and Vasishth, S. (2015). Retrieval interference in reflexive processing: experimental evidence from Mandarin, and computational modeling. *Front. Psychol.* 6:617. doi: 10.3389/fpsyg.2015.00617

Jones, M., and Love, B. C. (2011). Bayesian fundamentalism or enlightenment? *On the explanatory status and theoretical contributions of Bayesian models of cognition. Behav. Brain Sci.* 34, 169–188. doi: 10.1017/S0140525X10003134

Just, M. A., and Carpenter, P. A. (1992). A capacity theory of comprehension: individual differences in working memory. *Psychol. Rev.* 99, 122–149. doi: 10.1037/0033-295X.99.1.122

Knops, A., Thirion, B., Hubbard, E. M., Michel, V., and Dehaene, S. (2009). Recruitment of an area involved in eye movements during mental arithmetic. *Science* 324, 1583–1585. doi: 10.1126/science.1171599

Kuhl, P. K. (2004). Early language acquisition: cracking the speech code. *Nat. Rev. Neurosci.* 5, 831–843. doi: 10.1038/nrn1533

Kush, D. (2013). *Respecting Relations: Memory Access and Antecedent Retrieval in Incremental Sentence Processing.* Dissertation, University of Maryland.

Kush, D., Lidz, J., and Phillips, C. (2015). Relation-sensitive retrieval: evidence from bound variable pronouns. *J. Mem. Lang.* 82, 18–40. doi: 10.1016/j.jml.2015.02.003

Laufer, B. (1998). The development of passive and active vocabulary in a second language: same or different? *Appl. Linguist.* 19, 255–271.

Legge, G. E., Klitz, T. S., and Tjan, B. S. (1997). Mr. *Chips: an ideal-observer model of reading. Psychol. Rev.* 104, 524–553.

Levelt, W. J. M. (1999). Models of word production. *Trends Cogn. Sci.* 3, 223–232. doi: 10.1016/S1364-6613(99)01319-4

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition* 106, 1126–1177. doi: 10.1016/j.cognition.2007.05.006

Levy, R. (2013). "Memory and surprisal in human sentence comprehension," in *Sentence Processing,* ed R. P. G. Van Gompel (Hove: Psychology Press), 78–114.

Lewis, R. L., and Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cogn. Sci.* 29, 375–419. doi: 10.1207/s15516709cog0000_25

Lewis, R. L., Vasishth, S., and Van Dyke, J. A. (2006). Computational principles of working memory in sentence comprehension. *Trends Cogn. Sci.* 10, 447–454. doi: 10.1016/j.tics.2006.08.007

Luo, H., and Poeppel, D. (2007). Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron* 54, 1001–1010. doi: 10.1016/j.neuron.2007.06.004

Ma, W. J. (2012). Organizing probabilistic models of perception. *Trends Cogn. Sci.* 16, 511–518. doi: 10.1016/j.tics.2012.08.010

MacDonald, M. C. (2013). How language production shapes language form and comprehension. *Front. Psychol.* 4:226. doi: 10.3389/fpsyg.2013.00226

MacDonald, M. C., and Christiansen, M. H. (2002). Reassessing working memory: comment on Just and Carpenter (1992) and Waters and Caplan (1996). *Psychol. Rev. 109*, 35–54; discussion 55–74. doi: 10.1037/0033-295x.109.1.35

MacDonald, M. C., Pearlmutter, N. J., and Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychol. Rev.* 101, 676–703. doi: 10.1037/0033-295X.101.4.676

Maloney, L. T., and Zhang, H. (2010). Decision-theoretic models of visual perception and action. *Vision Res.* 50, 2362–2374. doi: 10.1016/j.visres.2010.09.031

Markman, A. B., and Otto, A. R. (2011). Cognitive systems optimize energy rather than information. *Behav. Brain Sci.* 34:207. doi: 10.1017/S0140525X11000355

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco, CA: W. H. Freeman.

Marslen-Wilson, W. D., and Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cogn. Psychol.* 10, 29–63. doi: 10.1016/0010-0285(78)90018-X

Martin, A. E., and McElree, B. (2008). A content-addressable pointer mechanism underlies comprehension of verb-phrase ellipsis. *J. Mem. Lang.* 58, 879–906. doi: 10.1016/j.jml.2007.06.010

Martin, A. E., and McElree, B. (2009). Memory operations that support language comprehension: evidence from verb-phrase ellipsis. *J. Exp. Psychol. Learn. Mem Cogn.* 35, 1231–1239. doi: 10.1037/a0016271

Martin, A. E., and McElree, B. (2011). Direct-access retrieval during sentence comprehension: evidence from Sluicing. *J. Mem. Lang.* 64, 327–343. doi: 10.1016/j.jml.2010.12.006

Martin, A. E., Nieuwland, M. S., and Carreiras, M. (2012). Event-related brain potentials index cue-based retrieval interference during sentence comprehension. *Neuroimage* 59, 1859–1869. doi: 10.1016/j.neuroimage.2011.08.057

Martin, A. E., Nieuwland, M. S., and Carreiras, M. (2014). Agreement attraction during comprehension of grammatical sentences: ERP evidence from ellipsis. *Brain Lang.* 135, 42–51. doi: 10.1016/j.bandl.2014.05.001

McElree, B. (2000). Sentence comprehension is mediated by content-addressable memory structures. *J. Psycholinguist Res.* 29, 111–123. doi: 10.1023/A:1005184709695

McElree, B. (2006). Accessing recent events. *Psychol. Learn. Motiv. Adv. Res. Theory* 155–200. doi: 10.1016/S0079-7421(06)46005-9

McElree, B., and Dosher, B. A. (1989). Serial position and set size in short-term memory: the time course of recognition. *J. Exp. Psychol. Gen.* 118, 346–373. doi: 10.1037/0096-3445.118.4.346

McElree, B., and Dosher, B. A. (1993). Serial retrieval processes in the recovery of order information. *J. Exp. Psychol. Gen.* 122, 291–315. doi: 10.1037/0096-3445.122.3.291

McElree, B., Foraker, S., and Dyer, L. (2003). Memory structures that subserve sentence comprehension. *J. Mem. Lang.* 48, 67–91. doi: 10.1016/S0749-596X(02)00515-6

McMurray, B., and Jongman, A. (2011). What information is necessary for speech categorization? *Harnessing variability in the speech signal by integrating cues computed relative to expectations. Psychol. Rev.* 118, 219–246. doi: 10.1037/a0022325

Menenti, L., Gierhan, S. M. E., Segaert, K., and Hagoort, P. (2011). Shared language: overlap and segregation of the neuronal infrastructure for speaking and listening revealed by functional MRI. *Psychol. Sci.* 22, 1173–1182. doi: 10.1177/0956797611418347

Miller, G. A., and Chomsky, N. (1963). "Finitary models of language users," in *Handbook of Mathematical Psychology,* ed D. Luce (New York, NY: John Wiley & Sons), 412–419.

Nairne, J. S. (2002). The myth of the encoding-retrieval match. *Memory* 10, 389–395. doi: 10.1080/09658210244000216

Nieuwland, M. S., Martin, A. E., and Carreiras, M. (2012). Brain regions that process case: evidence from basque. *Hum. Brain Mapp.* 33, 2509–2520. doi: 10.1002/hbm.21377

Norris, D. (2006). The Bayesian reader: explaining word recognition as an optimal Bayesian decision process. *Psychol. Rev.* 113, 327–357. doi: 10.1037/0033-295X.113.2.327

Norris, D. (2013). Models of visual word recognition. *Trends Cogn. Sci.* 17, 517–524. doi: 10.1016/j.tics.2013.08.003

Norris, D., and Kinoshita, S. (2012). Reading through a noisy channel: why there's nothing special about the perception of orthography. *Psychol. Rev.* 119, 517–545. doi: 10.1037/a0028450

Numminen, J., Salmelin, R., and Hari, R. (1999). Subject's own speech reduces reactivity of the human auditory cortex. *Neurosci. Lett.* 265, 119–122. doi: 10.1016/S0304-3940(99)00218-9

Öztekin, I., and McElree, B. (2007). Proactive interference slows recognition by eliminating fast assessments of familiarity. *J. Mem. Lang.* 57, 126–149. doi: 10.1016/j.jml.2006.08.011

Pickering, M. J., and Clark, A. (2014). Getting ahead: forward models and their place in cognitive architecture. *Trends Cogn. Sci.* 18, 451–456. doi: 10.1016/j.tics.2014.05.006

Pickering, M. J., and Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behav. Brain Sci. 27*, 169–190; discussion 190–226. doi: 10.1017/S0140525X04000056

Pickering, M. J., and Garrod, S. (2007). Do people use language production to make predictions during comprehension? *Trends Cogn. Sci.* 11, 105–110. doi: 10.1016/j.tics.2006.12.002

Pickering, M. J., and Garrod, S. (2013). An integrated theory of language production and comprehension. *Behav. Brain Sci.* 36, 329–347. doi: 10.1017/S0140525X12001495

Poeppel, D. (2014). The neuroanatomic and neurophysiological infrastructure for speech and language. *Curr. Opin. Neurobiol.* 28, 142–149. doi: 10.1016/j.conb.2014.07.005

Pouget, A., Pouget, A., Dayan, P., Dayan, P., Zemel, R., and Zemel, R. (2000). Information processing with population

codes. *Nat. Rev. Neurosci.* 1, 125–132. doi: 10.1038/35039062

Puchalla, J. L., Schneidman, E., Harris, R. A., and Berry, M. J. (2005). Redundancy in the population code of the retina. *Neuron* 46, 493–504. doi: 10.1016/j.neuron.2005.03.026

Rabagliati, H., and Bemis, D. K. (2013). Prediction is no panacea: the key to language is in the unexpected. *Behav. Brain Sci.* 36, 372–373. doi: 10.1017/S0140525X12002671

Rao, R. P., and Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* 2, 79–87. doi: 10.1038/4580

Rauschecker, J. P. (2012). Ventral and dorsal streams in the evolution of speech and language. *Front. Evol. Neurosci.* 4:7. doi: 10.3389/fnevo.2012.00007

Rauschecker, J. P., and Scott, S. K. (2009). Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. *Nat. Neurosci.* 12, 718–724. doi: 10.1038/nn.2331

Rauschecker, J. P., and Tian, B. (2000). Mechanisms and streams for processing of "what" and "where" in auditory cortex. *Proc. Natl. Acad. Sci. U.S.A.* 97, 11800–11806. doi: 10.1073/pnas.97.22.11800

Reed, A. V. (1973). Speed-accuracy trade-off in recognition memory. *Science* 181, 574–576.

Rehder, B. (2011). Taking the rationality out of probabilistic models. *Behav. Brain Sci.* 34, 210–211. doi: 10.1017/S0140525X11000422

Rothschild, G., Nelken, I., and Mizrahi, A. (2010). Functional organization and population dynamics in the mouse primary auditory cortex. *Nat. Neurosci.* 13, 353–360. doi: 10.1038/nn.2484

Saffran, J. R., Aslin, R. N., and Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science* 274, 1926–1928.

Samuel, A. G. (2001). Knowing a word affects the fundamental perception of the sounds within it. *Psychol. Sci.* 12, 348–351. doi: 10.1111/1467-9280.00364

Samuel, A. G., and Kraljic, T. (2009). Perceptual learning for speech. *Atten. Percept. Psychophys.* 71, 1207–1218. doi: 10.3758/APP.71.6.1207

Schneidman, E., Bialek, W., and Berry, M. J. (2003). Synergy, redundancy, and independence in population codes. *J. Neurosci.* 23, 11539–11553. doi: 10.1523/JNEUROSCI.5319-04.2005

Singer, W. (2013). Cortical dynamics revisited. *Trends Cogn. Sci.* 17, 616–626. doi: 10.1016/j.tics.2013.09.006

Smith, N. J., and Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition* 128, 302–319. doi: 10.1016/j.cognition.2013.02.013

Stecker, G. C., Harrington, I. A., and Middlebrooks, J. C. (2005). Location coding by opponent neural populations in the auditory cortex. *PLoS Biol.* 3:e78. doi: 10.1371/journal.pbio.0030078

Steedman, M. (1996). *Surface Structure and Interpretation.* Cambridge: MIT press.

Steedman, M. (2000). *The Syntactic Process*. Cambridge: MIT press.

Stephens, G. J., Silbert, L. J., and Hasson, U. (2010). Speaker-listener neural coupling underlies successful communication. *Proc. Natl. Acad. Sci. U.S.A.* 107, 14425–14430. doi: 10.1073/pnas.1008662107

Swinney, D. A. (1979). Lexical access during sentence comprehension: (Re)consideration of context effects. *J. Verbal Learn. Verbal Behav.* 18, 645–659. doi: 10.1016/S0022-5371(79)90355-4

Szabolcsi, A. (1989). "Noun phrases and clauses: is DP analogous to IP or CP," in *Proceedings of the Colloquium on Noun Phrase Structure* (Manchester).

Szabolcsi, A. (2003). "Binding on the fly: cross-sentential anaphora in variable-free semantics," in *Resource-Sensitivity, Binding and Anaphora*, eds G.-J. M. Kruijff and R. T. Oehrle (Dordrecht: Springer), 215–227.

Tabor, W., Galantucci, B., and Richardson, D. (2004). Effects of merely local syntactic coherence on sentence processing. *J. Mem. Lang.* 50, 355–370. doi: 10.1016/j.jml.2004.01.001

Tabor, W., and Hutchins, S. (2004). Evidence for self-organized sentence processing: digging-in effects. *J. Exp. Psychol. Learn. Mem. Cogn.* 30, 431–450. doi: 10.1037/0278-7393.30.2.431

Tabor, W., Juliano, C., and Tanenhaus, M. (1997). Parsing in a dynamical system: an attractor-based account of the interaction of lexical and structural constraints in sentence processing. *Lang. Cogn. Process.* 12, 211–271. doi: 10.1080/016909697386853

Tanner, D., Nicol, J., and Brehm, L. (2014). The time-course of feature interference in agreement comprehension: multiple mechanisms and asymmetrical attraction. *J. Mem. Lang.* 76, 195–215. doi: 10.1016/j.jml.2014.07.003

Toscano, J. C., and McMurray, B. (2010). Cue integration with categories: weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cogn. Sci.* 34, 434–464. doi: 10.1111/j.1551-6709.2009.01077.x

Townsend, J. T., and Ashby, G. F. (1983). *Stochastic Modeling of Elementary Psychological Processes*. (Cambridge: Cambridge University Press).

Treisman, A. (1999). Solutions to the binding problem: progress through controversy and convergence. *Neuron* 24, 105–110. doi: 10.1016/S0896-6273(00)80826-0

Trommershauser, J., Kording, K., and Landy, M. S. (eds.). (2011). *Sensory Cue Integration.* New York, NY: Oxford University Press.

Van Berkum, J. J. A., Brown, C. M., Zwitserlood, P., Kooijman, V., and Hagoort, P. (2005). Anticipating upcoming words in discourse: evidence from ERPs and reading times. *J. Exp. Psychol. Learn. Mem. Cogn.* 31, 443–467. doi: 10.1037/0278-7393.31.3.443

Van Dyke, J. A. (2007). Interference effects from grammatically unavailable constituents during sentence processing. *J. Exp. Psychol. Learn. Mem. Cogn.* 33, 407–430. doi: 10.1037/0278-7393.33.2.407

Van Dyke, J. A., and Johns, C. L. (2012). Memory interference as a determinant of language comprehension. *Linguist. Lang. Compass* 6, 193–211. doi: 10.1002/lnc3.330

Van Dyke, J. A., and McElree, B. (2011). Cue-dependent interference in comprehension. *J. Mem. Lang.* 65, 247–263. doi: 10.1016/j.jml.2011.05.002

Vosse, T., and Kempen, G. (2000). Syntactic structure assembly in human parsing: a computational model based on competitive inhibition and a lexicalist grammar. *Cognition* 75, 105–143. doi: 10.1016/S0010-0277(00)00063-9

Wagers, M. W., Lau, E. F., and Phillips, C. (2009). Agreement attraction in comprehension: representations and processes. *J. Mem. Lang.* 61, 206–237. doi: 10.1016/j.jml.2009.04.002

Wickelgren, W. A. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta Psychol.* 41, 67–85. doi: 10.1016/0001-6918(77)90012-9

Xiang, M., Dillon, B., and Phillips, C. (2009). Illusory licensing effects across dependency types: ERP evidence. *Brain Lang.* 108, 40–55. doi: 10.1016/j.bandl.2008.10.002

Xing, J., and Heeger, D. J. (2000). Center-surround interactions in foveal and peripheral vision. *Vision Res.* 40, 3065–3072. doi: 10.1016/S0042-6989(00)00152-8

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2016 Martin. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 07 May 2015
doi: 10.3389/fpsyg.2015.00590

Grammatical number processing and anticipatory eye movements are not tightly coordinated in English spoken language comprehension

*Brian Riordan[1], Melody Dye[2] and Michael N. Jones[2]**

[1] Aptima, Inc., Fairborn, OH, USA, [2] Department of Psychological and Brain Sciences, Indiana University, Bloomington, IN, USA

Recent studies of eye movements in world-situated language comprehension have demonstrated that rapid processing of morphosyntactic information – e.g., grammatical gender and number marking – can produce anticipatory eye movements to referents in the visual scene. We investigated how type of morphosyntactic information and the goals of language users in comprehension affected eye movements, focusing on the processing of grammatical number morphology in English-speaking adults. Participants' eye movements were recorded as they listened to simple English declarative *(There are the lions.)* and interrogative *(Where are the lions?)* sentences. In Experiment 1, no differences were observed in speed to fixate target referents when grammatical number information was informative relative to when it was not. The same result was obtained in a speeded task (Experiment 2) and in a task using mixed sentence types (Experiment 3). We conclude that grammatical number processing in English and eye movements to potential referents are not tightly coordinated. These results suggest limits on the role of predictive eye movements in concurrent linguistic and scene processing. We discuss how these results can inform and constrain predictive approaches to language processing.

Keywords: grammatical number, eye movements, sentence comprehension, spoken word recognition, visual world paradigm

OPEN ACCESS

Edited by:
Matthew Wagers, University of California, Santa Cruz, USA

Reviewed by:
Cynthia Lukyanenko, The Pennsylvania State University, USA
Nikole Patson, Ohio State University, USA

****Correspondence:***
Michael N. Jones, Department of Psychological and Brain Sciences, Indiana University, 1101 East 10th Street, Bloomington, IN 47404, USA
jonesmn@indiana.edu

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: *02 January 2015*
Accepted: *21 April 2015*
Published: *07 May 2015*

Citation:
Riordan B, Dye M and Jones MN (2015) Grammatical number processing and anticipatory eye movements are not tightly coordinated in English spoken language comprehension. Front. Psychol. 6:590. doi: 10.3389/fpsyg.2015.00590

Introduction

In the study of spoken language comprehension, the discovery that language processing is closely coordinated with patterns of eye movements represents a major advance for the discipline (Tanenhaus and Trueswell, 2006). Not only does the the visual context influence how the unfolding linguistic input is structured (Tanenhaus et al., 1995), but fixations to referents in the visual scene have been shown to reflect the fine-grained time course of spoken word recognition (e.g., Magnuson et al., 2007).

When processing linguistic and visual input simultaneously, listeners rapidly integrate across information streams, making anticipatory eye movements to likely referents. For example, Altmann and Kamide (1999) demonstrated that when listeners encounter verbs such as *eat*, they shift their visual attention to edible objects. Kamide et al. (2003) further demonstrated that listeners can integrate morphosyntactic and semantic information at the verb to drive eye movements to

likely referents. Other work has demonstrated anticipatory looking behavior during thematic role assignment (Dahan and Tanenhaus, 2004; Knoeferle and Crocker, 2006).

These findings are consistent with a host of related experimental results suggesting that, like other aspects of human cognition, language comprehension and production are incremental, predictive processes. In making predictive inferences about upcoming speech or text, communicators draw on multiple sources of linguistic information, ranging over lexical, semantic, and discourse levels (for reviews, see Pickering and Garrod, 2007; Ramscar et al., 2010). This has been demonstrated empirically in a number of ways. For instance, in reading, more predictable items are processed faster and more efficiently (McDonald and Shillcock, 2003; Hare et al., 2009), and in speech production tasks, such items are uttered more quickly, often in a reduced form (Gahl et al., 2012), with fewer disfluencies (Arnold et al., 2007). Eye movement studies complement these traditional experimental domains, furnishing a rich picture of how various linguistic factors conspire to affect processing in real time (Huettig et al., 2011).

Grammatical Gender

One important question that the visual world paradigm has begun to answer, is how syntactic agreement patterns assist comprehension processes. Agreement is thought to establish local and global coherence by linking temporally separated elements in discourse. However, precisely how it accomplishes this is an active area of research. A key line of enquiry concerns the influence of grammatical gender on lexical access. Gender systems are obligatory morphological systems found in many languages, which group nouns into a small number of mutually exclusive classes, and mark neighboring words – such as articles and adjectives – for agreement. In Romance languages, like French and Spanish, nouns are typically divided into two separate classes: masculine and feminine. Other major languages, such as Russian and German, add a third neuter category, and more are possible; Swahili has six (Corbett, 1991).

While historically gender has been viewed as an arbitrary or superfluous system (see Kilarski, 2007 for a review), there is an accumulating body of evidence to indicate otherwise. For one, while gender systems are not always semantically transparent, neither are they opaque to their speakers; there are typically multiple, converging linguistic cues to class membership (Frigo and McDonald, 1998). Further, gender systems may confer distinct advantages for native speakers. A leading hypothesis is that gender information reduces the lexical search space, delimiting the set of nouns to gender-consistent possibilities (but see Friederici and Jacobsen, 1999 for alternative proposals). On this view, speakers use gender to guide lexical access, helping them better predict upcoming nouns in discourse, as well as likely referents in the visual scene. This suggests that gender should both facilitate processing (when the marker is consistent with a following noun) and inhibit it (when the marker mismatches). Supporting evidence comes from a variety of sources, including lexical decision (Grosjean et al., 1994), naming times (Schriefers, 1993), word repetition (Bates et al., 1996), artificial grammar learning (Arnon and Ramscar, 2012), and ERP, where gender agreement violations have been found to produce neural error responses to the mismatch (Wicha et al., 2004; Van Berkum et al., 2005).

Yet perhaps the strongest support for the 'limited search' hypothesis comes from tasks that illuminate the time course of spoken language comprehension. In auditory gating paradigms, subjects hear short sequences in which a word fragment appears, and are asked to produce the target word. In a study of native French speakers, Grosjean et al. (1994) found that when gender information was provided, subjects correctly identified the target at shorter durations, and with greater confidence. More importantly, an inspection of subject errors revealed that gender information not only significantly reduced the number of misidentifications (both in terms of types and tokens), but also limited errors to gender-consistent candidates. Indeed, "in the presence of gender marking, no word candidate ever (had) the wrong gender" (Grosjean et al., 1994; p. 594). Similarly, in tip-of-the-tongue (TOT) states, Italian subjects can reliably guess the gender of the noun they are trying to retrieve, even when they cannot produce it (Vigliocco et al., 1997).

These findings are paralleled in studies of visual search. Dahan et al. (2000) investigated how gender-marked definite articles influenced the looking behavior of French-speaking participants. Subjects viewed a visual display with four possible referents, and heard instructions such as *Cliquez sur le bouton* (*Click on the$_{masc}$ button*). When gender information was provided at the determiner, listeners rapidly shifted their attention to gender-consistent referents, ignoring potential phonological competitors. Lew-Williams and Fernald (2007) reported a comparable result for Spanish-speakers, finding that both children and adults are faster to orient to the correct referent on trials when nouns of different genders are displayed than on trials showing nouns of the same gender (see also Weber and Paris, 2004; van Heugten and Shi, 2009).

Taken together, these results support the conclusion that grammatical gender does not merely prime lexical candidates, but rather restricts the space of subsequent possibility. However, the studies reviewed here focus exclusively on several closely related Romance languages. There is also evidence to suggest that the function and strength of gender, as a morphosyntactic cue, may vary significantly by language (see, e.g., Miozzo and Caramazza, 1999). This is quite clearly the case when it comes to grammatical number.

Grammatical Number

Grammatical number offers another promising domain of investigation for eye movement research. If gender is a widespread feature of the world's languages, number is nearly universal. In the simplest number systems, a noun's morphological form is modified to represent the numerosity of its referents, indicating whether the noun references a single entity or multiple entities, and neighboring words are marked for agreement (Corbett, 2000). In English, number is obligatory, and typically indicated by the presence or absence of a terminal sibilant *+s* (*cat*/*cats*), with several phonologically related families of irregulars (*mouse*/*mice*). A theoretical distinction is often drawn

between *count nouns*, which alternate freely between singular and plural forms, and *mass nouns*, which are treated as a single, indivisible set, regardless of numerosity. Compare, for instance, the usage of the semantically related pairs *noodles*$_{\text{count}}$/*pasta*$_{\text{mass}}$, *colds*$_{\text{count}}$/*flu*$_{\text{mass}}$, and *jobs*$_{\text{count}}$/*work*$_{\text{mass}}$.

As with grammatical gender, number information may be a potentially useful resource for predicting upcoming referents. Listeners appear to process grammatical number information quickly and automatically. Grammatical number violations are registered particularly rapidly, a conclusion that has been established through reading times (Wagers et al., 2009) and ERP (Pulvermüller and Shtyrov, 2003; Barber and Carreiras, 2005). Complementary results have been reported in TOT paradigms, where English-speakers have been found to reliably discriminate the appropriate sentential contexts for count nouns, even on failure to retrieve them (Vigliocco et al., 1999). Collectively, these findings imply that available agreement information scaffolds prediction of upcoming items in discourse.

If this is the case, simply hearing the string *Look, there are some*— might serve to restrict gaze to plural objects in a visual display. This is precisely what Kouider et al. (2006) found in a study of English-speaking children. On critical trials, toddlers saw pictures of novel objects on two screens; one picture depicted a single object and the other, multiple copies of the same object. Children heard sentences such as *Look, there are some blickets!* Beginning at 24 months, children were able to use the number marking on the copula and the indefinite article to launch anticipatory eye movements to the correct picture. Similar findings have been reported for French (Robertson et al., 2012). Complicating this picture, however, Johnson et al. (2005) report that in a picture selection task, English-speaking toddlers fail to use verb agreement marking as a cue to subject number (see Brandt-Kobele and Höhle, 2010 for a parallel finding in German).

Thus, despite some promising results, there is reason to suspect that grammatical number may not be as consistently informative about upcoming referents as grammatical gender. A variety of different theoretical accounts provide for different representations for gender and number (see discussion in Barber and Carreiras, 2005). One hypothesis is that whereas gender information is a property of the lexical item, stored in its lexical representation, number is is an independent morphological feature that combines with the stems of lexical items. These representational differences have processing consequences in models of lexical retrieval: gender information is retrieved with lexical access, while number information is involved only in a postlexical process of grammatical agreement as part of integration with the context. On this account, because grammatical number information does not directly activate lexical representations, processing of this information should only be weakly reflected in eye movements to referents in the visual scene.

Another source of difference may arise from number and gender's very different relations to semantics (Eberhard et al., 2005). Speaking broadly, a noun's number specification tends to be semantically motivated, reflecting the numerosity of the referent. By contrast, a noun's gender specification tends to be semantically arbitrary, with little obvious correspondence between the conceptual properties of the referent and its noun class, and substantial cross-linguistic variation. Thus, whereas number tends to be an extrinsic, inflectional feature that is highly responsive to semantics, gender tends to be intrinsic and non-inflectional, with comparatively limited interaction with semantics (see Vigliocco et al., 2005). This suggests that as a predictive cue, number may be less informative in languages in which semantic factors strongly bias agreement patterns.

For this reason, it is important consider the distributional facts of the language under study: namely, English. In number agreement in English, the mapping between inflection and semantics is highly context-dependent, and is difficult to capture with simple, easily generalizable rules (Huddleston and Pullum, 2002). To grasp this point, it is helpful to consider just how far the language departs from a highly simplified case, in which agreement is computed solely as a function of a referent's numerosity (singular/plural) and its semantic type (count/mass), and in which the semantic type distinction is clear-cut (e.g., mass nouns always refer to an undifferentiable whole).

The first complication is that, on inspection, there are certain systematic mismatches between syntax and semantics. For instance, mass nouns like *furniture* and *clothing* can be notionally plural while behaving like singulars (as when, e.g., there are multiple *pieces* of *furniture* or *articles* of *clothing* present), while pluralia tantum like *scissors* and *binoculars* can be notionally singular while behaving like plurals (as when there is a singular *pair* of *scissors* or *set* of *binoculars*). Nor is nominal inflection always a reliable guide to syntactic behavior, as evidenced by nouns whose meaning contravenes their marking, such as *news* (always singular), *police* (always plural), or *sheep* (which has the same singular and plural form).

Another wrinkle is that there is no straightforward way in which to tag nouns as countable, or not. While certain nouns fall on opposite ends of the *count*/*mass* spectrum, most nouns can behave in either way, depending on the semantic context (e.g., *I would like to buy a cake*/*I would like some more cake*). Further, countable nouns are not themselves a uniform class, and many show lexically specific preferences for (or restrictions on) the quantifiers they pair with. More broadly, item differences appear to be graded and distributional in kind, rather than rule-based and categorical (Baldwin and Bond, 2003). This suggests that agreement must be computed with reference to the entire noun phrase (NP), rather than simply the noun itself (Allan, 1980).

Finally, subject-verb agreement conventions are subject to variation both within and between speakers, and are closely influenced by semantics (Haskell and MacDonald, 2003; Eberhard et al., 2005). Singular collectives can take plural verbs (*the faculty are deliberating*/*neither of them are happy*) and plural quantities can take singular verbs (*ninety days is a long time*). In addition to these 'legal' alternations, agreement errors are common; speakers are especially prone to interference when the main verb is proximate to a noun with a different number than its head noun, as in *The key to the cabinets were missing* (Bock and Miller, 1991). In short, grammatical number in English is a highly complex system, in which agreement and marking conventions furnish, at best, an incomplete guide to the numerosity of the referent.

In the studies presented here, we sought to establish whether English-speaking adults make use of the partial information afforded by grammatical number to drive eye movements to likely referents, in contexts in which the predictive cue validity of number should be relatively weak. In online comprehension of both declarative and interrogative sentences, listeners first encountered grammatical number marking on the copula, in constructions such as *There are the cars* and *Where are the cars*? In addition, listeners heard sentences that incorporated multiple cues to number, such as *There are some cars*, in which the indefinite article was also marked.

Experiment 1

We recorded participants' eye movements as they listened to declarative and interrogative sentences. Following Lew-Williams and Fernald (2007), participants were exposed to two types of trials. On *same-number* trials, participants saw two pictures that each had the same number of object exemplars. On these trials, participants could not determine the target referent until the onset of the noun. On *different-number* trials, the two pictures differed in the number of exemplars depicted. On these trials, participants could use grammatical number information that preceded the noun to quickly orient toward the correct referent. If grammatical number information is rapidly exploited in sentence comprehension, participants should be faster to fixate the picture that matches the linguistic input on different-number trials than on same-number trials.

Method

Participants

Thirty native English speakers with normal or corrected-to normal vision participated for course credit.

Stimuli and Design

Noun targets were 16 object names with early age-of-acquisition. The words were divided into two sets of eight. Across participants, each set of eight words appeared in each condition. Within each set, no words shared the same initial phoneme. The noun targets were inserted in simple declarative and interrogative sentences. Sentences were of the form *There/Where [copula] [article] [noun]*.

Two conditions varied the number of grammatical number cues in the sentences. In the definite determiner condition, both declarative and interrogative sentences included the definite determiner *the*. In this condition, the grammatical number information was only available on the copula. In the indefinite determiner condition, all sentences included an indefinite determiner, *a* or *some*. Here, grammatical number information was available on both the copula and the indefinite determiner.

There were 64 total test trials in each condition (see **Table 1**). Half of the trials were *same-number* trials, and half were *different-number* trials. In addition, half of the trials were sentences with singular number, and half with plural number. Within each condition, the target referent appeared equally often in the left and right locations. Each participant was exposed to half of the total stimuli in each condition (32 trials per condition), and eight filler trials. Thus participants saw a total of 80 trials during the experiment.

Sentences were recorded by a female speaker using a natural speech rate. All sentences employed the uncontracted form of the copula. Across sentences, the mean duration of copulas was 152 ms (range = 100–225), the mean duration of determiners was 151 ms (range = 50–275), and the mean duration of nouns was 591 ms (range = 300–800 ms).

The visual stimuli were drawn from Rossion and Pourtois (2004). To form plural versions of each stimulus, four copies of each individual image were reduced in size and concatenated. The total surface area of the singular and plural images was identical. **Figure 1** depicts an example visual display for a different-number trial.

Procedure

Participants were instructed to click on the picture that was mentioned in the sentence (Weber and Paris, 2004). They were told to listen normally; no time constraints were imposed. As they listened, participants' eye movements were recorded using a desktop-mounted SR Research EyeLink eyetracker sampling at 1000 Hz. Each trial began with the presentation of a fixation dot for 750 ms. There was 2000 ms preview time before sentence onset. Using the fixation dot as a cursor, participants clicked on the picture that matched the sentence. The trial ended with

TABLE 1 | Composition of test trials in Experiment 1.

Condition		Trial type		Example
		Same-number	Different-number	
Definite	Singular	16	16	*There is the lion.*
	Plural	16	16	*There are the lions.*
Indefinite	Singular	16	16	*There is a lion.*
	Plural	16	16	*There are some lions.*

Across the definite and indefinite conditions, each participant was exposed to half of the test items.

FIGURE 1 | Example visual display from a different-number trial.

the mouse click. Each participant completed both the definite and indefinite conditions. Sentence order was randomized within condition, and the order of presentation of the conditions was counterbalanced across participants.

Analysis

The primary dependent variable was reaction time (RT) to initiate a saccade to the target referent (Lew-Williams and Fernald, 2007). We calculated RT as the latency of the first saccade or fixation that marked the start of an uninterrupted series of fixations on the target referent until the mouse click that ended the trial. RT was measured from copula onset.

Only trials that met the following conditions were included in the analysis. First, the participant must not have been fixating the target referent at the onset of the copula. Second, a saccade to or fixation on the target referent could not occur prior to 200 ms after the copula onset – approximately the earliest time a saccade could have been launched to the target referent after the copula onset (Altmann and Kamide, 2004). Third, RT must have occurred before 700 ms after the onset of the noun.

Results and Discussion

Figure 2 presents the time course of looking at each object in the display as the linguistic input unfolds in the definite condition. The curves represent the mean proportion of fixations to target objects on same-number trials versus different-number trials beginning with the start of the sentence. Participants shifted to the target object as the unfolding utterance allowed them to identify the correct picture. The trajectory of fixations is very similar across trial types, indicating that participants did not reliably use the grammatical number information encoded on the form of the copula to anticipate the target referent.

Figure 3 shows the time course of fixations for the two trial types in the indefinite condition. In this condition, too, the trajectory of fixations is similar across same-number and different-number trials. Participants did not make use of the two grammatical number cues preceding the noun – the copula and the indefinite article – to anticipate the correct referent.

These findings were confirmed with the RT analyses. Because sentence lengths varied with the type of copula (*is* vs. *are*) and the type of determiner (definite vs. indefinite, and within indefinite determiners, *a* vs. *some),* participants' processing of the grammatical number information is likely to have varied across sentence types. Therefore, we report separate RT analyses by sentence type. Mean RT was calculated both by-subjects (F_1) and by-items (F_2). **Table 2** presents the results of within-subjects ANOVAs for each comparison. Although there were trends toward faster RT on different-number vs. same-number trials, in no case were these differences reliable in the expected direction.

To explore the degree to which participants made anticipatory eye movements to the correct picture, we calculated the percentage of trials in which participants launched saccades to the target before they could process the noun (estimated as 200 ms after noun onset). Participants anticipated the target on only 35.1% of distracter-initial trials in the definite condition, and 39.6% of trials in the indefinite condition.

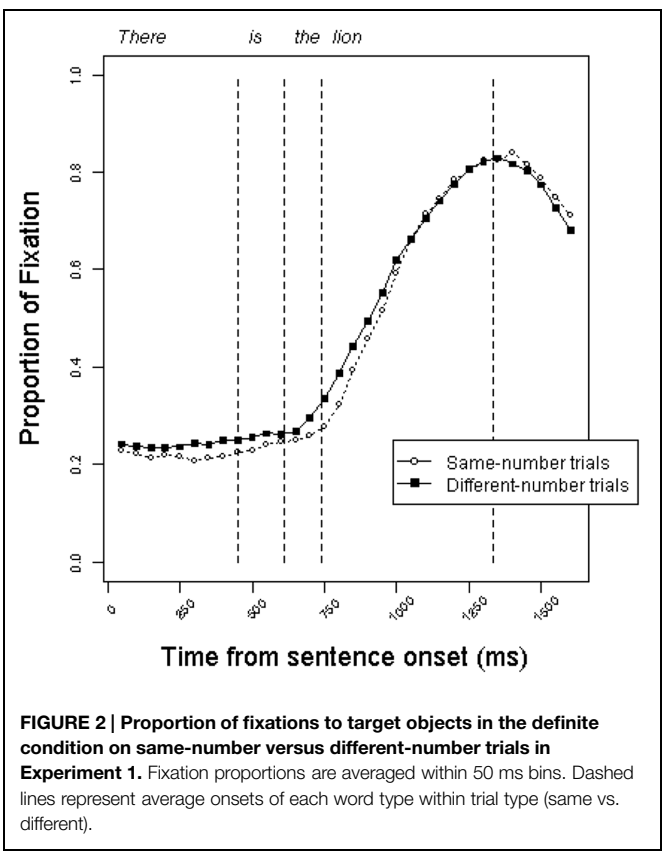


FIGURE 2 | Proportion of fixations to target objects in the definite condition on same-number versus different-number trials in Experiment 1. Fixation proportions are averaged within 50 ms bins. Dashed lines represent average onsets of each word type within trial type (same vs. different).

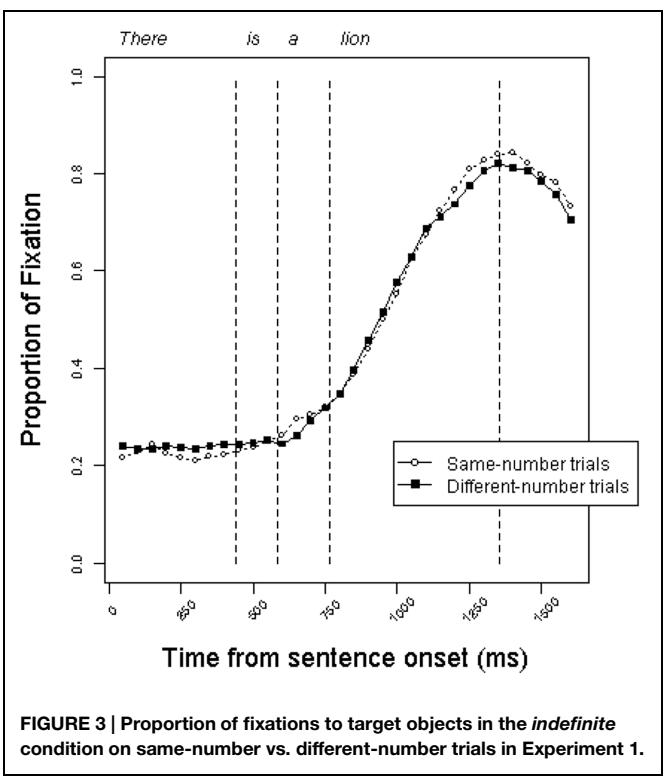


FIGURE 3 | Proportion of fixations to target objects in the *indefinite* condition on same-number vs. different-number trials in Experiment 1.

TABLE 2 | Experiment 1 reaction time (RT) analyses.

Condition		Same-number *M* (SD)	Different-number *M* (SD)	F_1	*p*	F_2	*p*
Definite	Singular	572 (109)	553 (142)	0.003	0.95	1.564	0.22
	Plural	574 (118)	555 (146)	0.132	0.72	0.003	0.96
Indefinite	Singular	524 (128)	572 (117)	0.778	0.38	8.836	0.006[1]
	Plural	613 (129)	560 (136)	0.017	0.90	0.400	0.53

[1] *In this case, RT on different-number trials was slower than on same-number trials.*

These results suggest that adults listening normally to simple declarative and interrogative sentences do not exploit grammatical number information to launch anticipatory eye movements to likely referents. We think it is unlikely that this null finding is due to a lack of power, given the consistent findings across both subjects and items, and the large number of exposures to each sentence type for each subject. Further, power analysis suggested sufficient observations for adequate sensitivity. However, it is possible that the surface structure led to strategic processing: anticipating that all sentences would have similar word order, participants may have adopted a strategy of simply waiting for the noun before shifting their gaze to the correct referent. Experiment 2 evaluated this possibility using the same stimuli and design as Experiment 1, but participants were instructed to select the correct referent as quickly as possible. Under these conditions, participants should use the grammatical number information on the copula and indefinite determiner to quickly orient to the correct picture.

Experiment 2

Method

Participants

Thirty native English speakers (not from Experiment 1) with normal or corrected-to-normal vision participated for course credit.

Stimuli and Design

Identical to Experiment 1.

Procedure

Participants were instructed to click on the picture that was mentioned in the sentence as quickly as possible without sacrificing accuracy. Otherwise, the procedure was identical to Experiment 1.

Results

An ANOVA with Experiment as a between-subjects factor revealed that the change in instructions had a dramatic effect on RTs: Experiment 2 RTs ($M = 496$, SD $= 121$) were faster than Experiment 1 RTs ($M = 566$, SD $= 129$) [$F_1(1,454) = 35.9$, $p < 0.001$; $F_2(1,252) = 40.8$, $p < 0.001$]. The percentage of trials on which participants launched saccades to the target before they could process the noun also increased: 51.9% of trials in the definite condition and 49.9% of trials in the indefinite condition.

Figures 4 and **5** present the time course of mean fixation proportions to the target pictures in the definite and indefinite conditions, respectively. Surprisingly, the trajectory of fixation proportions is similar to those in Experiment 1. The curves

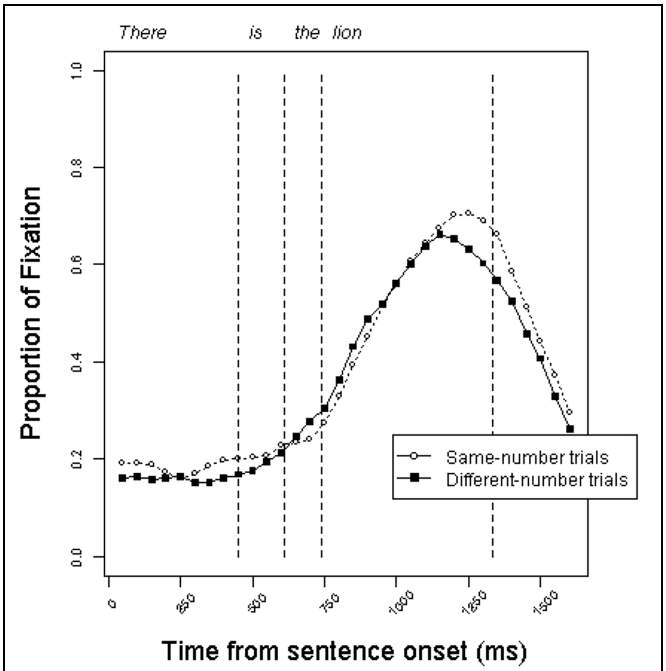


FIGURE 4 | Proportion of fixations to target objects in the definite condition on same-number vs. different-number trials in Experiment 2.

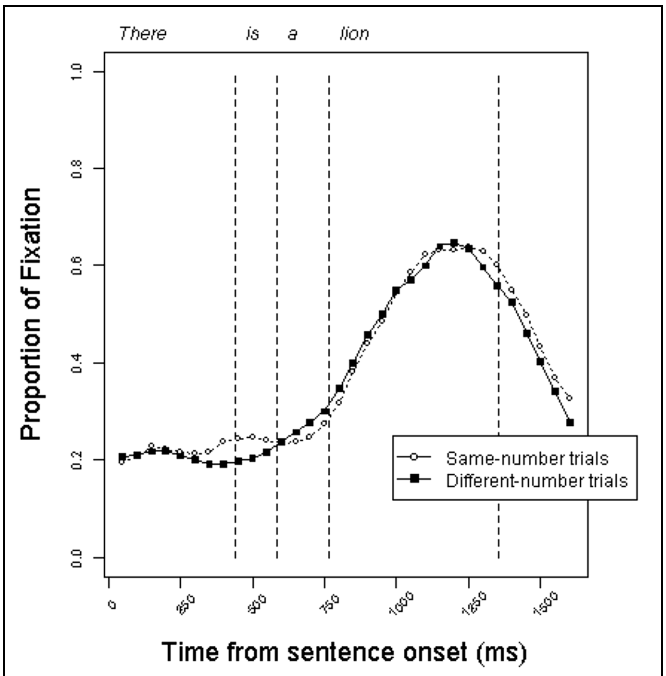


FIGURE 5 | Proportion of fixations to target objects in the *indefinite* condition on same-number vs. different-number trials in Experiment 2.

do not give an indication of anticipatory eye movements on different-number trials relative to same-number trials.

The RT analyses are presented in **Table 3**. As in Experiment 1, although there was a trend toward faster processing in the different-number trials, this impression was not statistically reliable in any of the analyses. This was true for both the definite and indefinite conditions, despite the difference in grammatical number information that was available to participants. The results of Experiment 2 corroborate the results of Experiment 1, suggesting that the result of Experiment 1 was not an artifact of strategic processing.

However, a potential concern still remains with Experiments 1 and 2. Since only declarative and interrogative sentences were used for the stimuli, it is possible that the results reflect strategies specific to the sentence types rather than a more general phenomenon of grammatical number processing in online language processing. Experiment 3 was designed to investigate this possibility using a similar design to Experiments 1 and 2 but with a wider range of sentence types.

Experiment 3

Method

Participants

Twenty native English speakers (not from Experiments 1 or 2) with normal or corrected-to-normal vision participated for course credit.

Stimuli and Design

Noun targets were 30 object names selected from McRae et al. (2005). These targets appeared in five conditions spanning auxiliary verbs in questions, declarative sentences, and demonstrative determiners. Each condition had singular and plural sentence versions, making 10 sentence sets. Three words were assigned to each sentence set and targets and distracter images were drawn from within the words in the sentence set. Distracters could not share the same initial phoneme as targets. Each target appeared in both same and different grammatical number conditions in separate trials of the experiment, yielding 60 unique grammatical number trials for each participant.

Because word types differ and length and word tokens differ in length with each utterance, across utterances, there is variation in the start and end of windows of interest. Therefore, it is common to align utterances based on the start of a window of interest for the purpose of analysis. An extension of this methodology to multiple windows of interest within an utterance involves resynchronizing at the start of each window (Altmann and Kamide, 1999). However, these techniques are only valid when the length of window, and word tokens within each window, are relatively homogeneous. Simply aligning utterances in this case runs the risk of glossing over utterance-specific eye movement behavior. Since our interest is in comparing the likelihood of launching a saccade based on information contained in function words, which are often phonetically reduced and of variable length, we chose to enforce an alignment of windows of interest across utterances by fixing the length of each window as shown in **Table 4**. Tokens shorter than the length of the window were followed by a short silence extending to the end of the window.

In addition to the grammatical number sentences, 60 new sentences were constructed using feature-target pairs selected from McRae et al. (2005) from 10 different feature types in order to compare anticipatory saccades as a function of feature type. However, these results will not be discussed in the current article. In order to ensure that participants did not develop an expectancy that target words would come later in the sentence, 60 filler sentences were created such that the first word was always the target referent. Target words for the filler sentences were the words from the feature experiment and filler sentences were generic sentences with plural subjects. The predicates of the filler sentences were features of the target word, but these features were different from the stimuli used in the feature experiment. Distracters could not share the same initial phoneme as targets. In 45 trials, both target and distracter images were plural. On the other 15 trials, the target image was singular while the distracter was plural. This was done to ensure that across the experiment participants did not develop expectations about the type of sentence they would hear based on the number-composition (i.e., target = singular, distracter = plural; etc.) of the image.

Procedure

Participants were required to make a saccade to an area of size 100 × 100 pixels surrounding a fixation dot in the center of the screen in order to initiate the sentence. This served to bring participants' fixations to a uniform location before the start of the sentence. Once a saccade was registered to the center interest

TABLE 3 | Experiment 2 RT analyses.

Condition		Same-number *M* (SD)	Different-number *M* (SD)	F_1	*p*	F_2	*p*
Definite	Singular	507 (119)	487 (119)	2.034	0.16	0.617	0.44
	Plural	492 (106)	453 (103)	1.297	0.26	0.134	0.72
Indefinite	Singular	503 (103)	491 (119)	0.728	0.40	1.637	0.21
	Plural	535 (136)	497 (155)	0.013	0.91	0.951	0.34

TABLE 4 | Experiment 3 alignment of different sentence types.

Window onset (ms)	0	300	600	900	1500	1900	2300
	is	there	a	dog			
	are	there	some	dogs			
		does	the	dog	have	brown	fur
		do	the	dogs	have	brown	fur
	there	is	the	dog			
	there	are	the	dogs			
	there	is	a	dog			
	there	are	some	dogs			
			that	dog	is	black	
			those	dogs	are	black	

area, there was a 300 ms pause, then the sentence was played. Otherwise, the procedure was identical to Experiment 2.

Results

The probability of initiating a saccade to the target object during a period starting 200 ms after the first word with grammatical number information and ending 150 ms after the onset of the target word was calculated for each participant by summing the number of trials in which a saccade to the target during this period occurred and dividing by the total number of trials. Since eye movements take approximately 180–200 ms to program, this is the critical period in which anticipatory eye movements could occur in response to the grammatical number information. Probabilities were calculated across all sentence types for each participant.

The anticipatory eye movement analyses are presented in **Table 5**. As in Experiments 1 and 2, no significant difference was observed between same-number and different- number trials, either for singular or plural sentences. The results of Experiment 3 further support the results of Experiments 1 and 2, suggesting that the results observed in these experiments were not due to the effect of strategic processing for different sentence types.

General Discussion

Many studies have demonstrated the important role that prediction plays in language processing. Prediction has been central to the study of world situated language comprehension, with demonstrations of anticipatory eye movements in response to a variety of different kinds of linguistic information. However, the three experiments presented here failed to find evidence that eye movements are tightly coordinated with the processing of morphosyntactic information. Listeners did not respond reliably faster on trials where grammatical number cues were informative about the identity of the upcoming referent relative to trials where grammatical number cues were uninformative. This was true both under natural listening conditions (Experiment 1) and when emphasizing a speeded response (Experiment 2). In addition, listeners were no more likely to look at the upcoming referent when grammatical number cues were informative as compared to trials where grammatical number was uninformative, using a mixed variety of sentence types (Experiment 3).

Our adults participants, all native English-speakers, presumably had considerable previous experience with the distributional structure of their mother tongue, and could use that knowledge to anticipate discourse as it unfolded (Haskell et al., 2010). That they did not capitalize on number as a predictive cue, even under speeded conditions, suggests that number has low cue validity; though verb number was a reliable guide to conceptual number in our experiments, this is not true of the language at large. This dovetails nicely with theoretical work indicating that in sentence processing, English speakers pay relatively little attention to subject–verb agreement marking in establishing numerosity, instead relying on word order to resolve key dependencies (MacWhinney et al., 1984), and with a raft of findings indicating that cue validity is key to attentional orienting.

TABLE 5 | Experiment 3 anticipatory eye movement analyses.

Condition	Same-number *M* (SD)	Different-number *M* (SD)	*F*	*p*
Singular	0.246 (0.217)	0.307 (0.218)	1.824	0.193
Plural	0.329 (0.229)	0.321 (0.248)	0.031	0.863

These results also complement that of Knoeferle and Crocker (2006), who found only a weak effect of tense and auxiliary words on eye movements. They found that auxiliary verbs such as *will* and *being* alone did not affect eye movements, but may have made the processing of the following verb and thematic role assignment faster. Knoeferle and Crocker (2006) concluded that there is generally a close coordination of scene processing and utterance comprehension, but this may be less so for words that only indirectly affect processing.

The finding that adult English-speakers do not reliably use grammatical number information to direct eye movements contrasts with the findings of Kouider et al. (2006) for young children (but see Johnson et al., 2005). As our experiments demonstrated, the nature of the task can have a large impact of the speed of eye movements in relation to linguistic input. Thus, the difference in findings could be attributed to differences in task, stimuli, or experimental procedure. A more interesting possibility is that novice and experienced English-language comprehenders differ qualitatively in their looking behavior during language comprehension.

Given the simplified nature of child-directed speech, adults may be more attuned to the range of possible continuations of the utterance following an opening such as *There is a. . .* For example, sentences with the singular copula *is* followed by the indefinite article *a* can be associated with plural referents, as when the referent is a collective noun, e.g., *There is a group of ducks in the water*. Thus, more experience with language in a variety of communicative contexts, and specifically with more complex NPs, may reduce adults' confidence in grammatical number morphology as a reliable cue to the identity of the upcoming referent. Indeed, because grammatical number information may not always be reliable, adults may make use of a form of "good-enough" processing (Ferreira et al., 2002) in these cases, computing an underspecified semantic expectation for possible referents (Sanford and Sturt, 2002).

This may be particularly true of certain constructions, such as the simple declaratives and interrogatives employed here, where grammatical number is only ever a partial guide to the numerosity of the referent. Naturally, there are many cases in which grammatical and conceptual number *do* align in such expressions, as was true of the sentences in our experiments. However, adults will also have been exposed to many instances in which grammatical number is highly unreliable as a predictive cue. For example, it will always be ambiguous for concrete mass nouns (*Where is the luggage she brought?*) and pluralia tantum (*There are some tongs on the counter*), where the number of the referent is left unspecified. Similarly, it will often be misleading when the verb is followed by a NP, and agreement is struck with the NP rather than the noun itself (*There is a herd of sheep*).

Varied conventions are not the only issue. A pair of large-scale corpus studies of British English confirms that agreement errors are quite common in declarative expressions, particularly in spoken language (Breivik and Martínez-Insua, 2008). Indeed, teenage speakers fail to achieve number agreement between the verb and post-verbal NP in more than a fifth of such utterances. The fact that number is not consistently informative in these contexts may help explain the growing tendency to omit number marking from them altogether (Meechan and Foley, 1994). In speech, English-speakers increasingly opt for the grammaticalized variants – *There's* and *Where's* – using these forms interchangeably with both singular and plural referents (*There's two ladies outside*).

It is not surprising then, that our participants did not rely on the number information encoded at the copula and determiner. Our null results argue against the notion that number in English is systematically informative about the numerosity of upcoming referents (see also Humphreys and Bock, 2005). More broadly, these results suggest that caution must be exercised in attempting to generalize the results of any one study – in any one language – to other studies in other languages, or to draw sweeping conclusions about the function of features like gender or number (MacWhinney et al., 1984). There is now an accumulating body of research attesting to cross-linguistic differences in morphosyntatic processing, showing systematic variation in number (Vigliocco et al., 1996; Berg, 1998) and gender processing (Miozzo and Caramazza, 1999; Schriefers and Teruel, 2000). Even within the same language, agreement processes may vary depending on the particulars of the construction (Kreiner et al., 2013), or the specific task demands (Brandt-Kobele and Höhle, 2010).

Cross-linguistic differences are to be expected. Languages vary widely in their "degree and specificity of morphological encoding" (Lupyan and Dale, 2010, p. 2), with some languages, like German, relying heavily on inflectional morphology to convey information, and others, like English, leaving more to the surrounding context—achieving lexically, what morphologically rich languages achieve through obligatory marking. In related work, Ramscar et al. (2015) have proposed that prenominal adjectives, in English, play a similar role to grammatical gender marking, in German. Both assist predictive processing; the difference is that one system is deterministic (only a certain set of nouns can legally follow the masculine article *der*), while the other is probabilistic (the distribution of nouns that follow *massive* and *moist* is markedly different, but not mutually exclusive). Thus, a possibility left open here is that rather than employing a rigid grammatical device, English simply relies on a more graded, semantically based means of specifying conceptual numerosity. This is consistent with the proposal that, in English, countability is a characteristic of NPs, rather than nouns (Allan, 1980), and that semantic principles selectively bias English agreement patterns (Berg, 1998).

In sum, English-speaking adults have difficulty consistently making use of grammatical number information to direct eye movements when processing simple declarative and interrogative sentences. This result indicates that the link between eye movements and linguistic processing is variable, depending especially on the linguistic information involved and the goals of language users.

Acknowledgments

This research was supported by NSF BCS-1056744 to MJ. BR was supported by NICHD (T32 HD07475). All three experiments reported here were conducted in accordance with Indiana University IRB 07-11661 "Eye Movements in Reading and Information Processing."

References

Allan, K. (1980). Nouns and countability. *Language* 56, 541–567. doi: 10.2307/414449

Altmann, G. T., and Kamide, Y. (1999). Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition* 73, 247–264. doi: 10.1016/S0010-0277(99)00059-1

Altmann, G. T. M., and Kamide, Y. (2004). "Now you see it, now you don't: mediating the mapping between language and the visual world," in *The Interface of Language, Vision, and Action: Eye Movements and the Visual World*, eds J. M. Henderson and F. Ferreira (New York, NY: Psychology Press), 347–386.

Arnold, J. E., Kam, C. L. H., and Tanenhaus, M. K. (2007). If you say thee uh you are describing something hard: the on–line attribution of disfluency during reference comprehension. *J. Exp. Psychol. Learn. Mem. Cogn.* 33, 914–930. doi: 10.1037/0278-7393.33.5.914

Arnon, I., and Ramscar, M. (2012). Granularity and the acquisition of grammatical gender: how order of acquisition affects what gets learned. *Cognition* 122, 292–305. doi: 10.1016/j.cognition.2011.10.009

Baldwin, T., and Bond, F. (2003). "Learning the countability of English nouns from corpus data," in *Proceedings of 41st Annual Meeting of the Association for Computational Linguistics: ACL–2003*, Stroudsburg, PA, 463–470. doi: 10.3115/1075096.1075155

Barber, H., and Carreiras, M. (2005). Grammatical gender and number agreement in Spanish: an ERP comparison. *J. Cogn. Neurosci.* 17, 137–153 doi: 10.1162/0898929052880101

Bates, E., Devescovi, A., Hernandez, A., and Pizzamiglio, L. (1996). Gender priming in Italian. *Percept. Psychophys.* 58, 992–1004. doi: 10.3758/BF03206827

Berg, T. (1998). The resolution of number conflicts in English and German agreement patterns. *Linguistics* 36, 41–70. doi: 10.1515/ling.1998.36.1.41

Bock, K., and Miller, C. A. (1991). Broken agreement. *Cogn. Psychol.* 23, 45–93. doi: 10.1016/0010-0285(91)90003-7

Brandt-Kobele, O.-C., and Höhle, B. (2010). What asymmetries within comprehension reveal about asymmetries between comprehension and production: the case of verb inflection in language acquisition. *Lingua* 120, 1910–1925. doi: 10.1016/j.lingua.2010.02.008

Breivik, L. E., and Martínez-Insua, A. E. (2008). Grammaticalization, subjectification and non–concord in English declarative sentences. *Engl. Stud.* 89, 351–362. doi: 10.1080/001383808020 11321

Corbett, G. G. (1991). *Gender*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9781139166119

Corbett, G. G. (2000). *Number*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9781139164344

Dahan, D., Swingley, D., Tanenhaus, M. K., and Magnuson, J. S. (2000). Linguistic gender and spoken–word recognition in French. *J. Mem. Lang.* 42, 465–480. doi: 10.1006/jmla.1999.2688

Dahan, D., and Tanenhaus, M. K. (2004). Continuous mapping from sound to meaning in spoken–language comprehension: immediate effects of verb–based thematic constraints. *J. Exp. Psychol. Learn. Mem. Cogn.* 30, 498–513. doi: 10.1037/0278-7393.30.2.498

Eberhard, K. M., Cutting, J. C., and Bock, K. (2005). Making syntax of sense: number agreement in sentence production. *Psychol. Rev.* 112, 531–559. doi: 10.1037/0033-295X.112.3.531

Ferreira, F., Bailey, K. G. D., and Ferraro, V. (2002). Good–enough representations in language comprehension. *Curr. Dir. Psychol. Sci.* 11, 11–15. doi: 10.1111/1467-8721.00158

Friederici, A. D., and Jacobsen, T. (1999). Processing grammatical gender during language comprehension. *J. Psycholinguist. Res.* 28, 467–484. doi: 10.1023/A:1023264209610

Frigo, L., and McDonald, J. L. (1998). Properties of phonological markers that affect the acquisition of gender–like subclasses. *J. Mem. Lang.* 39, 218–245. doi: 10.1006/jmla.1998.2569

Gahl, S., Yao, Y., and Johnson, K. (2012). Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech. *J. Mem. Lang.* 66, 1–18. doi: 10.1016/j.jml.2011.11.006

Grosjean, F., Dommergues, J. Y., Cornu, E., Guillelmon, D., and Besson, C. (1994). The gender–marking effect in spoken word recognition. *Percept. Psychophys.* 56, 590–598. doi: 10.3758/BF03206954

Hare, M., Elman, J. L., Tabaczynski, T., and McRae, K. (2009). The wind chilled the spectators, but the wine just chilled: sense, structure, and sentence comprehension. *Cogn. Sci.* 33, 610–628. doi: 10.1111/j.1551-6709.2009.01027.x

Haskell, T. R., and MacDonald, M. C. (2003). Conflicting cues and competition in subject–verb agreement. *J. Mem. Lang.* 48, 760–778. doi: 10.1016/S0749-596X(03)00010-X

Haskell, T. R., Thornton, R., and MacDonald, M. C. (2010). Experience and grammatical agreement: statistical learning shapes number agreement production. *Cognition* 114, 151–164. doi: 10.1016/j.cognition.2009.08.017

Huddleston, R., and Pullum, G. K. (2002). *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

Huettig, F., Rommers, J., and Meyer, A. S. (2011). Using the visual world paradigm to study language processing: a review and critical evaluation. *Acta Psychol.* 137, 151–171. doi: 10.1016/j.actpsy.2010.11.003

Humphreys, K. R., and Bock, K. (2005). Notional number agreement in English. *Psychon. Bull. Rev.* 12, 689–695. doi: 10.3758/BF03196759

Johnson, V. E., de Villiers, J. G., and Seymour, H. N. (2005). Agreement without understanding? The case of third person singular /s/. *First Lang.* 25, 317–330. doi: 10.1177/0142723705053120

Kamide, Y., Scheepers, C., and Altmann, G. T. M. (2003). Integration of syntactic and semantic information in predictive processing: cross–linguistic evidence from German and English. *J. Psycholinguist. Res.* 32, 37–55. doi: 10.1023/A:1021933015362

Kilarski, M. (2007). "On grammatical gender as an arbitrary and redundant category," in *Proceedings of the History of Linguistics 2005: Selected Papers from the 10th International Conference on the History of Language Sciences (ICHOLS X)*, Vol. 112, ed. D. Kilbee (Amsterdam: John Benjamins Publishing Company), 24–36.

Knoeferle, P., and Crocker, M. W. (2006). The coordinated interplay of scene, utterance, and world knowledge: evidence from eye tracking. *Cogn. Sci.* 30, 481–529. doi: 10.1207/s15516709cog0000_65

Kouider, S., Halberda, J., Wood, J., and Carey, S. (2006). Acquisition of English number marking: the singular-plural distinction. *Lang. Learn. Dev.* 2, 1–25. doi: 10.1207/s15473341lld0201_1

Kreiner, H., Garrod, S., and Sturt, P. (2013). Number agreement in sentence comprehension: the relationship between grammatical and conceptual factors. *Lang. Cogn. Process.* 28, 829–874. doi: 10.1080/01690965.2012.667567

Lew-Williams, C., and Fernald, A. (2007). Young children learning Spanish make rapid use of grammatical gender in spoken word recognition. *Psychol. Sci.* 18, 193–198. doi: 10.1111/j.1467-9280.2007.01871.x

Lupyan, G., and Dale, R. (2010). Language structure is partly determined by social structure. *PLoS ONE* 5:e8559. doi: 10.1371/journal.pone.0008559

MacWhinney, B., Bates, E., and Kliegl, R. (1984). Cue validity and sentence interpretation in English, German, and Italian. *J. Verbal Learn. Verbal Behav.* 23, 127–150. doi: 10.1016/S0022-5371(84)90093-8

Magnuson, J. S., Dixon, J. A., Tanenhaus, M. K., and Aslin, R. N. (2007). The dynamics of lexical competition during spoken word recognition. *Cogn. Sci.* 31, 133–156. doi: 10.1080/03640210709336987

McDonald, S. A., and Shillcock, R. C. (2003). Low-level predictive inference in reading: the influence of transitional probabilities on eye movements. *Vis. Res.* 43, 1735–1751. doi: 10.1016/S0042-6989(03)00237-2

McRae, K., Cree, G. S., Seidenberg, M. S., and McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behav. Res. Methods* 37, 547–559. doi: 10.3758/BF03192726

Meechan, M., and Foley, M. (1994). On resolving disagreement: linguistic theory and variation—"There's bridges". *Lang. Var. Change* 6, 63–85. doi: 10.1017/S0954394500001587

Miozzo, M., and Caramazza, A. (1999). The selection of determiners in noun phrase production. *J. Exp. Psychol. Learn. Mem. Cogn.* 25, 907–922. doi: 10.1037/0278-7393.25.4.907

Pickering, M. J., and Garrod, S. (2007). Do people use language production to make predictions during comprehension? *Trends Cogn. Sci.* 11, 105–110. doi: 10.1016/j.tics.2006.12.002

Pulvermüller, F., and Shtyrov, Y. (2003). Language outside the focus of attention: the mismatch negativity as a tool for studying higher cognitive processes. *Prog. Neurobiol.* 79, 49–71. doi: 10.1016/j.pneurobio.2006.04.004

Ramscar, M., Dye, M., Milin, P., and Futrell, R. (2015). "The social evolution and communicative function of noun classification (abstract)," in *Proceedings of the 37th Annual Conference of the Cognitive Science Society*, eds R. Dale, C. Jennings, P. Maglio, T. Matlock, D. Noelle, A. Warlaumont, et al. (Pasadena, CA: Cognitive Science Society).

Ramscar, M., Yarlett, D., Dye, M., Denny, K., and Thorpe, K. (2010). The effects of feature-label-order and their implications for symbolic learning. *Cogn. Sci.* 34, 909–957. doi: 10.1111/j.1551-6709.2009.01092.x

Robertson, E. K., Shi, R., and Melançon, A. (2012). Toddlers use the number feature in determiners during online noun comprehension. *Child Dev.* 83, 2007–2018. doi: 10.1111/j.1467-8624.2012.01828.x

Rossion, B., and Pourtois, G. (2004). Revisiting Snodgrass and Vanderwart's object pictorial set: the role of surface detail in basic-level object recognition. *Perception* 33, 217–236. doi: 10.1068/p5117

Sanford, A., and Sturt, P. (2002). Depth of processing in language comprehension: not noticing the evidence. *Trends Cogn. Sci.* 6, 382. doi: 10.1016/S1364-6613(02)01958-7

Schriefers, H. (1993). Syntactic processes in the production of noun phrases. *J. Exp. Psychol. Learn. Mem. Cogn.* 19, 841–850. doi: 10.1037/0278-7393.19.4.841

Schriefers, H., and Teruel, E. (2000). Grammatical gender in noun phrase production: the gender interference effect in German. *J. Exp. Psychol. Learn. Mem. Cogn.* 26, 1368–1377. doi: 10.1037/0278-7393.26.6.1368

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., and Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science* 268, 1632–1634. doi: 10.1126/science.7777863

Tanenhaus, M. K., and Trueswell, J. C. (2006). "Eye movements and spoken language comprehension," in *Handbook of Psycholinguistics*, 2nd Edn, eds M. J. Traxler and M. A. Gernsbacher (Amsterdam: Elsevier Press).

Van Berkum, J. J. A., Brown, C. M., Zwitserlood, P., Kooijman, V., and Hagoort, P. (2005). Anticipating upcoming words in discourse: evidence from ERPs and reading times. *J. Exp. Psychol. Learn. Mem. Cogn.* 31, 443–467. doi: 10.1037/0278-7393.31.3.443

van Heugten, M., and Shi, R. (2009). French-learning toddlers use gender information on determiners during word recognition. *Dev. Sci.* 12, 419–425. doi: 10.1111/j.1467-7687.2008.00788.x

Vigliocco, G., Antonini, T., and Garrett, M. F. (1997). Grammatical gender is on the tip of italian tongues. *Psychol. Sci.* 8, 314–317. doi: 10.1111/j.1467-9280.1997.tb00444.x

Vigliocco, G., Butterworth, B., and Garrett, M. F. (1996). Subject–verb agreement in Spanish and English: differences in the role of conceptual constraints. *Cognition* 61, 261–298. doi: 10.1016/S0010-0277(96)00713-5

Vigliocco, G., Vinson, D. P., Martin, R. C., and Garrett, M. F. (1999). Is "count" and "mass" information available when the noun is not? An investigation of tip of the tongue states and anomia. *J. Mem. Lang.* 40, 534–558. doi: 10.1006/jmla.1998.2626

Vigliocco, G., Vinson, D. P., Paganelli, F., and Dworzynski, K. (2005). Grammatical gender effects on cognition: implications for language learning and language use. *J. Exp. Psychol. Gen.* 134, 501–520. doi: 10.1037/0096-3445.134.4.501

Wagers, M. W., Lau, E. F., and Phillips, C. (2009). Agreement attraction in comprehension: representations and processes. *J. Mem. Lang.* 61, 206–237. doi: 10.1016/j.jml.2009.04.002

Weber, A., and Paris, G. (2004). "The origin of the linguistic gender effect in spoken–word recognition: evidence from non–native listening," in *Proceedings of the 26th Annual Meeting of the Cognitive Science Society*, eds K. Forbus, D. Gentner, and T. Tegier (Mahwah, NJ: Erlbaum).

Wicha, N. Y. Y., Moreno, E. M., and Kutas, M. (2004). Anticipating words and their gender: an event-related brain potential study of semantic integration, gender expectancy, and gender agreement in Spanish sentence reading. *J. Cogn. Neurosci.* 16, 1272–1288. doi: 10.1162/0898929041920487

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2015. Riordan, Dye and Jones. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 16 June 2015
doi: 10.3389/fpsyg.2015.00766

An interference account of the missing-VP effect

Jana Häussler[1]* *and Markus Bader*[2]

[1] Department of Linguistics, University of Potsdam, Potsdam, Germany, [2] Department of Linguistics, Goethe University Frankfurt, Frankfurt, Germany

Sentences with doubly center-embedded relative clauses in which a verb phrase (VP) is missing are sometimes perceived as grammatical, thus giving rise to an illusion of grammaticality. In this paper, we provide a new account of why missing-VP sentences, which are both complex and ungrammatical, lead to an illusion of grammaticality, the so-called missing-VP effect. We propose that the missing-VP effect in particular, and processing difficulties with multiply center-embedded clauses more generally, are best understood as resulting from interference during cue-based retrieval. When processing a sentence with double center-embedding, a retrieval error due to interference can cause the verb of an embedded clause to be erroneously attached into a higher clause. This can lead to an illusion of grammaticality in the case of missing-VP sentences and to processing complexity in the case of complete sentences with double center-embedding. Evidence for an interference account of the missing-VP effect comes from experiments that have investigated the missing-VP effect in German using a speeded grammaticality judgments procedure. We review this evidence and then present two new experiments that show that the missing-VP effect can be found in German also with less restricting procedures. One experiment was a questionnaire study which required grammaticality judgments from participants without imposing any time constraints. The second experiment used a self-paced reading procedure and did not require any judgments. Both experiments confirm the prior findings of missing-VP effects in German and also show that the missing-VP effect is subject to a primacy effect as known from the memory literature. Based on this evidence, we argue that an account of missing-VP effects in terms of interference during cue-based retrieval is superior to accounts in terms of limited memory resources or in terms of experience with embedded structures.

Keywords: sentence parsing, center embedding, grammatical illusion, missing-VP effect, cue-based retrieval, interference, German

OPEN ACCESS

Edited by:
Matthew Wagers, University of California, Santa Cruz, USA

Reviewed by:
Clinton L. Johns, Haskins Laboratories, USA
Manuel Gimenes, University of Poitiers, France

***Correspondence:**
Jana Häussler, Department of Linguistics, University of Potsdam, Karl-Liebknecht-Straße 24–25, 14476 Potsdam, Germany
jana.haeussler@uni-potsdam.de

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: *12 January 2015*
Accepted: *22 May 2015*
Published: *16 June 2015*

Citation:
Häussler J and Bader M (2015) An interference account of the missing-VP effect. Front. Psychol. 6:766. doi: 10.3389/fpsyg.2015.00766

1. Introduction

Some sentences are more difficult to process than other sentences, and some sentences are so complex that they exceed the processing capacity of the human parser and thereby lead to processing overload. A striking illustration of the parser's limited capacity is provided by sentences with multiple center-embedding as illustrated by the example in (1) from Frazier (1985).

(1) The patient the nurse the clinic had hired admitted met Jack.

Sentences with multiple center-embedding have long been known to be difficult to process (Chomsky and Miller, 1963; Miller and Chomsky, 1963; Miller and Isard, 1964; Bever, 1970; Kimball, 1973, e.g.,). Sentences with two degrees of center-embedding can still be comprehended under certain conditions, as demonstrated by the following sentence from Bever (1974), in which the subject of the most deeply embedded relative clause is a first-person pronoun and not a lexical NP.

(2) The reporter who everyone *that I met* trusts said the president won't resign yet.

Sentences with two degrees of center-embedding are also produced from time to time, at least in written language (cf. Karlsson, 2007). With two levels of center-embedding, the maximum degree of center-embedding is already reached, however, and sentences with three or more degrees of center-embedding seem to be beyond the capacity of human parsing and human sentence production.

In comparison to sentence (1), the closely related sentence in (3) seems much easier to understand.

(3) The patient the nurse the clinic had hired met Jack.

Sentence (3) is an example of the so-called *missing-VP effect*, a term coined by Gibson and Thomas (1999) for the observation that people often fail to notice the lack of a verb phrase in sentences involving multiple center-embedding. The effect was first discussed by Frazier (1985), who attributes the observation to Janet Fodor.

Missing-VP sentences contain two degrees of center-embedding and an uncontroversial ungrammaticality. Each of these properties alone should suffice to decrease the acceptability of such sentences, and when the two properties occur together, a highly degraded sentence should result. However, instead of being perceived as highly degraded, the acceptability of such sentences is as high or even higher as the acceptability of corresponding complete and thereby grammatical sentences. This was first demonstrated by Gibson and Thomas (1999) in a rating study examining sentences like (4)[1].

(4) The ancient manuscript that the graduate student who the new card catalog [$_{VP3}$ had confused a great deal] [$_{VP2}$ was studying in the library] [$_{VP1}$ was missing a page].

Sentences were either complete or were missing one of VP1, VP2, or VP3 and had to be rated for their intuitive complexity. While sentences with either missing VP1 or missing VP3 were rated as being significantly more complex than complete sentences, the ratings for sentences with missing VP2 did not differ significantly from the ratings for complete sentences. Later research by Christiansen and MacDonald (2009) and Vasishth et al. (2010) showed that sentences in which VP2 is missing are more often perceived as grammatical and are easier to process than corresponding complete sentences with two degrees of center-embedding. Gimenes et al. (2009) have found similar results for French, another SVO language. The only SOV language for which evidence on the missing-VP effect exists seems to be German, but this evidence is mixed. Since our experiments investigate German sentences, we postpone a discussion of the missing-VP effect in this language to Section 4.

The missing-VP effect belongs to a small class of grammatical illusions—sentences which tend to be perceived as grammatical despite containing an undisputed ungrammaticality. In their review of grammatical illusions, Phillips et al. (2011, p. 166) exclude the missing-VP effect from further consideration because examples as in (3) "differ from the others discussed here in the respect that they plausibly reflect complexity-induced overload, and it is not clear what parse is assigned to such dramatically ill-formed sentences". We take the view that the missing-VP effect reflects complexity-induced overload to be uncontroversial. However, there are competing conceptions as to the source of parsing overload. The major aim of this paper is to provide an account of the missing-VP effect that follows much recent work in cognitive psychology and psycholinguistics claiming that overload is mainly a matter of interference during memory retrieval. Based on this hypothesis, we will argue that the parse assigned to missing-VP sentences differs minimally from the parse assigned to corresponding complete sentences. The only difference is that for complete sentences, all VP slots of the syntactic representation are filled by lexical material whereas for missing-VP sentences one of the VP slots remains empty. These claims are based on a review of prior experimental investigations of the missing-VP effect in German and on two new experiments, which were run with the additional aim of resolving some contradictions that concern the status of the missing-VP effect in German.

The organization of this paper is as follows. In Sections 2 and 3, we discuss two approaches to capacity limitations of cognitive processes and how they might account for the missing-VP effect. Section 2 introduces the resource account of capacity limitations and Section 3 the interference account. Section 4 reviews evidence from German favoring the interference account over the resource account. Some concerns regarding this evidence is addressed by two experiments that are presented in Sections 5 and 6. Section 7 concludes with a general discussion of the experimental results.

2. Resource Accounts of the Missing-VP Effect

The parser is not alone in being capacity limited. Most if not all cognitive abilities share this property. For example, our ability for mental calculations is restricted to a small subset of numbers, our ability to recall lists of unrelated items is limited to lists of no more than seven or eight items, and so on. Limitations of this kind are often attributed to a working memory system of limited

[1]In Gibson and Thomas (1999), VPs are numbered according to their linear position in the sentence string. In later publications (Christiansen and MacDonald, 2009; Gimenes et al., 2009; Vasishth et al., 2010), VPs are numbered according to their hierarchical position in the phrase-structure tree. In this paper, we adopt the latter numbering. VP1 is the VP of the matrix clause (S1), VP2 is the VP of the upper relative clause (S2), and VP3 is the VP of the lower relative clause (S3). The VPs therefore appear in the linear order VP3 VP2 VP1.

capacity. The question then becomes why working memory has such a severely limited capacity. Over time, this question has received various answers (see overviews in Oberauer and Kliegl, 2001; van Dyke and Johns, 2012).

Before we take a closer look at these answers, let us first get clear about the tasks that have to be accomplished in order to parse sentences successfully. By definition, a parser takes the words of an input string and constructs a syntactic structure for them. In the following, we assume the syntactic structure to be a conventional phrase-structure tree. Given the strong evidence that human parsing proceeds in an incremental way, the parser's task can be divided into two major subtasks (see Just and Carpenter, 1992; Gibson, 1998). First, the parser must store the syntactic structure, which is incremented word-by-word, in some kind of temporary buffer. Secondly, the parser must integrate each word of the input string into the unfolding syntactic structure as soon as the word is encountered. This subtask can be decomposed further. First, the parser must find a place within the ongoing syntactic structure where the word can be attached. Second, the word must be connected to words that are already part of the ongoing syntactic structure. For example, a verb must be connected to its arguments for thematic role and case assignment and for checking agreement requirements.

With the distinction between storage and integration at hand, we now come back to the question of why human parsing is subject to severe capacity limitations. For a long time, the dominant approach to capacity limitations was based on the claim that cognitive processes draw on a limited pool of processing resources. Applied to the issue of sentence parsing, the Resource Hypothesis states that the parser can use only a fixed amount of resources for the storage and computation of syntactic structures. When the available resources do not suffice for processing sentences of high complexity, processing overload results. An influential theory of sentence comprehension building on the Resource Hypothesis is the Capacity Theory of Just and Carpenter (1992). According to the Capacity Theory, each individual has a fixed amount of processing resources available for processing language. These resources can be allocated flexibly to the storage of intermediate syntactic structures and the incremental integration of words into the intermediate structure built thus far. The assumption that the parser must use a fixed pool of resources for both storage and processing is shared by the Syntactic Prediction Locality Theory (SPLT) of Gibson (1998) and its successor, the Dependency Locality Theory (DLT) of Gibson (2000).

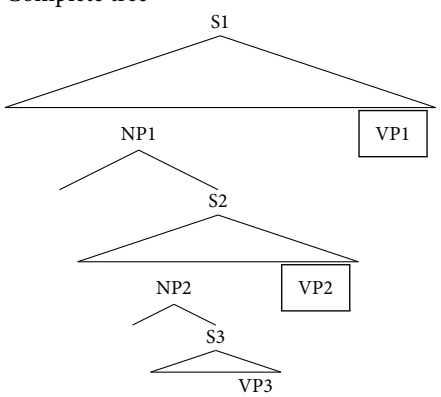
According to resource-based theories, sentences with a high degree of center-embedding cannot be successfully parsed because they require more resources than are available. This suggests an explanation of the missing-VP effect along the following lines. When the parser is processing a sentence of high syntactic complexity, it may be short of running out of resources. In such a case, the parser can try to proceed by forgetting some part of the structure built thus far, thereby freeing resources needed to continue the ongoing parsing process. Two implementations of this idea are the *Disappearing Syntactic Nodes Hypothesis* of Frazier (1985) and the *High Memory Cost Pruning Hypothesis* of Gibson and Thomas (1999). Both implementations share the idea that the phrase structure tree is cut down under conditions of high memory load. For reasons of space, we only give the High Memory Cost Pruning Hypothesis below.

(5) *The High Memory Cost Pruning Hypothesis* (Gibson and Thomas, 1999, p. 231)
At points of high memory complexity, forget the syntactic prediction(s) associated with the most memory load.

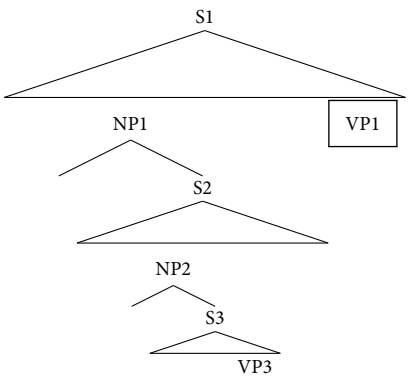
The High Memory Cost Pruning Hypothesis was formulated within the SPLT of Gibson (1998). According to the SPLT's definition of storage cost, the prediction of VP2 is associated with the most memory load in sentences with doubly center-embedded relative clauses (see Gibson and Thomas, 1999; Vasishth et al., 2010, for details). The prediction of VP2 is therefore forgotten. Instead of the complete tree shown on the left side in (6), the incomplete tree on the right side is available for the parser at the point where the two final VPs of a missing-VP sentence are about to be integrated.

The first VP that the parser encounters is put into the open slot for VP3 and the next VP into the slot for VP1. In a sentence with missing VP2, all VPs of the input string can thus be successfully integrated. Because the slot for VP2 is no longer present in the syntactic representation, the parser fails to notice the lack of a VP. In the case of complete sentences with a doubly center-embedded relative clause, the representation on the right side of (6) does not provide an attachment site for each VP. Such sentences should

(6) Complete tree



Tree with VP2 pruned



thus be more difficult to process than sentences with a missing VP. This was not the case in the off-line ratings reported in Gibson and Thomas (1999), but on-line evidence obtained by Christiansen and MacDonald (2009) and Vasishth et al. (2010) shows that sentences in which VP2 is missing are easier to process than corresponding sentences in which VP2 is present.

A conceptual drawback of the High Memory Cost Pruning Hypothesis is that it does not follow from independently motivated principles of storage or computation. Resource-based accounts of capacity limitations typically assume that trace decay is an important source of storage limitations. When applied to sentences with double center-embedding, the prediction of VP1 should be less available than the prediction of VP2 because it was introduced earlier and had therefore more time for decay. This is just the opposite of what the Pruning Hypothesis claims and is accordingly not compatible with the findings for missing-VP sentences. Additional machinery is therefore necessary in order to derive that the prediction of VP2 is pruned but not the prediction of VP1. For example, the parser must somehow be able to calculate the memory load of each prediction in order to prune the one with the most memory load. These calculations are heavily theory dependent. As discussed in more detail in Vasishth et al. (2010), the memory-load definitions of the DLT do no longer predict that VP2 is pruned, although—as also shown in Vasishth et al. (2010)—it is possible to adapt the Pruning Hypothesis to the particular properties of the DLT.

We will not dwell further on this issue because resource-based accounts of capacity limitations in general and the concept of trace decay have fallen into disreputation, both for theoretical reasons (e.g., Navon, 1984, MacDonald and Christiansen, 2002) and for lack of empirical support (e.g., Oberauer and Kliegl, 2001). Two influential alternatives to the resource-based view are the *interference account* and the *experience-based account* (further alternatives are discussed in Oberauer and Kliegl, 2001). In the next section, we propose an explanation of the missing-VP effect that is based on the interference account. The experience-based account is discussed in the final section[2].

3. An Interference Account of the Missing-VP Effect

The interference account is based on the observation that the retrieval of material from working memory becomes less reliable in the presence of similar material. In order to apply the interference account to the process of sentence parsing, we have to take a closer look at the steps that lead to the integration of new words into the unfolding syntactic representation. A crucial first step for successful integration is the retrieval of the correct attachment site for the word that is to be integrated. If the retrieval cues used for this purpose match more than a single attachment site, finding the proper place for attaching the next word can become more difficult.

While interference from similar items can make the integration of new items more difficult, whether interference does indeed occur depends on the particular syntactic configuration. Consider first the situation that obtains in sentences with double center-embedding at the point where the verb of the most deeply embedded relative clause has to be integrated, that is, *met* in sentence (7).

(7) [$_{S1}$ [*The reporter*] [$_{S2}$ who [**everyone**] [$_{S3}$ that I [$_{V3}$ met]] [$_{V2}$ **trusts**] …
[$_{S1}$ NP1 [$_{S2}$ NP2 [$_{S3}$ NP3]

The syntactic representation built up to this point contains three clauses which still need a VP, namely the matrix clause S1 and the two relative clauses S2 and S3. Despite the existence of three potential attachment sites, the integration of the verb *met* into the embedded relative clause will not be disturbed by the presence of other potential attachment sites. The reason for this is that the most recently read subject is in the focus of attention and therefore immediately available for integration (see McElree, 2006, for the notion of focal attention as used in the memory literature).

The situation changes when the parser encounters the verb of the higher relative clause, *trusts* in sentence (5). After the most deeply embedded relative clause has been processed, the ongoing phrase-structure representation still contains two possible attachment sites for a verb. In this case, choosing the correct attachment site is not so easy for the parser because due to the intervening relative clause S3, the parser faces the task of switching back to a clause that is no longer in the focus of attention. Because there are two such clauses and each contains a slot for a VP, retrieving the correct integration site is difficult due to interference from the competing integration site.

Experimental evidence that attaching a word into the current clause is qualitatively different from attaching it to a clause that has been interrupted by one or more embedded clauses has been provided by McElree et al. (2003). In an experiment using a response-signal speed-accuracy tradeoff procedure, the verb either occurred adjacent to the head noun of the subject NP or was separated from it by either one or two relative clauses plus an additional PP in some cases. The results show that sentences in which the verb was adjacent to the subject head noun were associated with a higher asymptotic accuracy and also with a faster retrieval speed. This suggests that integrating a word with the immediately preceding word has a special status, in accordance with findings from the memory literature that only the most recent item is in focal attention (see McElree, 2006). However, Foraker and McElree (2011) and McElree and Dyer (2013) cite unpublished data by McElree and Wagers that argue against a too narrow definition of focal attention for the purpose of sentence parsing. McElree and Wagers found that an intervening relative clause removes the subject head noun

[2]Resource-based and interference-based accounts are not necessarily incompatible with each other. For example, cognitive architectures such as ACT-R and its relatives SOAR and 3CAPS usually include both assumptions—a limited amount of resources as well as interference from similar items—but differ with regard to the component they emphasize.

from focal attention but an intervening PP (8a) or an intervening adverbial (8b) do not.

(8) a. The crowd gasped as the driver *of the ambulance* fainted.
b. The crowd gasped as the driver *abruptly* fainted.

At the current state of knowledge, it does not seem to be possible to come up with a precise definition of the scope of focal attention. What can be concluded from the literature is that an intervening relative clause removes material preceding it from focal attention whereas at least some non-clausal constituents do not. We therefore propose the *Discrimination Hypothesis* in (9) (for related ideas, see Bader et al., 2003; Lewis and Vasishth, 2005; and Bader, 2015).

(9) The *Discrimination Hypothesis*
The integration of new material becomes difficult when an intervening clause separates the word that is to be integrated next from the required integration site and an incorrect but similar integration site competes for attachment.

When processing a sentence with doubly center-embedded relative clauses, two potential integration sites are available at the point where the second verb has to be integrated into the ongoing syntactic representation. According to the *Discrimination Hypothesis*, these two integration sites are difficult to discriminate. When the second verb has to be integrated, the parser may therefore retrieve S1 as integration site for V2 instead of S2. This will result in a syntactic representation in which the verb slot of S1 is filled whereas the verb slot of S2 remains empty. Because complete sentences exhibiting double center-embedding and corresponding missing-VP sentences diverge only after the second verb, such an ill-formed representation can arise for both of them. A mis-attachment of V2 to the verb slot of S1 can therefore happen in both cases. In complete sentences, this can give rise to the well-known processing difficulties of sentences with double center-embedding. For missing-VP sentences, this can lead to an illusion of grammaticality.

Crucially, the Discrimination Hypothesis does not claim that anything is deleted from the ongoing syntactic representation of a sentence. The only claim is that the parser sometimes attaches a word to an incorrect attachment site. As a result, the syntactic structure for a sentence can contain a node that has not been filled with lexical material. In order for a grammatical illusion to occur, the parser must not detect that a VP slot is still empty after the last word of the sentence has been processed. We therefore have to complement the Discrimination Hypothesis with appropriate assumptions concerning the processes that check whether a sentence obeys all syntactic constraints or not[3].

What could be the reason that the parser at times fails to detect that a sentence is incomplete? To begin with, consider sentence (10), a variant of sentence (4) of Gibson and Thomas (1999).

(10) [$_{S1}$ A page was missing in the ancient manuscript
[$_{S2}$ that the graduate student [$_{S3}$ who the new card catalog had confused a great deal]]].

In contrast to the original example, the higher relative clause appears in a sentence final position in (10) and is thus no longer center-embedded. In sentence (10), VP2 (the VP of the higher relative clause) is missing, but in this case the resulting ungrammaticality seems easy to detect[4]. This shows that a missing VP goes unnoticed only under conditions of high processing load. The question then is what these conditions are. One major issue concerns the absence of a grammatical illusion when VP1 is missing, that is, in sentences as in (11) [partially repeated from (4)].

(11) The manuscript that the graduate student who the new card catalog had confused a great deal *was studying in the library*.

a. Integration of the final verb *was studying* as VP1 into S1:

[$_{S1}$ *The manuscript* [$_{S2}$ that the graduate student [...] Δ] *was studying in the library*.]

b. Integration of the final verb as VP2 into S2:

[$_{S1}$ The manuscript [$_{S2}$ *that the graduate student* [...] *was studying in the library*.] Δ]

Sentences of this type were rated as highly complex in the experiment of Gibson and Thomas (1999). There are at least two alternative reasons for this. First, the final VP in (11) is integrated into S1 as VP1, giving rise to the configuration in (11a), which is complex due to its semantic implausibility. Alternatively, the final VP could be integrated into S2 as VP2, resulting in the configuration in (11b). This configuration will be perceived as complex only when one detects that the initial NP remains without a VP.

According to the Discrimination Hypothesis, VP2 and VP1 are both available as attachment sites when the final VP in (11) is about to be integrated. The finding of Gibson and Thomas (1999) that a missing-VP effect occurs when VP2 is missing but not when VP1 is missing can be accounted for in an interference-based framework by recourse to the notion of primacy. The opposite notion, namely recency, has already been made responsible for the fact that integrating V3 into S3 is not subject to interference and therefore unproblematic. S1 in sentences like (11) is the first clause not only in a hierarchical sense but also in a temporal sense. It therefore enjoys the advantage of primacy that is well-known from studies of memory retrieval (e.g., Knoedler et al., 1999). This advantage can have two consequences. First, it can cause V2 to be integrated more readily into S1 than into S2, resulting in configuration (11a). Second, it can ease the detection of a missing VP1 in case V2 was correctly integrated into S2, as in (11b). If this reasoning is on the right

[3]Note that similar assumptions would also be necessary for accounts assuming the deletion of a VP from the phrase structure representation. Pruning of a predicted VP leaves a subject NP without corresponding predicate within the ongoing syntactic structure. Thus, for a missing-VP sentence to be judged as grammatical, a pruning account has to claim that the parser overlooks the dangling subject NP.

[4]For English, Gibson and Thomas (1999) cite an unpublished acceptability experiment by Gibson and Kaan as providing evidence for this claim. In an unpublished experiment using the procedure of speeded grammaticality judgments, we found that German sentences corresponding to (10) were rejected as ungrammatical in almost 90% of the trials.

track, it should be possible to find evidence for a missing-VP effect for VP1 when the primacy advantage is taken away from S1. Evidence of this kind is discussed in the next section.

4. Evidence for the Interference Account

Experimental evidence for the interference account presented in the preceding section comes from an investigation of the missing-VP effect in German by Bader et al. (2003). In contrast to Gibson and Thomas (1999), who had participants rate the complexity of sentences on a scale from 1 to 5, Bader et al. (2003) required participants to give a binary grammaticality judgment. The rationale behind this decision was as follows. The defining property of a grammatical illusion is that a sentence is perceived as grammatical despite containing an undisputed ungrammaticality. Thus, the most straightforward way to test whether an ungrammatical sentence causes a grammatical illusion or not is to have native speakers judge its grammaticality. If the sentence is judged as grammatical, we can conclude that it caused a grammatical illusion.

Things are more complicated because we cannot expect that the illusion of grammaticality will arise on each single occasion. For some of the grammatical illusions that are discussed in Phillips et al. (2011), judgment data are available, showing that the strength of such illusions can vary considerably. For example, sentences with a negative polarity item and a negation not c-commanding the polarity item give rise to a negative polarity illusion. In a judgment experiment by Drenhaus et al. (2005), such sentences were erroneously accepted as grammatical in 30% of the time, which is only 10% more than for sentences with a negative polarity item and no negation at all. For the case illusion reported in Bader et al. (2000) and Meng and Bader (2000), the false acceptance rate was about 40% for sentences in which the verb assigned dative case to an NP which was case-ambiguous but not compatible with dative case (Bader et al., 2000). When this NP was made more complex by adding a relative clause, the false acceptance rate increased to a value of about 60% (Meng and Bader, 2000). Grammatical illusions are thus not an all-or-nothing matter, but a probabilistic phenomenon instead. Grammatical illusions do not differ from semantic illusions in this respect. For example, when testing the Moses illusion by means of a truth judgment task, Erickson and Mattson (1981) found that the sentence *Moses took two animals of each kind on the Ark.* was judged as true by 41% of the participants who possessed the relevant knowledge. Thus, semantic illusions are probabilistic too.

The particular procedure for obtaining grammaticality judgments used by Bader et al. (2003) was the procedure of speeded grammaticality judgments. Sentences were presented visually one word at a time. Participants were asked to judge sentences as either grammatical or ungrammatical as quickly as possible. A time limit of 2000 ms starting at the offset of the last word was imposed in order to encourage fast decisions. On average, participants responded even faster. Using this method, Bader et al. (2003) investigated whether the evidence provided by Gibson and Thomas (1999) can be replicated for German. The experiments provided two major results. First, participants accepted sentences with a missing VP as grammatical in a substantial number of trials. This shows that the grammatical illusion caused by a missing VP also occurs in German—at least when participants have to judge sentences under time pressure. The second major finding concerns the difference between sentences in which VP1 is missing and sentences in which VP2 is missing. In accordance with the initial observation in Frazier (1985), Gibson and Thomas (1999) found that the missing VP effect occurs when VP2 is missing but not when VP1 is missing. The same was found for German sentences which were similar to the sentences investigated by Gibson and Thomas in that the head noun of the highest relative clause was part of a main clause. When this noun was part of an embedded that-clause, however, participants often accepted incomplete sentences whether VP1 or VP2 was missing.

Thus, missing-VP sentences in which the final VP was syntactically and semantically compatible with attachment to either S1 or S2 were accepted most of the time. A relevant example is provided in (12).

(12) *Missing VP, singular S2 subject*

	Klaus	hat	mir	erzählt,							
	K.	has	me	told							
S1	dass	**jemand**		die	Sängerin,						
	that	someone.SG		the	singer.SG						
S2	*die*	den	Moderator,								
	who.SG	the	moderator								
S3	der	das	Interview	trotz	einer	Grippe	führen	musste,			
	who	the	interview	despite	a	flu	conduct	must			
S1/S2	***beleidigt***	***hat***.									
	insult	has.SG									

final verb in S2: "Klaus told me that someone Δ the singer who insulted the moderator who had to conduct the interview despite a flu"

final verb in S1: "Klaus told me that someone insulted the singer who Δ the moderator who had to conduct the interview despite a flu"

In (12), a verb with an animate subject and an animate direct object is required for completion of both S1 and S2. Since the clause-final verb *beleidigt hat* ("insulted has") meets both requirements, it can be attached to either S1 or S2. For such sentences, the acceptance rate reached a high value of about 75%, which is even slightly higher than for complete sentences. When syntactic or semantic constraints only allowed attachment to either S1 or S2, missing-VP sentences were accepted significantly less often, although still about half of the time. In the context of other types of grammatical illusions, missing VPs thus give rise to a rather strong illusion.

There is a caveat, however. Grammaticality judgments provide the most direct way of testing whether participants experience a grammatical illusion, but they are not without problems. This holds in particular when judgments must be given under time pressure, as in the experiments of Bader et al. (2003). Without further evidence, it cannot be excluded that the grammatical illusion found by Bader et al. was caused by the strict timing conditions imposed by the procedure of speeded grammaticality judgments. In order to address this issue, Experiment 1 replicates one experiment of Bader et al. (2003) using a judgment procedure that neither limits the time to process a sentence nor the time for giving a judgment.

An even more serious issue was brought about by Vasishth et al. (2010) who investigated the missing-VP effect in both English and German by recording reading times. A German example from Vasishth et al. (2010) is shown in (13).

(13) [$_{S1}$ Der Anwalt, [$_{S2}$ den der Zeuge, [$_{S3}$ den der Spion [$_{VP3}$ betrachtete,]] [$_{VP2}$ schnitt,]] [$_{VP1}$ überzeugte den Richter.]]
the lawyer who the witness who the spy watched avoided convinced the judge
"The lawyer that the witness that the spy watched avoided convinced the judge."

The study included complete sentences as in (13) as well as incomplete sentences in which the intermediate verb [= *schnitt* in (13)] was missing. In a self-paced reading experiment and in an eye-tracking experiment, Vasishth and colleagues found longer reading times for the last verb (*überzeugte*) and the following NP in incomplete sentences compared to complete sentences. For English, in contrast, Vasishth and colleagues found the opposite pattern. Reading times for the last verb were longer in complete sentences than in incomplete sentences. The authors take the reading time increase in German as indicating that their participants noticed the ungrammaticality. Based on the crosslinguistic difference between German and English, Vasishth et al. conclude that the German reader's parser is more adapted to keeping track of upcoming verbs due to the verb-final nature of German.

Using again a speeded grammaticality judgment procedure, Bader (2015) found evidence for a missing-VP effect in sentences structurally similar to those investigated by Vasishth et al. (2010). When taking the whole literature into account, we arrive at the generalization that a missing-VP effect was found for German when using the method of speeded grammaticality judgments but not when using reading time measures. Experiment 2 therefore uses a self-paced reading procedure that does not require any grammaticality judgments at all in order to test whether the missing-VP effect also occurs under more natural reading conditions for the kind of sentences for which only judgment data are available so far.

5. Experiment 1

Experiment 1 has two aims. The first aim concerns the question of whether the illusion of grammaticality caused by missing-VP sentences also occurs when participants are not set under time pressure. To answer this question, Experiment 1 obtained grammaticality judgments without time limits on either reading a sentence or judging its grammaticality.

If the missing-VP effect is indeed independent of time constraints on reading and judging a sentence, the next question is whether we can replicate the finding of Bader et al. (2003) that a grammatical illusion can arise not only when VP2 is omitted but also when VP1 is omitted. In the prior literature, sentences with a missing VP1 were rarely investigated after Gibson and Thomas (1999) found that such sentences are rated as highly complex. In Bader et al. (2003), the missing-VP1 effect was restricted to sentences in which S1 is an embedded clause. The second aim of Experiment 1 is therefore to examine whether a missing-VP1 effect arises and whether it depends on the clause type of the corresponding S1.

To test this question, Experiment 1 adopts the design and materials of Experiment 2 in Bader et al. (2003). Experiment 1 varies the clause type of S1 such that S1 is either an embedded complement clause as in (14) or a main clause as in (15). In addition, the experiment varies whether VP1 or VP2 is omitted as indicated in (14) and (15) by crossing.

Two subprocesses within the human parser are crucially involved when sentences with a missing VP elicit a grammatical illusion. First, either S1 or S2 is retrieved as integration site for the final VP. Second, the resulting structure is accepted as grammatical despite the lack of a VP. When S1 is selected for integration, a missing-VP effect arises when the lack of VP2 goes unnoticed. When S2 is selected for integration, a missing-VP effect arises when the lack of VP1 goes unnoticed. The clause type of S1 could influence both the likelihood of retrieving the wrong attachment site and the likelihood of noticing the lack of a VP. It will thereby determine the probability that a missing-VP effect is observed.

Why should clause type of S1 matter? When S1 is a main clause, it is the first clause and might benefit from primacy effects as observed in the literature on memory retrieval (for a recent overview, see Knoedler et al., 1999). Adding a level

(14) *Relative clauses within a complement clause*

Ich habe gehört
I have heard
S1 dass seit heute Mittag die Praktikantin,
that since today noon the intern
S2 die den Systemabsturz,
who the system_crash
S3 der die Technikerin für etliche Stunden beschäftigt hatte,
which the engineer for several ours occupied had
S2 [$_{VP2}$ ~~*verursacht*~~ ~~*hat*~~],
caused has
S1 [$_{VP1}$ ~~*verschwunden*~~ ~~*ist*~~.]
disappeared.PTCP is

Complete: "I have heard that the intern who caused the system crash which occupied the engineer for several hours disappeared since noon."
Missing VP2: "I have heard that the intern who Δ the system crash which occupied the engineer for several hours disappeared since noon."
Missing VP1: "I have heard that the intern who caused the system crash which occupied the engineer for several hours Δ since noon."

(15) *Relative clauses within a main clause*

S1 Seit heute Mittag ist die Praktikantin,
Since today noon is the intern
S2 die den Systemabsturz,
who the system_crash
S3 der die Technikerin für etliche Stunden beschäftigt hatte,
who the engineer for several hours occupied had
S2 [$_{VP2}$ ~~*verursacht*~~ ~~*hat*~~],
caused has
S1 [$_{VP1}$ ~~*verschwunden*~~.]
disappeared.PTCP

Complete: "Since noon, the intern who caused the system crash which occupied the engineer for several hours is missing."
Missing VP2: "Since noon, the intern who Δ the system crash which occupied the engineer for several hours is missing."
Missing VP1: "Since noon, the intern who caused the system crash which occupied the engineer for several hours is Δ."

of embedding changes the accessibility of S1. As an embedded clause, S1 is no longer the first clause but occurs in an intermediate position in the sequence of clauses. S2, in contrast, occurs always in an intermediate position between at least two clauses, namely S1 and S3 regardless of the type of S1. A primacy advantage of S1 in sentences in which S1 is a main clause could affect the processing of missing-VP sentences in two ways: First, it might increase the probability of integrating the final VP into S1 and thus decrease the probability of integrating it into S2. Second, in case the final VP was integrated into S2, the primacy advantage might increase the probability of detecting that S1 is missing a verb. These two possible consequences of the increased salience of S1 do not exclude each other. Both could jointly prevent a missing-VP effect in sentences in which S1 is a main clause and VP1 is missing. In these sentences, integration of VP2 into S1 results in a syntactic and semantic conflict, which prevents the illusion of grammaticality. Attachment to S2 would leave S1 with a missing VP, which will be noted thanks to the salience of S1. Thus, primacy predicts that the likelihood of a grammatical illusion in missing-VP2 sentences depends on the level of embedding of S1.

In missing-VP2 sentences, a grammatical illusion arises when the remaining VP1 is correctly integrated into S1 and the lack of a VP in S2 goes unnoticed, or when VP1 is integrated into S2 and the lack of a VP in S1 goes unnoticed. Primacy effects might increase the chance of S1 integration and thereby increase the likelihood of a grammatical illusion (under the assumption that the likelihood of detecting a missing VP in S2 is independent of the status of S1). But at the same time, primacy would increase the chance of detecting a missing VP in S1 when VP2 is correctly integrated into S2. Taken together, primacy predicts a lower rate of grammatical illusions in sentences in which S1 is a main clause and VP1 is missing.

5.1. Method

5.1.1. Participants

Twenty-four students at the University of Konstanz participated in Experiment 1. In this and the following experiment, all participants were native speakers of German and were naive with respect to the purpose of the experiment. They were either paid or received course credit for participation in the experiment.

5.1.2. Materials

The materials for Experiment 1 consisted of 30 sentences that were taken from Bader et al. (2003). Each sentence appeared in six versions according to the two factors Clause Type and Structure. The factor Clause Type varied the type of the matrix clause of the higher relative clause. This was either an embedded complement clause as in (14) or a main clause as in (15). The factor Structure manipulated whether the sentence was complete or not. In case it was not complete, either VP2 or VP1 was missing as indicated in (14) and (15) by crossing.

Sentences in the condition "main clause" consisted of three clauses: a main clause (S1), a relative clause (S2) center-embedded into the main clause, and a second relative clause (S3) center-embedded into the first relative clause. All main clauses started with an adverbial followed by the finite auxiliary and the subject NP. This subject NP was modified by the first relative clause. This relative clause was a subject-initial relative clause whose object NP was modified by the second relative clause. Each relative clause ended in a lexical verb followed by an auxiliary whereas the main clause ended in a lexical verb only because the main-clause auxiliary occurred already in the second position of the sentence. Sentences in the condition "embedded clause" contained one more level of embedding and thus consisted of four clauses: a short main clause, a complement clause (S1) and two center-embedded relative clauses (S2 and S3). The short main clause always preceded the complement clause. In complete sentences, all three verbs were present. In missing-VP sentences, either VP2 (lexical verb and auxiliary) or VP1 [lexical verb and auxiliary in the condition "embedded clause" and just lexical verb in the condition "main cause," in which the auxiliary appeared in the main clause, cf. (15)] was missing. The lexical verbs in VP1 and VP2 were always compatible with an animate subject and insofar compatible with both S1 and S2. However, their syntactic properties prevent them from being interchangeable: V1 was an intransitive verb while V2 was transitive.

The sentences were distributed across six lists using a Latin square design. Each list contained only a single version of each sentence and an equal number of sentences in each condition. The experimental lists were interspersed in a list of about 260 filler sentences for Experiment 1. The majority of filler sentences was from unrelated experiments. Each participant saw only one list.

5.1.3. Procedure

Participants received a questionnaire on which the experimental sentences were printed. They were asked to judge the grammaticality of each sentence on the questionnaire by marking one of the two options "grammatical" or "ungrammatical" printed beneath each sentence. Participants could spend as much time as they wanted on reading the sentences and giving their judgments. On average, they needed about 45–50 min to complete the questionnaire.

5.2. Results

For each participant and item, we recorded the grammaticality judgment. **Table 1** shows the results in terms of acceptance rates. All statistical analyses reported in this paper were computed using the statistics software R, version 2.14.2 (R Development Core Team, 2012). Responses were analyzed by means of linear mixed-effects logistic regression using the R-package *lme4* (Bates and Maechler, 2010). Forward difference coding was used for the experimental factors. That is, they were coded in such a way that all contrasts tested whether the means of adjacent factor levels were significant. Contrasts were specified as follows. For the factor Clause Type, the mean results in the condition "main clause" are contrasted with the mean results in the condition "embedded clause." For the factor Structure, two contrasts were defined. The first one compares complete sentences to sentences with a missing VP2 and the second one compares sentences with a missing VP2 to sentences with a missing VP1. Since not all possible contrasts can be tested within one model, we chose the contrasts such that the condition with the highest acceptance rates (complete sentences) is compared to the condition with intermediate acceptance rates (missing VP2), which in turn is compared to the condition with the lowest acceptance rates (missing VP1). If both contrasts turn out to be significant, we can conclude that the remaining contrast (complete sentences vs. sentences missing VP1) is significant as well. We included participants and items as crossed random effects. Following the advice given in Barr et al. (2013), we first computed a model containing the full factorial design in the random slopes. Since this model did not converge, we dropped the interaction term from the random sentence factor, which resulted in a converging model. For each contrast, **Table 2** shows the estimate, the standard error, the resulting z-value and the corresponding p-value.

The factor Clause Type was significant, with sentences in the condition "embedded clause" being judged as grammatical more often than sentences in the condition "main clause" (52 vs. 41%). The two contrasts of the factor Structure were also significant. Complete sentences received higher acceptance rates than missing-VP2 sentences (81 vs. 37%) which in turn received higher acceptance rates than missing-VP1 sentences (37 vs. 22%). Of the two interactions, only the one involving the second contrast of the factor Structure was significant. This reflects the

TABLE 1 | Acceptance rates in Experiment 1.

	Complete	Missing V2	Missing V1
Main clause	81 (5.6)	33 (7.1)	10 (4.0)
Embedded clause	81 (5.4)	41 (8.5)	33 (7.5)

Standard error (by participants) is given in parentheses.

TABLE 2 | Mixed-effects model for the judgment results of Experiment 1.

Contrast	Estimate	Std. error	*z*-value	Pr(>|*z*|)
Clause Type	−3.585	0.834	−4.297	< 0.001
Structure 1	−1.311	0.524	−2.502	< 0.05
Structure 2	0.903	0.435	2.083	< 0.05
Structure 1 × Clause Type	0.419	0.721	0.581	0.562
Structure 2 × Clause Type	1.834	0.771	2.377	< 0.05

finding that for complete and missing-VP2 sentences, the factor Clause Type did not have much of an effect whereas for missing-VP1 sentences, sentences in the condition "embedded clause" were more often accepted as grammatical than sentences in the condition "main clause."

Pairwise comparison were computed in order to explore the interaction more closely. Sentences with a missing VP1 received significantly fewer grammatical judgments than sentences with a missing VP2 when S1 was a main clause (33 vs. 10%, $z = 4.71$, $p < 0.001$). In sentences with an embedded S1, in contrast, the difference between missing VP2 and missing VP1 was not significant (41 vs. 33%, $z = 1.49$, $p = 0.14$). Furthermore, sentences with a missing VP1 were judged as grammatical significantly less often when S1 was a main clause than when S1 was an embedded clause (10 vs. 33%, $z = 4.82$, $p < 0.001$). For sentences with a missing VP2, the contrast between main and embedded S1 clause was marginally significant (33 vs. 41%, $z = 1.65$, $p = 0.10$).

5.3. Discussion

Experiment 1 has yielded two major results. First of all, although participants had unlimited time for reading and judging a sentence, sentences in which a VP was missing were accepted as grammatical in a substantial number of cases. Though the observed missing-VP effects were somewhat weaker in the current experiment than in the corresponding speeded grammaticality judgment experiment from Bader et al. (2003), the questionnaire results closely replicate the pattern from the speeded grammaticality judgments study (correlation coefficient for grand means: $r = 0.94$, $p < 0.01$; for items means per condition: $r = 0.31$, $p < 0.001$). Moreover, the average acceptance rate for missing-VP sentences in the questionnaire study was still 29% despite the lack of time pressure. The mean acceptance rate was even higher when we excluded main clauses with a missing VP1. For these sentences, no missing-VP effect was expected, and in accordance with this expectation, they were rejected as ungrammatical in about 90% of the time. The finding that the other missing-VP sentences are accepted as grammatical to a substantial degree despite the lack of time constraints corroborates the existence of the missing-VP effect in German. Given the interaction of Clause Type and Structure, the missing-VP effect cannot be attributed to an undifferentiated tendency to accept sentences of this type as grammatical. In sum, participants experience a grammatical illusion with missing-VP sentences not only when put under time pressure, but also when they have as much time as they want. The possibility to reread sentences and to engage in deliberate reasoning reduces the missing-VP effect, but it does not eliminate it.

The second major finding of Experiment 1 is that a missing-VP effect for VP2 is independent of clause type whereas a missing-VP effect for VP1 is restricted to sentences in which S1 is an embedded clause. Sentences lacking a VP in their main clause were reliably rejected as ungrammatical with a 90% rejection rate. This finding is compatible with the proposal that primacy effects make it easier to spot the lack of a VP in the first clause, i.e., the main clause, of a complex sentence.

For sentences with a missing VP2, clause type had no effect. The lack of a difference between main and embedded clauses indicates that properties of S1 did not affect the probability of detecting that VP2 was missing. This is expected under the primacy perspective since S2 is always an embedded clause. Promoting S1 to a main clause brings S1 into first position but leaves S2 in an intermediate position.

The clause type of S1 had also no effect for complete sentences. The finding of identical acceptance rates for main and embedded matrix clauses confirms earlier claims that clausal embedding does not cause increased processing costs as long as clauses are embedded in sentence final position (see Gibson, 1998; Gibson and Thomas, 1999). Erroneous integration of VP2 or VP1 into the wrong clause might occur from time to time but is easily detected because of the other VP. If VP2 is erroneously attached to S1, the subsequent verb (VP1) signals the error. An attempt to attach VP1 to S2 will fail because the verb slot of S2 is already occupied by VP2.

In sum, Experiment 1 has shown that the grammatical illusion caused by a missing VP2 is a robust phenomenon which is not affected by whether S1 is a main clause or an embedded clause. A missing VP1, in contrast, causes a grammatical illusion only when S1 is an embedded clause. An interesting question raised by these findings is whether the same holds for English. Since our account did not appeal to any special properties of German, it predicts that a missing VP1 should also cause a grammatical illusion in an English sentence as in (16), which is identical to the original example of Gibson and Thomas (1999) with the exception that S1 is now an embedded clause.

(16) I believe that the ancient manuscript that the graduate student who the new card catalog [$_{VP3}$ had confused a great deal] [$_{VP2}$ was studying in the library].

6. Experiment 2

In contrast to the SVO languages English and French, all experiments demonstrating a missing-VP effect in German relied on some form of grammaticality judgments, either under time constrained conditions (Bader et al., 2003; Bader, 2015) or without time limitations (Experiment 1). The only study that investigated the missing-VP effect in German using on-line reading measures (selfpaced reading and eye tracking) is Vasishth et al. (2010), and this study failed to find evidence for a grammatical illusion in German whereas it found such evidence for English. Based on the current evidence, it can thus not be excluded that in an SOV language like German a missing-VP

effect only occurs when explicit grammaticality judgments are required but not when participants simply process sentences for the purpose of comprehension.

A different possibility is suggested by the results of Bader (2015). These results show that the likelihood of a missing-VP effect in German is modulated by the syntactic configuration in which the center-embedded relative clauses occur. The sentences from Vasishth et al. (2010) contain the relative clauses in the initial position of the main clause whereas the sentences in the current study contain the relative clauses in a sentence-medial position. Using the same speeded grammaticality judgment task as Bader et al. (2003), Bader (2015) found a higher acceptance rate for missing-VP sentences in the latter configuration. The lack of a reading time advantage for missing-VP sentences in the experiments of Vasishth et al. (2010) may thus be due to a weak missing-VP effect in sentences in which the relative clauses belong to a sentence initial NP. If so, we expect that reading time evidence for a missing-VP effect can be found for sentences for which the missing-VP effect is more likely to occur. Experiment 2 tests this prediction by collecting reading times for sentences that, like the sentences investigated in Experiment 1, contain the relative clauses in a sentence medial position.

In the sentences in Experiment 2, S1 is an embedded clause. The sentences are thus structurally similar to the sentences in the "embedded clause" condition of Experiment 1. An example is given in (17). Incomplete sentences were derived by dropping VP2, as indicated by crossing in (17). The subject of S2 is either a singular or a plural NP. The verb of S2, which is only present in complete sentences, is accordingly either a singular or a plural verb. The verb of S1 is always present and always marked for singular in agreement with the subject of S1.

(17) *Example sentences of Experiment 2*

	Ich	glaube					
	I	think					
S1	dass	man	(den	**Direktor**,	/ die	**Direktoren**,)	
	that	one	the	principal.SG	the	principals.PL	
S2	(**der**	/ **die**)	den	Schulrat,			
	who.SG	who.PL	the	schools.inspector			
S3	der	das	Projekt	absegnen	soll,		
	who.SG	the	project	approve	should		
S2	~~alarmiert~~	(~~hat,~~	/	~~haben,~~)			
	alarmed	has.SG		have.PL			
S1	belogen	**hat**,					
	lied.to	has					
	um	von	dem	eigentlichen	Problem	abzulenken	
	for	from	the	actual	problem	distract	

"I think that the principal who ~~alarmed~~ the schools inspector who was supposed to approve the project was lied to in order to distract him from the actual problem."

If German was immune to the missing-VP effect, as claimed by Vasishth et al. (2010), reading times should be longer in incomplete sentences compared to complete sentences. The increase should start after the final verb since only then does it become evident that no further verbs are coming and thus a verb is missing. If, on the other hand, the missing-VP effect is present in German too, longer reading times are predicted for the final verb in complete sentences. This prediction is made both by the Pruning Hypothesis and by the Discrimination Hypothesis. Hence, the purpose of Experiment 2 is not to decide between the two hypotheses. Instead, the aim is more modest. The main objective of Experiment 2 is to test whether the missing-VP effect in German can be observed in online measures like reading times at all.

In addition, Experiment 2 tests whether the effect of number reported by Bader et al. (2003) also occurs in on-line reading times. In complete sentences with a plural S2 subject and therefore a plural verb V2, the attempt to integrate V2 into S1 results in a fleeting agreement violation which should increase reading times. Moreover, the integration of the actual verb of S1 then becomes difficult because the verb slot of S1 is already filled by the preceding verb. In incomplete sentences with a plural S2 subject, integration of the final verb, which is always singular, into S2 results in an agreement violation.

6.1. Method

6.1.1. Participants

Twenty-four students at the University of Konstanz participated in Experiment 2. They were paid for participation or received course credit.

6.1.2. Materials

We constructed 20 sentences each in four versions. An example is given in (17). All sentences started with a short main clause followed by a complement clause introduced by the complementizer *dass* ("that"). This complement clause (S1) contained an indefinite pronoun as subject and a definite NP as the object followed by a relative clause (S2) modifying the object. The relative clause contained another relative clause (S3), again modifying the object. Due to the clause-final position of verbs in embedded clauses in German, the verbs for S3, S2, and S1 occur in a row after the object of S3. As before, the final

position of each clause was filled by a verb cluster consisting of a lexical verb and an auxiliary. The final verb was followed by an adjunct clause in order to minimize wrap-up effects and to provide space for potential spillover effects. Two factors were fully crossed resulting in four conditions. The factor Structure varied whether the sentences were complete or incomplete; in incomplete sentences the intermediate verb (VP2) was omitted. The factor S2-Subject varied the number specification of the head noun of the higher relative clause and thereby of the subject of this relative clause. If present, VP2 matched the S2 subject in number.

For each sentence, we designed a question that probed understanding of the sentence. The example in (18) gives the probe question for (17).

(18) Hat der Direktor falsche Informationen erhalten?
has the principal wrong information received
"Did the principal receive wrong information?"

As in the example, all probe questions asked for an event involving the subject of S2 which is at the same time the object of S1. Low attachment of VP1 and subsequent interpretation of V1 as the verb of S2 would result in a wrong answer to the probe question. Half of the questions required a positive answer, the other half required a negative answer.

The experimental stimuli were distributed over four lists using a Latin square design. Each participant saw only one list. The order of items in a list was pseudo-randomized for each participant individually. In addition to the experimental stimuli, an experimental session included 94 filler sentences. Most of them served as experimental stimuli in unrelated experiments. Filler sentences were always grammatical and covered a variety of syntactic constructions. The order of filler sentences and experimental items was arranged in such a way that no two experimental items followed each other.

6.1.3. Procedure

Experiment 2 used a word-by-word non-cumulative self-paced reading procedure. Participants read sentences on a computer screen using a moving window display in which all non-space characters of the sentence were initially replaced by underlines (Just et al., 1982). Participants pressed a key on the keyboard to see each new word of the sentence. On each key press, a new word was uncovered and the previous word was again replaced by underlines. The time between successive key presses was recorded automatically. Once the last word of the sentence had been reached, pressing the key again cleared the screen and revealed the word "Frage" ("question"). The next key press produced the question which had to be answered by pushing the "j"-key for "Ja" ("yes") or the "n"-key for "Nein" ("no"). Participants received no feedback for their answers. To become acquainted with the procedure, participants read four training sentences before the experiment started.

6.2. Results

Despite the complexity of the sentences, participants answered probe questions with an overall accuracy of 87%. There were only minimal differences between conditions (range 84–89%). A statistical analysis using a mixed effects model did not find significant effects.

Reading times >2000 ms were removed from the analysis. This affected <1% of the data. The remaining mean reading times are summarized in **Table 3**. In accordance with Vasishth et al. (2010), we log-transformed raw reading times before fitting linear mixed effects models to the data. Contrasts were coded as follows. The contrast for the factor S2-Subject compares sentences with a singular S2 subject to sentences with a plural S2 subject. The contrast for the factor Structure compares complete sentences to sentences missing the second verb cluster. Fixed effects results for the models are given in **Table 4**. All models reported in the table contain the full factorial design in the crossed random slopes for participants and items. Since degrees of freedom can only be estimated in linear mixed effects models (Baayen, 2008), we report estimates, standard errors and t-values but no p-values. An absolute t-value of 2 or greater indicates significance at the α-level 0.05. We also computed residual reading times (Ferreira and Clifton, 1986) and repeated all analyses; the results were similar as for the log-transformed raw reading times.

For VP3 and VP2, joint reading times for the lexical verb and the auxiliary are virtually identical across conditions (VP3 in sentences with a singular S2 subject: 946 ms, with plural S2 subject: 931 ms; VP2 in sentences with a singular S2 subject: 1029 ms, with plural S2 subject: 1038 ms). The statistical models indicate no significant effect. For VP1, however, reading times are longer in complete sentences (1066 ms in complete sentences, 953 in incomplete sentences). Reading times for individual words reveal that the effect occurs at the lexical verb (550 vs. 483 ms). Numerically, the effect is still visible at the auxiliary but no longer significant (513 vs. 477 ms). At the next word, the effect is gone. The factor S2-subject had no effect at all.

6.3. Discussion

The major finding of Experiment 2 is that reading times for the final verb were shorter in incomplete sentences compared to complete sentences. Thus, the missing-VP effect observed in prior judgment experiments occurs as well when participants only have to read for meaning. The difference between the current results and the results of Vasishth et al. (2010) can be attributed to structural differences between the respective sentence materials. As discussed above, the missing-VP effect is weaker when the relative clauses modify an NP in sentence initial position, as in the study of Vasishth et al. (2010). If readers experience a grammatical illusion in only a subset of trials, it may well be that any reading time advantage resulting from trials eliciting a grammatical illusion is offset by a reading time penalty for trials in which readers detect the ungrammaticality.

In contrast to the finding in Bader et al. (2003), the number manipulation had no effect in Experiment 2. We surmise that this difference reflects the fact that Experiment 3 of Bader et al. (2003), but not Experiment 2 of the current study, involved an explicit grammaticality judgment. Since no judgment was required in Experiment 2, the temporary ungrammaticality that might have arisen in conditions with a plural S2 subject

TABLE 3 | Mean reading times in experiment 2.

Subject S2	Structure	VP3	VP2	VP1			Post-VP1
				V1	Aux1	V1 + Aux1	
Singular subject in S2	Complete	940	1029	544	510	1058	852
Singular subject in S2	Missing	952		481	482	958	829
Plural subject in S2	Complete	928	1038	556	516	1074	844
Plural subject in S2	Missing	934		484	476	951	872

TABLE 4 | Fixed effects of mixed-effect models for reading times in experiment 2.

	Structure			Subject S2			Structure × Subject S2		
	Est.	SE	*t*	Est.	SE	*t*	Est.	SE	*t*
VP3	−0.008	0.02	−0.34	0.001	0.03	0.03	−0.010	0.05	−0.20
VP2				0.005	0.03	0.13			
VP1	0.090	0.03	3.12	−0.016	0.03	−0.55	−0.003	0.05	−0.06
V1	0.097	0.04	2.76	−0.030	0.04	−0.83	−0.005	0.06	−0.08
Aux1	0.047	0.03	1.45	−0.005	0.03	−0.14	−0.004	0.06	−0.06
Post-VP1	0.005	0.02	0.20	−0.019	0.02	−0.82	0.050	0.04	1.21

could be internally repaired by the parser without any overtly observable effect.

7. General Discussion

This paper has presented an interference account of the missing-VP effect, that is, the observation that sentences in which a VP is missing can give rise to an illusion of grammaticality. This account is based on experimental investigations of the missing-VP effect in German. While prior reports of the missing-VP effect in German relied on speeded grammaticality judgments, the experiments reported in this paper show that the missing-VP effect is rather robust with regard to the experimental procedure. In particular, the missing-VP effect is so strong that it also occurs when participants have to judge sentences without time pressure, and it occurs as well when participants simply have to read sentences for meaning.

The finding of missing-VP effects in German points to the cross-linguistic generality of this kind of grammatical illusion. It is not confined to languages with SVO order but is found in languages with SOV order too. This suggests that the source of the effect is not language-specific but results from more general mechanisms that apply across languages. Interference during cue-based retrieval is a promising candidate for such a general mechanism. It provides a unified account of how sentences with double center-embedding—whether complete or incomplete—are processed. In sentences with double center-embedding, the parser faces two competing attachment sites for the second verb, as illustrated in (19).

(19) $[_{S1}$ NP1 … $[_{S2}$ NP2 … $[_{S3}$ NP3 … VP3] … **VP2** … (VP1)

Processing of NP1 causes the creation of a sentence node and thereby leads to the expectation of a verb. Similar expectations result from the processing of NP2 and NP3. Integration of VP3 fills the open verb slot of S3. After processing of S3, the next verb generates retrieval cues that call for a sentence with an open verb slot. Since both S1 and S2 fit this cue, interference arises and hampers the correct integration of the second verb into S2. As a result, the second verb is occasionally integrated into the wrong clause, namely S1, and thereby analyzed as VP1. Erroneous integration of VP2 into S1 entails difficulties for the subsequent integration of VP1 in complete sentences and it contributes to the illusion of grammaticality in missing-VP sentences. To make the illusion perfect, the lack of lexical material in the VP2 slot must go unnoticed. A failure to detect the missing VP is especially likely because the incomplete clause (S2) is no longer the current clause as soon as the parser returns to the higher clause, what it does when attaching the final verb to S1. This reasoning also explains why the status of S1 (main clause vs. embedded clause) had no effect for the likelihood of a missing-VP2 effect. Since the clause lacking VP2 is always an embedded clause, its processing must be completed when the last verb is integrated into the higher clause.

Since nothing is ever deleted according to our account, S1 and S2 are always available as attachment sites and therefore as targets for retrieval. The additional finding of a grammatical illusion when VP1 is missing indicates that the VP slot of S2 is retrieved for integration in some of the cases. In contrast to cases of a missing VP2, a grammatical illusion for a missing VP1 was observed only when S1 was an embedded clause but not when S1 was a main clause. We have argued that this finding is a primacy effect. When S1 is a main clause and thereby occurs in sentence initial position, the probability of erroneously attaching VP2 to it increases as does the probability of detecting that VP1 is missing in case VP2 has correctly been attached to S2. Taken together, this prevents the occurrence of a missing-VP effect for VP1 in main clauses.

Two alternatives to an interference-based account of the missing-VP effect are the resource-based account of Gibson and Thomas (1999) and the experience-based account of Christiansen and MacDonald (2009). The resource-based account of Gibson and Thomas (1999), which was already discussed above, is based on the idea that the parser has only a limited amount of resources available for storage and integration. Their Pruning Hypothesis proposes pruning as a last resort mechanism to free resources and thereby to avoid an overload of the parser. After deletion of VP2, the second verb can only be integrated into S1, creating the illusion of completeness in missing-VP sentences. The assumption of VP2-pruning is disconfirmed by the finding that omitting VP1 can lead to a missing-VP effect as well under certain circumstances. In addition, the Pruning Hypothesis is not attractive from a theoretical point of view. Pruning is a mechanism specific for situations with high memory load and has to be stipulated. Interference, on the other hand, is a general phenomenon that follows from cue-based retrieval. Similarity-based interference arises whenever two or more items in a memory representation are similar to each other. Interference can emanate from an item preceding the target item (proactive interference) or from an item following the target item (retroactive interference). Under the Discrimination Hypothesis, the missing-VP effect is an instance of proactive interference. Interference has been shown to be effective in explaining various phenomena in language comprehension (cf. van Dyke and Johns, 2012; Gordon and Lowder, 2012). We conclude that an interference-based explanation of the missing-VP effect is both empirically and conceptually more adequate than a resource-based explanation.

An experience-based account of the missing-VP effect was proposed by Christiansen and MacDonald (2009). This account draws on earlier work by Christiansen and Chater (1999) who proposed a connectionist model of recursion in natural language. This model is cast as a simple recurrent network (Elman, 1990) that learns from experience to predict the next word of a sentence from the words processed so far. Simulations by Christiansen and MacDonald (2009) show that when processing an English sentence with double center-embedding, the model expects only a single verb after it has encountered V3, as in a missing VP sentence, and not two verbs, as in a corresponding complete sentence. This approach was extended to German by Engelmann and Vasishth (2009). The model that they trained for German predicts that missing-VP sentences do not give rise to a grammatical illusion in German. Based on the experimental evidence from Vasishth et al. (2010), Engelmann and Vasishth (2009) conclude that an experience-based account of the missing-VP effect is superior to a memory-based account (e.g., the Pruning Hypothesis of Gibson and Thomas, 1999) because only the former account predicts that the missing-VP effect is present in English but absent in German.

With regard to the difference between SVO- and SOV-languages, the main thrust of the experience-based account has been succinctly summarized by Vasishth et al. (2010, p. 558): "One consequence of German head-finality is that—due to the relatively frequent occurrence of head-final structures—predictions of upcoming verbs may have more robust memory representations in German than in English. This could result in reduced susceptibility to forgetting the upcoming verb's prediction, even in the face of increased memory load." As the results of the present study show, this conclusion is premature. When presented with missing-VP sentences, native speakers of German experience a grammatical illusion as well. Furthermore, native speakers also produce such sentences from time to time. In an ongoing analysis of the deWaC corpus[5], we found a number of authentic missing-VP sentences. A small selection of such examples is provided in **Table 5**.

Such examples make two points. First, the missing-VP effect is not restricted to language comprehension but occurs in language production as well. Second, the missing-VP effect is not merely a laboratory phenomenon. Since this is evidence from German, we can conclude that the verb-final nature of German does not lead to memory structures that prevent the missing-VP effect from occurring. At face value, this contradicts experience-based accounts which have derived the absence of a missing-VP effect in German from corpus-based simulations. However, drawing strong conclusions at this point would be premature. For example, the training corpus used by Engelmann and Vasishth (2009) for their simulation is not described in detail, which leaves the possibility that their training input did not include all relevant syntactic configurations. Additional simulations are necessary in

[5]DeWac is the German part of Wacky, a family of large corpora built by web crawling (Baroni et al., 2009). DeWac contains 1.7 billion tokens of text which is POS tagged and lemmatized (using TreeTagger). Partial results from an ongoing corpus study of complete and incomplete doubly embedded relative clauses in the DeWac corpus can be found in Bader (2015).

TABLE 5 | Authentic examples of the missing-VP effect from the deWac corpus.

| | | |

order to address the issues raised above, but this is beyond the scope of the present paper.

The seeming contradiction between the evidence presented by Vasishth et al. (2010) on the one hand and the evidence provided by Bader et al. (2003) and the analysis of the deWac corpus on the other hand was addressed by Bader (2015). Based on corpus evidence and on evidence from yet another experiment using the method of speeded grammaticality judgments, Bader (2015) showed that the strength of the missing-VP effect varies with the syntactic position occupied by the doubly center-embedded relative clause. The probability that a missing-VP effect occurs is smaller when the relative clauses occupy the initial position of a main clause, as in the sentences investigated by Vasishth et al. (2010) [see (13)], than when they are contained within the lower part of the clause, whether this is an embedded clause as in (14) or a main clause as in (15).

In sum, the results of the present study confirm the existence of the missing-VP effect in German and thereby show that the occurrence of this grammatical illusion does not depend on whether a language is SVO or SOV. The results challenge resource-based and experience-based accounts of the effect, but they lend further evidence to interference-based accounts of human parsing. In particular, the missing-VP effect adds to the existing evidence for proactive interference during language comprehension and supports cue-based parsing architectures.

Acknowledgments

Special thanks to the reviewers (Clinton L. Johns and Manuel Gimenes) for their valuable comments and suggestions. We acknowledge the support of the Open Access Publication Fund of the University of Potsdam.

References

Baayen, R. H. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge: Cambridge University Press.

Bader, M. (2015). "Complex center embedding in German - the effect of sentence position," in *Firm Foundations: Quantitative Approaches to Grammar and Grammatical Change*, eds Y. Versley and S. Featherston (Berlin: de Gruyter). (in press).

Bader, M., Bayer, J., and Häussler, J. (2003). "Explorations of center embedding and missing VPs," in *Poster, 16th CUNY Conference on Sentence Processing* (Cambridge, MA: MIT).

Bader, M., Meng, M., and Bayer, J. (2000). Case and reanalysis. *J. Psycholinguist. Res.* 29, 37–52. doi: 10.1023/A:1005120422899

Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The WaCky Wide Web: a collection of very large linguistically processed web-crawled corpora. *Lang. Resour. Eval. J.* 23, 209–226. doi: 10.1007/s10579-009-9081-4

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: keep it maximal. *J. Mem. Lang.* 68, 255–278. doi: 10.1016/j.jml.2012.11.001

Bates, D., and Maechler, M. (2010). *lme4: Linear Mixed-Effects Models Using S4 Classes*. R package version 0.999999-0. Available online at: http://CRAN.R-project.org/package=lme4

Bever, T. G. (1970). "The cognitive basis for linguistic structure," in *Cognition and the Development of Language*, ed J. R. Hayes (New York, NY: Wiley and Sons), 279–362.

Bever, T. G. (1974). "The ascent of the specious, or there's a lot we don't know about mirrors," in *Explaining Linguistic Phenomena*, ed D. Cohen (Washington, DC: Hemisphere), 173–200.

Chomsky, N., and Miller, G. A. (1963). "Introduction to the formal analysis of natural languages," in *Handbook of Mathematical Psychology*, Vol. II, eds R. D. Luce, R. R. Bush, and E. Galanter (New York; London; Sydney: Wiley), 269–321.

Christiansen, M. H., and Chater, N. (1999). Toward a connectionist model of recursion in human linguistic performance. *Cogn. Sci.* 23, 157–205. doi: 10.1207/s15516709cog2302_2

Christiansen, M. H., and MacDonald, M. C. (2009). A usage-based approach to recursion in sentence processing. *Lang. Learn.* 59, 126–161. doi: 10.1111/j.1467-9922.2009.00538.x

Drenhaus, H., Saddy, D., and Frisch, S. (2005). "Processing negative polarity items: when negation comes through the backdoor," in *Linguistic Evidence. Empirical, Theoretical and Computational Perspectives*, eds M. Reis and S. Kepser (Berlin: de Gruyter), 145–165.

Elman, J. L. (1990). Finding structure in time. *Cogn. Sci.* 14, 179–211. doi: 10.1207/s15516709cog1402_1

Engelmann, F., and Vasishth, S. (2009). "Processing grammatical and ungrammatical center embeddings in English and German: a computational model," in *Proceedings of 9th International Conference on Cognitive Modeling* (Manchester), 240–245.

Erickson, T. D., and Mattson, M. E. (1981). From words to meaning: a semantic illusion. *J. Verbal Learn. Verbal Behav.* 20, 540–551. doi: 10.1016/S0022-5371(81)90165-1

Ferreira, F., and Clifton, C. Jr. (1986). The independence of syntactic processing. *J. Mem. Lang.* 25, 348–368. doi: 10.1016/0749-596X(86)90006-9

Foraker, S., and McElree, B. (2011). Comprehension of linguistic dependencies: speed-accuracy tradeoff evidence for direct-access retrieval from memory. *Lang. Linguist. Compass* 5, 764–783. doi: 10.1111/j.1749-818X.2011.00313.x

Frazier, L. (1985). "Syntactic complexity," in *Natural Language Parsing. Psychological, Computational and Theoretical Perspectives*, eds D. R. Dowty, L. Karttunen, and A. Zwicky (Cambridge: Cambridge University Press), 129–189.

Gibson, E. (1998). Linguistic complexity: locality of syntactic dependencies. *Cognition* 68, 1–76. doi: 10.1016/S0010-0277(98)00034-1

Gibson, E. (2000). "The dependency locality theory: a distance-based theory of linguistic complexity," in *Image, Language, Brain. Papers From the First Mind Articulation Project Symposium*, eds A. Marantz, Y. Miyashita, and W. O'Neil (Cambridge, MA: MIT Press), 95–126.

Gibson, E., and Thomas, J. (1999). Memory limitations and structural forgetting: the perception of complex ungrammatical sentences as grammatical. *Lang. Cogn. Process.* 14, 225–248. doi: 10.1080/016909699386293

Gimenes, M., Rigalleau, F., and Gaonac'h, D. (2009). When a missing verb makes a French sentence more acceptable. *Lang. Cogn. Process.* 24, 440–449. doi: 10.1080/01690960802193670

Gordon, P. C., and Lowder, M. W. (2012). Complex sentence processing: a review of theoretical perspectives on the comprehension of relative clauses. *Lang. Linguist. Compass* 6, 403–415. doi: 10.1002/lnc3.347

Just, M. A., and Carpenter, P. A. (1992). A capacity theory of comprehension: individual differences in working memory? *Psychol. Rev.* 99, 122–149.

Just, M. A., Carpenter, P. A., and Wolley, J. D. (1982). Paradigms and processes in reading comprehension. *J. Exp. Psychol. Gen.* 111, 228–238. doi: 10.1037/0033-295x.99.1.122

Karlsson, F. (2007). Constraints on multiple center-embedding of clauses. *J. Linguist.* 43, 365–392. doi: 10.1017/S0022226707004616

Kimball, J. (1973). Seven principles of surface structure parsing in natural language. *Cognition* 2, 15–48. doi: 10.1016/0010-0277(72)90028-5

Knoedler, A. J., Hellwig, K. A., and Neath, I. (1999). The shift from recency to primacy with increasing delay. *J. Exp. Psychol. Learn. Mem. Cogn.* 25:474. doi: 10.1037/0278-7393.25.2.474

Lewis, R., and Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cogn. Sci.* 29, 375–419. doi: 10.1207/s15516709cog0000_25

MacDonald, M. C., and Christiansen, M. H. (2002). Reassessing working memory: comment on Just and Carpenter (1992) and Waters and Caplan (1996). *Psychol. Rev.* 109, 35–54. doi: 10.1037/0033-295X.109.1.35

McElree, B. (2006). "Accessing recent events," in *Psychology of Learning and Motivation*, Vol. 46, ed B. H. Ross (San Diego, CA: Academic Press), 155–200.

McElree, B., and Dyer, L. (2013). "Beyond capacity: the role of memory processes in building linguistic structure in real time," in *Language Down the Garden-Path: The Biological and Cognitive Basis for Linguistic Structure*, eds M. Sanz, I. Laka, and M. K. Tanenhaus (Oxford: Oxford University Press), 229–240.

McElree, B., Foraker, S., and Dyer, L. (2003). Memory structures that subserve sentence comprehension. *J. Mem. Lang.* 48, 67–91. doi: 10.1016/S0749-596X(02)00515-6

Meng, M., and Bader, M. (2000). Ungrammaticality detection and garden-path strength: evidence for serial parsing. *Lang. Cogn. Process.* 15, 615–666. doi: 10.1080/016909600750040580

Miller, G. A., and Chomsky, N. (1963). "Finitary models of language users," in *Handbook of Mathematical Psychology*, Vol. II, eds R. D. Luce, R. R. Bush, and E. Galanter (New York; London; Sydney: Wiley), 419–490.

Miller, G. A., and Isard, S. (1964). Free recall of self-embedded English sentences. *Inf. Control* 7, 292–303. doi: 10.1016/S0019-9958(64)90310-9

Navon, D. (1984). Resources – a theoretical soup stone? *Psychol. Rev.* 91, 216–234. doi: 10.1037/0033-295X.91.2.216

Oberauer, K., and Kliegl, R. (2001). Beyond resources: formal models of complexity effects and age differences in working memory. *Eur. J. Cogn. Psychol.* 13, 187–215. doi: 10.1080/09541440042000278

Phillips, C., Wagers, M. W., and Lau, E. W. (2011). "Grammatical illusions and selective fallibility in real-time language comprehension," in *Experiments at the Interfaces*, ed J. T. Runner (Bingley: Emerald), 153–186.

R Development Core Team. (2012). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

van Dyke, J. A., and Johns, C. L. (2012). Memory interference as a determinant of language comprehension. *Lang. Linguist. Compass* 6, 193–211. doi: 10.1002/lnc3.330

Vasishth, S., Suckow, K., Lewis, R. L., and Kern, S. (2010). Short-term forgetting in sentence comprehension: crosslinguistic evidence from verb-final structures. *Lang. Cogn. Process.* 25, 533–567. doi: 10.1080/01690960903310587

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2015 Häussler and Bader. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 15 December 2015
doi: 10.3389/fpsyg.2015.01841

Listeners Exploit Syntactic Structure On-Line to Restrict Their Lexical Search to a Subclass of Verbs

Perrine Brusini[1,2]*, *Mélanie Brun*[2,3], *Isabelle Brunet*[2,4] and *Anne Christophe*[2,4]

[1] Language, Cognition and Development Lab, Cognitive Neuroscience Department, Scuola Internazionale Superiore di Studi Avanzati, Trieste, Italy, [2] Laboratoire de Sciences Cognitives et de Psycholinguistique, École des Hautes Études en Sciences Sociales (EHESS), Centre National de la Recherche Scientifique, École Normale Supérieure (ENS), Paris, France, [3] Laboratoire Psychologie de la Perception, Université Paris Descartes, Paris, France, [4] Département d'Etudes Cognitives, Ecole Normale Supérieure - PSL Research University, Paris, France

Many experiments have shown that listeners actively build expectations about up-coming words, rather than simply waiting for information to accumulate. The online construction of a syntactic structure is one of the cues that listeners may use to construct strong expectations about the possible words they will be exposed to. For example, speakers of verb-final languages use pre-verbal arguments to predict on-line the kind of arguments that are likely to occur next (e.g., Kamide, 2008, for a review). Although in SVO languages information about a verb's arguments typically follows the verb, some languages use pre-verbal object pronouns, potentially allowing listeners to build on-line expectations about the nature of the upcoming verb. For instance, if a pre-verbal direct object pronoun is heard, then the following verb has to be able to enter a transitive structure, thus excluding intransitive verbs. To test this, we used French, in which object pronouns have to appear pre-verbally, to investigate whether listeners use this cue to predict the occurrence of a transitive verb. In a word detection task, we measured the number of false alarms to sentences that contained a transitive verb whose first syllable was homophonous to the target monosyllabic verb (e.g., target "dort" /dɔʁ/ *to sleep* and false alarm verb "dorlote" /dɔʁlɔt/ *to cuddle*). The crucial comparison involved two sentence types, one without a pre-verbal object clitic, for which an intransitive verb was temporarily a plausible option (e.g., "Il dorlote" / He *cuddles*) and the other with a pre-verbal object clitic, that made the appearance of an intransitive verb impossible ("Il le dorlote" / *He cuddles it*). Results showed a lower rate of false alarms for sentences with a pre-verbal object pronoun (3%) compared to locally ambiguous sentences (about 20%). Participants rapidly incorporate information about a verb's argument structure to constrain lexical access to verbs that match the expected subcategorization frame.

Keywords: linguistic expectation, verb argument structure, lexical search, on-line syntactic structure construction

OPEN ACCESS

Edited by:
Colin Phillips,
University of Maryland, USA

Reviewed by:
Hugh Rabagliati,
University of Edinburgh, UK
Brian Dillon,
University of Massachusetts Amherst, USA

***Correspondence:**
Perrine Brusini
pbrusini@gmail.com

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: *27 August 2015*
Accepted: *13 November 2015*
Published: *15 December 2015*

Citation:
Brusini P, Brun M, Brunet I and Christophe A (2015) Listeners Exploit Syntactic Structure On-Line to Restrict Their Lexical Search to a Subclass of Verbs. Front. Psychol. 6:1841. doi: 10.3389/fpsyg.2015.01841

INTRODUCTION

To understand spoken sentences, listeners have to process speech sounds, recognize words and morphemes, and decode the syntactic structure of the sentence to recover its meaning. All of these complicated processes seem effortless and are performed in a very short amount of time (Pylkkänen and Marantz, 2003; Poeppel et al., 2008). One way to explain the speed with which spoken language

is processed is to suppose that the human language parser is able to exploit the context—both linguistic and non-linguistic—to compute expectations about upcoming material, thus reducing the number of possible options available at any given point in time and anticipating the processes it is likely to have to complete next (e.g., Levy, 2008; Gibson et al., 2013).

Many experiments have shown that listeners indeed use context to build linguistic expectations (see e.g., Kamide, 2008; Rayner, 2009, for reviews). The use of lexico-semantic knowledge to anticipate upcoming words was one of the first studied phenomenon: comprehenders were shown to use the beginning of a sentence to look faster to likely referents (e.g., Altmann and Kamide, 1999), to read predictable words faster (e.g., Frisson et al., 2005), and they also displayed smaller N400 responses to content words that were made more likely by their preceding contexts (e.g., Kutas and Hillyard, 1980; Federmeier, 2007; Xiang and Kuperberg, 2015). Once they've generated a prediction about a possible upcoming content word, listeners can also build specific expectations regarding the phonological shape of the article that precedes it (DeLong et al., 2005), or the gender of a preceding adjective or article (Wicha et al., 2004; Van Berkum et al., 2005; Foucart et al., 2015). In addition, they are able to integrate other kinds of information within their anticipatory processes, such as the prosodic/rhythmic pattern of a sentence (Brown et al., 2011) or the phonological patterns typical for specific syntactic categories (Farmer et al., 2006).

Syntactic structure *per se* should be another good candidate for activating anticipatory linguistic processes: Indeed many studies have observed that participants are able to extrapolate syntactic information on the fly to build expectations regarding an upcoming word or structure (e.g., Boland et al., 1990; Konieczny, 2000; Kamide et al., 2003; Boland, 2005; Hare et al., 2009; Levy and Keller, 2013), and modeling work shows that probabilistic parsers trained on linguistic corpora account for a whole range of human experimental data (e.g., Jurafsky, 1996; Hale, 2001; Levy, 2008). For instance, comprehenders exploit a verb's argument structure to expect specific kinds of arguments (e.g., Boland et al., 1990), and use the selectional constraints imposed by the verb to predict probable referents (i.e., listeners expect an eatable entity after hearing the verb "to eat," Altmann and Kamide, 1999). In a verb-final language such as Japanese, a verb's argument structure can be exploited even before the verb is heard, so that the presence of an indirect object, for instance a DP marked with a goal marker, leads the parser to search for a patient before hearing the corresponding DP (Kamide et al., 2003).

This ability to build an argument structure before hearing its head (the verb), might be a specific adaptation from head-final languages comprehenders, to cope with the fact that verbs systematically appear after their arguments, and avoid lengthy delays. It could also, however follow from a general property of the human parser, which would use all the information available to constrain its ongoing syntactic structure on-line. A very recent paper by Omaki and colleagues addressed this precise issue (Omaki et al., 2015). They exploited filler-gap dependency completion in object relative clauses, as in "the city that the author chatted regularly about was named after an explorer," and measured participants' surprise upon hearing the intransitive verb "chatted" which cannot take a direct object (the preposition "about" provides a slot as an indirect object for "the city" and makes the sentence grammatical in the end). If participants wait until they process the verb in order to posit an object position for that verb, then they should show no surprise upon hearing an intransitive verb (since they haven't attempted yet to assign "the city" to a direct object position), and should simply wait longer until they find a suitable slot for the DP "the city." Instead, in three experiments, Omaki et al. observed delayed reading (interpreted as delayed processing) upon encountering an intransitive verb in such a position. These results thus strongly suggest that in English too, a verb-medial language, comprehenders posit an argument structure even before they have processed the verb.

In this paper, we will address this same question through a different angle. Rather than looking for the effect of specific content words on the predictability of upcoming words (integrating them within the on-going syntactic structure), we focus on purely structural effects: namely, we wonder whether participants are able to exploit the syntactic structure they've heard so far—irrespective of the semantic content of the specific lexical items involved—in order to build expectations as to the type of word that is likely to occur next. Previous work on this topic has yielded somewhat mixed results: within the noun phrase, Dahan et al. (2000) have shown that after hearing a gender-marked article, listeners successfully reduce their lexical search to gender-matching referents. When ambiguous words of different syntactic categories are involved (e.g., noun/verb, or adjective/noun), some studies have found that both meanings of a homophone are initially activated (Tanenhaus et al., 1979) while others found that context allowed listeners to completely ignore the unintended meaning, when one member of the homophone pair was a function word (Shillcock and Bard, 1993, e.g., "would" vs. "wood"), when the preceding linguistic context marked syntactic constituent boundaries through phrasal prosody (Millotte et al., 2008; de Carvalho et al., 2015), and when the visual context led listeners to expect either an adjective or a noun, for pragmatic reasons (Magnuson et al., 2008). Here, we focus on the level of the verb phrase, and wonder whether listeners are able to use elements from the subcategorization frame of a verb in order to constrain lexical access to upcoming verbs.

To do so, we tested whether the presence of a preverbal direct object pronoun blocks the activation of intransitive verb candidates, or not. In French, as in other Romance languages, the direct object of a verb is pronominalized as a clitic accusative pronoun ("le," "la," or "les") that rises in the left periphery of the transitive verb (Kayne, 1991). For example the DP complement of the French verb "manger" *to eat*, "la souris" in "Le chat mange la souris" *The cat eats the mouse* moves to the left-periphery of the verb when pronominalized, as in "Le chat **la** mange" *the cat eats **it***. This situation resembles the one in head-final languages such as Japanese; however, whereas all direct objects are preverbal in Japanese, whether they are pronominalized or not, in French only pronoun direct objects are pre-verbal, while full DP objects

appear post-verbally. Thus, in French the presence of pre-verbal objects is not a standard configuration (although it is reasonably frequent). Consequently, if French listeners are able to integrate the object pronoun clitic into the syntactic structure, and deduce on-line that this utterance calls for a transitive verb, this will confirm Omaki et al.'s finding that the fast integration of pre-verbal arguments is a general ability of the human parser, rather than a specific adaptation from listeners of head-final languages where objects systematically appear pre-verbally. Additionally, such a result would enlarge the growing body of evidence that the human parser is processing a wide variety of available cues in order to anticipate the linguistic material that might follow. In contrast, if French listeners do not make use of pre-verbal object pronouns to anticipate transitive verbs, this will suggest that the ability to anticipate upcoming materials is fine-tuned to the specific properties of the language being processed, such that only features that are usually relevant will be put to use by the parser.

We investigated this question using a false alarm paradigm with adult French speakers. Subjects were instructed to respond as quickly and accurately as possible to an intransitive monosyllabic verb (e.g., "dormir" *to sleep*, that surfaces as "il dort" /**dɔʁ**/ *he sleeps* when conjugated in the 3rd person singular present tense). While half the sentences did contain the target verb, the other half did not contain it, but contained instead a multisyllabic transitive verb that started with the same first syllable (e.g., "dorloter" *to cuddle,* that surfaces as "il dorlote" /**dɔʁ**l**ɔt**/ when conjugated in the 3rd person singular present tense). The number of false alarms triggered by this multisyllabic verb was the measure of interest here. To test the impact of the under-construction syntactic structure upon lexical access, we inserted this multisyllabic catch verb in two kinds of sentences: In the first experimental condition, it appeared immediately after the pronoun subject, and was followed by an object DP (e.g., "elle dorlote son nounours" *she cuddles her teddybear*); thus, at the point when the verb was processed, the available information was still compatible with an intransitive verb (only the pronoun subject had been heard), and the intransitive target verb was thus a plausible option to continue this sentence (e.g., "elle dort toute la nuit" *she sleeps through the night*). We expected this condition to trigger a baseline amount of false alarms. In the second condition, the catch verb appeared after a pronoun subject and an object clitic (e.g., "elle le dorlote toute la nuit" *She cuddles it through the night*). In that case, when listeners heard the beginning of the verb, they had already heard the clitic object: if they spontaneously integrate this clitic object on-line to the syntactic structure they are building, they should be able to reject the intransitive target verb as a possible continuation for that sentence, and should therefore exhibit a very low proportion of false alarms, close to zero. If, in contrast, lexical access is primarily based on the available phonological information (here, the first syllable of the catch verb, which matches the target verb), together perhaps with a coarse syntactic information (e.g., that a verb is expected), then the proportion of false alarms triggered should be roughly equal in both conditions, irrespective of the fact that the second one contains a pre-verbal object clitic and the first one does not.

MATERIALS AND METHODS

Participants

Twenty-five native speakers of French took part in this experiment and were paid 5€ for their participation. Two additional participants were tested but their data were discarded from the final analysis because their hit rate was too low (<30%). This work was approved by the local ethics committee (Paris Ile de France III), and all participants signed an informed consent form.

Stimuli

Using the LEXIQUE 3.55 database (New et al., 2001), we selected 14 pairs of verbs consisting of an intransitive monosyllabic verb (or, more precisely, a verb that could not take a direct object), and a multisyllabic transitive verb whose first syllable was homophonous to the intransitive monosyllabic verb. For example, the verb "dormir" *to sleep* was paired with the verb "dorloter" *to cuddle.* While both "il dorlote…" *he cuddles…* and "il le dorlote" *he cuddles it* are grammatical structures, the sequence "*il le dort" *he sleeps it* is ungrammatical (see Supplementary Material for a complete list of the verbs). Only the intransitive verb of each pair was used as a target in the word detection task.

Each pair of verbs was used to build one or several quadruplets of sentences, for a total of 31 quadruplets. Each quadruplet contained two HIT sentences that actually contained the target monosyllabic intransitive verbs (e.g., "Quand il fait nuit, elle dort tranquillement," *During the night, she sleeps peacefully* and "Quand il fait nuit, elle dort dans son lit," *During the night, she sleeps in her bed*), as well as two False Alarm sentences that contained the multisyllabic transitive verb. One of these false alarm sentences contained a pre-verbal object pronoun ("le," "la," or "les" *it masc, fem, plural*): this created an ungrammatical context for an intransitive verb (as in "Quand il fait nuit, elle **la** dorlote plus," *During the night, she cuddles it more*). The other false alarm sentence was locally ambiguous, in that the verb immediately followed a subject DP (which could be a personal pronoun), such that both members of a verb pair, the transitive and the intransitive one, were compatible with the structure heard so far (e.g., "Quand il fait nuit, elle dorlote sa poupée," *During the night, she cuddles her doll*). Crucially, all sentences from a quadruplet contained the same sequence of words before the verb phrase ("Quand il fait nuit, elle…" in the examples): this ensures that the only pre-verbal cue that can be used to constrain lexical access to the verb is the object clitic (when it is present). In other words, if the false alarm rate is greater for locally ambiguous sentences than for non-ambiguous sentences (with an object clitic), this can only be due to the fast integration of the pronoun object clitic which makes subjects discard the possibility of encountering an intransitive verb.

To check whether or not there were acoustic/prosodic differences across conditions on the syllable homophonous with the target word, we measured the duration and F0 of the first syllable of the multisyllabic verbs of the false alarm sentences (e.g., "dor" from "dorlote"). We compared these values with a Wilcoxon rank test (since visual inspection showed that the

duration and F0 values were not normally distributed).The results of this analysis are presented in **Table 1**. We observed a marginally significant effect of duration, with FA_CLI sentences tending to have somewhat shorter first syllables than FA_AMB sentences (about 10 ms). It is unlikely that such a small difference would trigger a major difference in False Alarm rates between the two conditions.

The 31 quadruplets thus amounted to a total of 124 test sentences. Experimental sentences were recorded by an expert speaker (the last author) and marked at the onset of the critical verb using PRAAT (Boersma and Weenink, 2015). Each participant heard each of the 124 experimental sentences once within two blocks of 62 sentences: for each quadruplet, one block contained one HIT and one ambiguous False Alarm sentence (FA_AMB), while the other block contained the other HIT sentence and the non-ambiguous False Alarm sentence, which featured a pre-verbal clitic (FA_CLI). Each block contained roughly the same amount of FA_AMB and FA_CLI sentences. In total, a subject was exposed to 62 HIT sentences, 31 FA_CLI and 31 FA_AMB sentences. Within each block, the order of the sentences was pseudo-randomized so as to avoid sequences of five or more false alarm sentences.

Procedure

Each participant was tested individually. They were instructed to press the spacebar from a computer keyboard as soon as they could identify the target verb. A trial began with the visual presentation of the target word, always an intransitive verb written in the infinitive form (1.5 s), followed by a black screen with a white fixation cross (1 s), then a sentence was played (the auditory stimuli were stored at a sampling rate of 22,050 Hz and were presented through headphones). The trial ended 2.5 s after the subject's response or after the end of the auditory presentation (whichever came first), and a new trial began immediately. Response times were measured from the onset of the target word. Speed and accuracy were emphasized in the instructions. Before the experiment began, participants performed a short training. If they gave an incorrect or delayed response during training (more than 1 s response time), a warning message appeared on the screen asking them to correct or speed up their response (depending on the situation). The whole experiment was run using the Psychotoolbox of Matlab (Kleiner et al., 2007) and lasted about 15 min including a pause (of about 2 min) between blocks.

Analysis

Since the false alarm responses were categorical (0 for no response, 1 for a FA), we used a logit model to analyze whether false alarms were distributed differently between the FA_CLI and FA_AMB conditions. We ran a mixed model analysis using R 3.2 and the lme4 package (v 1.1-6, based on Bates and Sarkar, 2007). Each false alarm F_{isc}, for a given item i (where an item represents the pair of False Alarm sentences from the same quadruplet, i varied between 1 and 31) and a given subject s (between 1 and 25), in a given Condition c (FA_CLI vs. FA_AMB), was modeled via an intercept β_0 reflecting the baseline probability of making a false alarm, and a slope estimate β_1 of the predictor variable C (Condition), reflecting the impact of the context on the probability of making a false alarm (either a locally ambiguous context, FA_AMB condition, or non-ambiguous context with a pre-verbal clitic that makes an intransitive verb unlikely, FA_CLI condition). Since we used the maximal random effect structure (recommended by Barr et al., 2013), we also included by-subjects and by-items intercepts (S_{0s} and I_{0i} allowing the baseline to vary from β_0 by a fixed amount for each subject s and each item i) and slopes (S_{1s} and $I_{1i,}$ respectively, allowing each subject and item to deviate from the population slope $\beta_1$in their sensitivity to Condition). The categorical predictor Condition C was coded as 0 for the ambiguous context (FA_AMB) and 1 for the object pronoun context (FA_CLI). The resulting equation for the model, taking into account a normally distributed error for each observation, e_{is}, is the following:

$$\text{Logit}(\text{P}(F_{is} = 1)) = \beta_0 + S_{0s} + I_{0i} + (\beta_1 + S_{1s} + I_{1i}).\, C + e_{is} \quad (1)$$

β estimates are given in log-odds (the space in which the logit models are fitted). To compare the probabilities of making a false alarm across the two levels of C (Conditions: ambiguous context vs. non-ambiguous object clitic pronoun context), we computed the difference: $\text{P}(F_{is} = 1; C = 0$ i.e., FA_AMB) – $\text{P}(F_{is} = 1; C = 1$ i.e., FA_CLI) by taking the inverse logit of the right-hand side of Equation (1).

We computed Wald's Z statistic using the mixed model described above. This statistic tests whether the estimates are significantly different from 0. Hence the intercept corresponds to the probability of making a false alarm when participants are exposed to an ambiguous context, while the slope corresponds to the modification in the probability of making a false alarm when participants hear an object clitic before the verb.

TABLE 1 | Mean and standard error for the duration and F0 of the first syllable of the multisyllabic carrier verb from the false alarm sentences (e.g., "dor" in "dorlote").

	FA_CLI mean (std. error)	FA_AMB mean (std. error)	Wilcoxon rank test $Z(p)$
Duration (ms)	173 (9.63)	183.5 (9.79)	−1.86 ($p = 0.062$)
Mean F0 (vowel) (Hz)	245.4 (4.89)	255.8 (4.16)	−0.975 ($p > 0.3$)

RESULTS

The Hit rate was 90.8%, with an overall False Alarm rate of 11.6% (averaged across FA_CLI and FA_AMB conditions), showing that participants performed the task adequately. To assess whether or not French listeners quickly integrate the presence of a clitic object pronoun in order to compute the probability of occurrence of a transitive vs. intransitive target verb, we compared the false alarms produced by subjects when they were exposed to non-ambiguous FA_CLI sentences containing a clitic object (as in "Quand il fait nuit, elle la dorlote plus" *During the night, she cuddles it more*) to ambiguous FA_AMB sentences that did not contain a clitic object (as in "Quand il fait nuit, elle dorlote sa poupée," *During the night, she cuddles her doll*). The mean proportion of false alarm responses is plotted in **Figure 1**. As can

be seen, subjects made many more false alarms to ambiguous sentences, presenting a syntactic context that is appropriate for both the transitive and intransitive verbs (20% false alarms, range 3.23–41.94%, by participants), than to non-ambiguous sentences featuring a clitic object pronoun: sentences of this type only triggered 3% of false alarms (range: 0–9.68%). This result was confirmed by our mixed model analysis exhibiting a main effect of the predictor Condition ($\beta = -3.01$; $z = -4.35$; $p = 1.4\text{e-}05$) corresponding to a decrease of 0.16 in the probability to make a false alarm when the participant heard an object pronoun clitic (FA_CLI condition) relative to when there was no object pronoun (FA_AMB condition).

Thus, the probability that participants would be influenced by the sound similarity between the target verb (intransitive) and the verb that was actually present in the sentence (multisyllabic transitive), was largely reduced by the presence of the pre-verbal object clitic. As pointed out by a reviewer, every sentence which exhibited a pre-verbal clitic object did not contain the target. As a result, one may wonder whether the reduced rate of false alarms on FA_CLI sentences with a pre-verbal object did not result from participants learning, over the course of the experiments, that clitic objects signaled sentences without a target. If this were the case, the difference in False Alarm rates between conditions should start at zero and increase with time, as participants start using the strategy. To examine this possibility, we checked whether the difference in false alarm rates increased between the two experimental conditions, over the course of the experiment. **Table 2** shows the percentage of False Alarms, in both conditions, for each quarter of the experiment (first 31 trials, trials 32–62, trials 63–93, and 94–124).

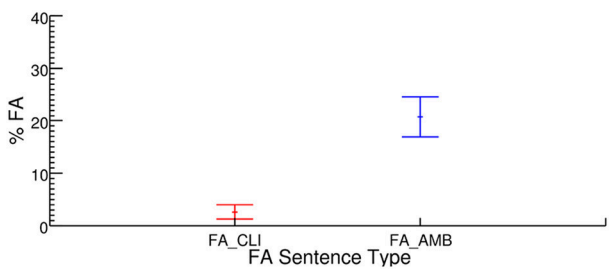


FIGURE 1 | Percentage of false alarms made by participants for each type of false alarm sentence presented, non-ambiguous FA sentences with a clitic (FA_CLI) in red (left-hand side) and ambiguous FA sentences (FA_AMB) in blue (right-hand side). Error bars represent the standard error of the mean (by participants).

TABLE 2 | Percentage of False Alarms, in both conditions (FA_AMB and FA_CLI), for the 4 quarters of the experiment.

	FA_AMB (%)	FA_CLI (%)
1st quarter	32	7.5
2nd quarter	23	2
3rd quarter	15	0.9
4th quarter	18	0.3

As can be seen, the difference in proportion of False Alarms between conditions does not increase with time. The only observable effect is a sharp decrease in overall False Alarm rate, as the experiment unfolds, suggesting that as participants became aware that they got caught on some of the false alarm sentences, they adopted a more conservative response bias. However, this pattern occurs for both types of False Alarms. In particular, there is already a massive difference between the FA_AMB and FA_CLI conditions in the first quarter of the experiment, suggesting that the lower rate of responses in the FA_CLI condition is present from the start, and therefore unlikely to be the result of a specific strategy developed as a function of the experiment (specifically, noticing that whenever a clitic is present the target is not there).

All in all, these results show that subjects dismiss the possibility that the target intransitive verb will occur, when they process the clitic object pronoun, an argument that is incompatible with the target intransitive verb.

DISCUSSION

We investigated the ability of French listeners to quickly compute the match or mismatch between the subcategory of an upcoming verb, and the presence or absence of a pre-verbal direct object pronoun. The logic here is that if French listeners are able to rapidly integrate the information conveyed by the object pronoun, they should be able to completely rule out an intransitive verb as a possible continuation of that sentence—since, by definition, an intransitive verb cannot take a direct object. We observed that participants' tendency to falsely detect a monosyllabic intransitive target verb was much higher when the multisyllabic carrier verb occurred in a syntactic context which was congruent with the target (e.g., "elle dorlote…" *she cuddles…*, FA_AMB condition), than when the carrier verb was preceded by an object pronoun (e.g., "elle le dorlote"/ *she cuddles it*, FA_CLI condition), making the syntactic context impossible for the target verb ("*elle le dort"/**she sleeps it*). This result shows that the participants integrated the clitic object pronoun on-line into the syntactic structure of the sentence, and inferred from this that the target intransitive verb would not follow.

This study confirms, with a different experimental technique, a different language, and in a different modality (listening vs. reading), the results obtained by Omaki et al. (2015) in English: They studied filler-gap dependency completion in object relative clauses and observed that processing of "chatted" was slowed down in a sentence such as "the city that the author chatted regularly about was named after an explorer," because there was a mismatch between the verb "chatted" which cannot take a direct object, and the implicit assumption that "the city" will be the direct object of the next encountered verb. Omaki et al. concluded from their series of three experiments that English-speaking comprehenders build argument structure before having heard the verb itself—just like Japanese-speaking comprehenders do, even though English is a verb-medial language. In their discussion, however they acknowledge the fact that their data are compatible with an alternative interpretation in which argument structure building does not occur ahead of the verb (p. 14). Under that alternative interpretation, participants would initially

access only very coarse category information about the verb, namely that it is a verb; at that first step, filler retrieval processes would be activated and an object filler would be posited for the verb; only later would finer-grained information about the verb subcategory be retrieved, transitivity information would then become available and reveal the mismatch between the filler and the verb. In our experimental design, this alternative processing strategy would have led to opposite effects: As we mentioned in our introduction, if participants initially generated expectations about lexical items on the basis of coarse category information (e.g., Verb, Noun), then the two contexts, with and without an object clitic, should have led to approximately the same number of false alarms. Indeed, both are equally good verb contexts, and should have led participants to occasionally respond too fast upon hearing a word starting with the target verb, in a verb position. The fact that participants made almost zero false alarms to the sentences with an object clitic shows that they were able to compute that this context was inappropriate for the target intransitive verb, even before they had started hearing the first phonemes of the verb itself. Taken together, the available experimental evidence thus suggests that comprehenders' ability to exploit pre-verbal arguments to constrain their interpretation of sentences, even before they have heard the verb itself, is not a specific adaptation to verb-final languages, but reflects instead a more general behavior of the human language parser.

Note that two mechanisms are compatible with the present set of results: either a predictive account, in which the preceding context is used to generate specific expectations about upcoming words, which are then matched with the input; or an integrative account, in which the preceding context is integrated very rapidly with the available phonological information. As we mentioned above, the fact that almost zero False Alarms were observed in the FA_CLI condition suggests that participants were able to compute that the target was unlikely to occur even before they heard its first syllable, which might be interpreted as evidence in favor of the predictive account over the integrative account. However, since the target verb was specified as a target before the sentence itself, it is likely that it was pre-activated before participants even started to process the sentence. As a result, the integration between context and target verb could start even before the first syllable of the carrier verb was heard, because the target verb was pre-activated. In other words, at every point in time, participants could try to work out whether their target word is likely or not in that context. When it is consistent with the context (as in FA_AMB sentences), it is likely to occur next, and hearing consistent phonological information probably results in the increased rate of False Alarms. When it is not consistent with the context (as in FA_CLI sentences), then it is unlikely to occur, and the processing of phonological information in the hope of finding the target verb may stop very early (and result in the almost-zero rate of false alarms we observed). In other words, the task we used makes it possible for participants to integrate the preceding context both with the phonological information as it becomes available, and the information about the target verb that was provided before the sentence began. All in all, both the predictive and the integrative interpretations account equally well for the present set of results[1].

An interesting particularity of the experimental paradigm we chose to use is that it allowed us to test participants' ability to use abstract syntactic information, in the absence of any semantic information conveyed by content words. Indeed, in the sentence quadruplets that were used, the first words of all four sentences, up to the critical word, were always identical, and the only difference was the presence or absence of the clitic object pronoun just before the critical verb. Because there was no difference whatsoever in the content words that were heard before the critical verb, participants' behavior was thus necessarily due to their processing of the syntactic role of the object pronoun. Thus, listeners can exploit the syntactic structure they are constructing on-line to restrict their lexical search to word candidates that fit this syntactic structure.

This conclusion might seem at odds with recent results from Chow et al. (2015), in which they conclude that comprehenders initially rely on the lexical meanings of arguments—but not their structural roles—to compute predictions about a likely upcoming verb (using sentences in which arguments were reversed, e.g., "which customer the waitress had served," vs. "which waitress the customer had served"). In the present experiment, we conclude that listeners exploit the structural role of an argument—e.g., direct object—to infer whether a target intransitive verb is plausible or not in that context. The apparent discrepancy here comes from the difference in experimental paradigms, which tested different kinds of inferences about the upcoming verb. In Chow et al. (2015) what was delayed was not really the computation of a structural role of an argument (e.g., subject, or direct object), but rather the computation of an argument's likely thematic role based on its structural role. For instance, if an argument occupies the subject position, is it the agent of the action or not? Often yes, but not necessarily (depending on the nature of the verb and the structure of the sentence). In the present experiment, simply knowing whether the verb takes a direct object or not (irrespective of the thematic role played by the referent occupying that position), was sufficient to constrain lexical access and eliminate the intransitive candidate. In Chow et al. (2015) participants had not only to find events involving waitresses (which might be fast), but also to find events in memory involving waitresses as agents (which might be slower). So overall, the available evidence suggests that structural roles are computed fast, and exploited on-line, so long as this does not involve an extra step (assigning thematic roles to arguments, and/or retrieving specific event types in memory).

If all the experiments presented above clearly point out the capacity of the linguistic parser to exploit various features of its input to anticipate different aspects of upcoming materials, it remains unclear to what extent these phenomena actually occur outside of the lab. Indeed, within experiments, participants are often placed in closed-choice situations, which restrict the number of possible anticipations that might be entertained. For the experiment reported here, participants are presented with the target verb before hearing a sentence, thus reducing considerably

[1]We thank a reviewer for helping us to clarify this important point.

the possible verbs that are expected. The same issue can be addressed to the "visual-word" paradigm, where participants are exposed to a finite number of images representing the sentence they are exposed to. As an aside, this is not such an artificial situation, as people in real-life situations will often have access to other elements of context, either visually (with a less impoverished visual context than in the visual-world paradigm), or through preceding linguistic materials (with a small set of words that can be made highly plausible by the preceding context). Even in reading experiments in which no visual or discourse context is available, participants are repeatedly exposed to sentences with similar syntactic structures, which may restrict the kind of materials they are led to expect—and participants do exploit this kind of experiment-specific information (see e.g., Gibson et al., 2013). However, all these results clearly show the capacity of the parser to rapidly integrate useful information to facilitate the process of on-line comprehension. Future research should investigate how this type of anticipation processes can be used in less constrained situations.

To conclude, the study reported here focused on the ability of French listeners to rapidly integrate syntactic cues to constrain lexical access and eliminate verb candidates on the basis of a mismatch between their sub-category and the syntactic context. Participants were shown to be able to take into account a very subtle cue, a clitic object pronoun, to infer that a target intransitive verb was unlikely to come next. This result proves that the human language parser can use subtle syntactic cues to constrain lexical access on-line, and restrict the lexical search to candidates that fit the ongoing syntactic structure. This study nicely aligns with previous data suggesting that each element from the input can be analyzed and exploited by listeners, on-line, to improve the precision of linguistic processing at all levels of linguistic analysis.

ACKNOWLEDGMENTS

This work was supported by the French Ministry of Research, the French Agence Nationale de la Recherche (grants n° ANR-2010-BLAN-1901, ANR-13-APPR-0012, ANR-10-IDEX-0001-02 PSL* and ANR-10-LABX-0087 IEC), the Fondation de France, as well as by the Région Ile-de-France, and European Research Council 269502, which supported PB while writing this manuscript. We thank I. Dautriche for suggestions on the manuscript and for help with data analysis.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpsyg.2015.01841

REFERENCES

Altmann, G. T. M., and Kamide, Y. (1999). Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition* 73, 247–264. doi: 10.1016/S0010-0277(99)00059-1

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: keep it maximal. *J. Mem. Lang.* 68, 255–278. doi: 10.1016/j.jml.2012.11.001

Bates, D., and Sarkar, D. (2007). *lme4: Linear Mixed-Effects Models Using S4 Classes*. (Version 0.9975-12) [R package]. Available online at: http://CRAN.R-project.org/

Boersma, P., and Weenink, D. (2015). *Praat: Doing Phonetics by Computer [Computer Program]*. Version 5.4.09, retrieved 1 June 2015. Available online at: http://www.praat.org/

Boland, J. E. (2005). Visual arguments. *Cognition* 95, 237–274. doi: 10.1016/j.cognition.2004.01.008

Boland, J. E., Tanenhaus, M. K., and Garnsey, S. M. (1990). Evidence for the immediate use of verb control information in sentence processing. *J. Mem. Lang.* 29, 413–432. doi: 10.1016/0749-596X(90)90064-7

Brown, M., Salverda, A. P., Dilley, L. C., and Tanenhaus, M. K. (2011). Expectations from preceding prosody influence segmentation in online sentence processing. *Psychon. Bull. Rev.* 18, 1189–1196. doi: 10.3758/s13423-011-0167-9

Chow, W.-Y., Smith, C., Lau, E., and Phillips, C. (2015). A "bag-of-arguments" mechanism for initial verb predictions. *Lang. Cogn. Neurosci.* doi: 10.1080/23273798.2015.1066832. [Epub ahead of print].

Dahan, D., Swingley, D., Tanenhaus, M. K., and Magnuson, J. S. (2000). Linguistic gender and spoken-word recognition in French. *J. Mem. Lang.* 42, 465–480. doi: 10.1006/jmla.1999.2688

de Carvalho, A., Dautriche, I., and Christophe, A. (2015). "Three-year-olds use prosody online to constrain syntactic analysis," in *38th Boston University Conference on Language Development* (Boston, MA).

DeLong, K. A., Urbach, T. P., and Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nat. Neurosci.* 8, 1117–1121. doi: 10.1038/nn1504

Farmer, T. A., Christiansen, M. H., and Monaghan, P. (2006). Phonological typicality influences on-line sentence comprehension. *Proc. Natl. Acad. Sci. U.S.A.* 103, 12203–12208. doi: 10.1073/pnas.0602173103

Federmeier, K. D. (2007). Thinking ahead: the role and roots of prediction in language comprehension. *Psychophysiology* 44, 491–505. doi: 10.1111/j.1469-8986.2007.00531.x

Foucart, A., Ruiz-Tada, E., and Costa, A. (2015). How do you know I was about to say "book"? Anticipation processes affect speech processing and lexical recognition. *Lang. Cogn. Neurosci.* 30, 768–780. doi: 10.1080/23273798.2015.1016047

Frisson, S., Rayner, K., and Pickering, M. J. (2005). Effects of contextual predictability and transitional probability on eye movements during reading. *J. Exp. Psychol. Learn. Mem. Cogn.* 31, 862–877. doi: 10.1037/0278-7393.31.5.862

Gibson, E., Bergen, L., and Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proc. Natl. Acad. Sci. U.S.A.* 110, 8051–8056. doi: 10.1073/pnas.1216438110

Hale, J. (2001). "A probabilistic earley parser as a psycholinguistic model," in *Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies*, Vol. 2 (Stroudsburg, PA: Association for Computational Linguistics), 1–8. doi: 10.3115/1073336.1073357

Hare, M., Elman, J. L., Tabaczynski, T., and McRae, K. (2009). The wind chilled the spectators, but the wine just chilled: sense, structure, and sentence comprehension. *Cogn. Sci.* 33, 610–628. doi: 10.1111/j.1551-6709.2009.01027.x

Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cogn. Sci.* 20, 137–194. doi: 10.1207/s15516709cog2002_1

Kamide, Y. (2008). Anticipatory processes in sentence processing. *Lang. Linguist. Compass* 2, 647–670. doi: 10.1111/j.1749-818X.2008.00072.x

Kamide, Y., Altmann, G. T. M., and Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *J. Mem. Lang.* 49, 133–156. doi: 10.1016/S0749-596X(03)00023-8

Kayne, R. S. (1991). Romance clitics, verb movement, and PRO. *Linguist. Inq.* 22, 647–686.

Kleiner, M., Brainard, D., and Pelli, D. (2007). "What's new in Psychtoolbox-3?," in *Perception 36 ECVP Abstract Supplement*.

Konieczny, L. (2000). Locality and parsing complexity. *J. Psycholinguist. Res.* 29, 627–645. doi: 10.1023/A:1026528912821

Kutas, M., and Hillyard, S. A. (1980). Reading senseless sentences: brain potentials reflect semantic incongruity. *Science* 207, 203–205. doi: 10.1126/science.7350657

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition* 106, 1126–1177. doi: 10.1016/j.cognition.2007.05.006

Levy, R. P., and Keller, F. (2013). Expectation and locality effects in German verb-final structures. *J. Mem. Lang.* 68, 199–222. doi: 10.1016/j.jml.2012.02.005

Magnuson, J. S., Tanenhaus, M. K., and Aslin, R. N. (2008). Immediate effects of form-class constraints on spoken word recognition. *Cognition* 108, 866–873. doi: 10.1016/j.cognition.2008.06.005

Millotte, S., René, A., Wales, R., and Christophe, A. (2008). Phonological phrase boundaries constrain on-line syntactic analysis. *J. Exp. Psychol. Learn. Mem. Cogn.* 34, 874–885. doi: 10.1037/0278-7393.34.4.874

New, B., Pallier, C., Ferrand, L., and Matos, R. (2001). Une base de données lexicales du français contemporain sur internet: LEXIQUE. *L'Année Psychol.* 101, 447–462. doi: 10.3406/psy.2001.1341

Omaki, A., Lau, E., White, D. I., Dakan, M., Apple, A., and Colin, P. (2015). Hyper-active gap filling. *Front. Psychol.* 6:384. doi: 10.3389/fpsyg.2015.00384

Poeppel, D., Idsardi, W., and van Wassenhove, V. (2008). Speech perception at the interface of neurobiology and linguistics. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 12, 1071–1086. doi: 10.1098/rstb.2007.2160

Pylkkänen, L., and Marantz, A. (2003). Tracking the time course of word recognition with MEG. *Trends Cogn. Sci.* 7, 187–189. doi: 10.1016/S1364-6613(03)00092-5

Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *Q. J. Exp. Psychol.* 62, 1457–1506. doi: 10.1080/17470210902816461

Shillcock, R. C., and Bard, E. (1993). "Modularity and the processing of closed-class words," in *Cognitive Models of Speech Processing: The Second Sperlonga Meeting,* eds G. T. M. Altmann and R. C. Shillcock (Hove: Lawrence Erlbaum Associates), 163–185.

Tanenhaus, M. K., Leiman, J. M., and Seidenberg, M. S. (1979). Evidence for multiple stages in the processing of ambiguous words in syntactic contexts. *J. Verbal Learn. Verbal Behav.* 18, 427–440. doi: 10.1016/S0022-5371(79)90237-8

Van Berkum, J. J. A., Brown, C. M., Zwitserlood, P., Kooijman, V., and Hagoort, P. (2005). Anticipating upcoming words in discourse: evidence from ERPs and reading times. *J. Exp. Psychol. Learn. Mem. Cogn.* 31, 443–467. doi: 10.1037/0278-7393.31.3.443

Wicha, N. Y. Y., Moreno, E. M., and Kutas, M. (2004). Anticipating words and their gender: an event-related brain potential study of semantic integration, gender expectancy, and gender agreement in Spanish sentence reading. *J. Cogn. Neurosci.* 16, 1272–1288. doi: 10.1162/0898929041920487

Xiang, M., and Kuperberg, G. (2015). Reversing expectations during discourse comprehension. *Lang. Cogn. Neurosci.* 30, 648–672. doi: 10.1080/23273798.2014.995679

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2015 Brusini, Brun, Brunet and Christophe. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 30 March 2016
doi: 10.3389/fpsyg.2016.00403

Dependency Resolution Difficulty Increases with Distance in Persian Separable Complex Predicates: Evidence for Expectation and Memory-Based Accounts

Molood S. Safavi[1], *Samar Husain*[2] *and Shravan Vasishth*[3]*

[1] International Doctorate for Experimental Approaches to Language and Brain (IDEALAB), University of Potsdam, Germany / University of Groningen, Netherlands /University of Trento, Italy / University of Newcastle, UK / Macquarie University Sydney, Australia, [2] Department of Humanities and Social Sciences, Indian Institute of Technology, New Delhi, India, [3] Department of Linguistics, University of Potsdam, Potsdam, Germany

Delaying the appearance of a verb in a noun-verb dependency tends to increase processing difficulty at the verb; one explanation for this locality effect is decay and/or interference of the noun in working memory. Surprisal, an expectation-based account, predicts that delaying the appearance of a verb either renders it no more predictable or more predictable, leading respectively to a prediction of no effect of distance or a facilitation. Recently, Husain et al. (2014) suggested that when the exact identity of the upcoming verb is predictable (strong predictability), increasing argument-verb distance leads to facilitation effects, which is consistent with surprisal; but when the exact identity of the upcoming verb is not predictable (weak predictability), locality effects are seen. We investigated Husain et al.'s proposal using Persian complex predicates (CPs), which consist of a non-verbal element—a noun in the current study—and a verb. In CPs, once the noun has been read, the exact identity of the verb is highly predictable (strong predictability); this was confirmed using a sentence completion study. In two self-paced reading (SPR) and two eye-tracking (ET) experiments, we delayed the appearance of the verb by interposing a relative clause (Experiments 1 and 3) or a long PP (Experiments 2 and 4). We also included a simple Noun-Verb predicate configuration with the same distance manipulation; here, the exact identity of the verb was not predictable (weak predictability). Thus, the design crossed Predictability Strength and Distance. We found that, consistent with surprisal, the verb in the strong predictability conditions was read faster than in the weak predictability conditions. Furthermore, greater verb-argument distance led to slower reading times; strong predictability did not neutralize or attenuate the locality effects. As regards the effect of distance on dependency resolution difficulty, these four experiments present evidence in favor of working memory accounts of argument-verb dependency resolution, and against the surprisal-based expectation account of Levy (2008). However, another expectation-based measure, entropy, which was computed using the offline sentence completion data, predicts reading times in Experiment 1 but not in the other experiments. Because participants tend to produce more ungrammatical continuations in the long-distance condition in Experiment 1, we suggest that forgetting due to memory overload leads to greater entropy at the verb.

Keywords: locality, expectation, surprisal, entropy, Persian, complex predicates, self-paced reading, eye-tracking

OPEN ACCESS

Edited by:
Matthew Wagers,
University of California, Santa Cruz, USA

Reviewed by:
Tim Hunter,
University of Minnesota, USA
Pouneh Shabani-Jadidi,
McGill University, Canada

***Correspondence:**
Shravan Vasishth
vasishth@uni-potsdam.de

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: *30 August 2015*
Accepted: *07 March 2016*
Published: *30 March 2016*

Citation:
Safavi MS, Husain S and Vasishth S (2016) Dependency Resolution Difficulty Increases with Distance in Persian Separable Complex Predicates: Evidence for Expectation and Memory-Based Accounts. Front. Psychol. 7:403. doi: 10.3389/fpsyg.2016.00403

1. INTRODUCTION

A long-standing claim in sentence processing is that increasing distance in a linguistic dependency, such as a noun-verb dependency, leads to greater processing difficulty (Chomsky, 1965; Just and Carpenter, 1992; Gibson, 2000; Lewis and Vasishth, 2005); it is common to refer to this increase in processing difficulty as the locality effect. One explanation for the locality effect is in terms of constraints imposed by working memory. According to one account, the Dependency Locality Theory (DLT; Gibson, 1998), the processing difficulty experienced when resolving a long dependency depends on the decay experienced by the noun; a related account by Lewis and Vasishth (2005) attributes the locality effect to decay and/or interference. Constraints on working memory may be a plausible explanation given that individuals' working memory capacity seems to affect the processes involved in dependency resolution (Caplan and Waters, 2013; Nicenboim et al., 2015). Although there is evidence consistent with the memory-based explanation in English, German, Chinese, Russian, and Hindi, (Hsiao and Gibson, 2003; Grodner and Gibson, 2005; Bartek et al., 2011; Vasishth and Drenhaus, 2011; Levy et al., 2013; Husain et al., 2014, 2015), research on some of these languages has also uncovered evidence that increasing noun-verb distance facilitates processing at the verb (Konieczny, 2000; Vasishth, 2003; Vasishth and Lewis, 2006; Jaeger et al., 2008; Vasishth and Drenhaus, 2011; Levy and Keller, 2013; Husain et al., 2014; Jäger et al., 2015). One explanation for these anti-locality effects is in terms of surprisal (Hale, 2001; Levy, 2008). Surprisal extends and formalizes the old idea of predictive sentence processing—which has been extensively investigated in the EEG literature (e.g., Kutas and Hillyard, 1984)—in terms of probabilistic parse continuations (also see Jurafsky, 1996). The surprisal account assumes that the comprehender maintains and uses linguistic knowledge probabilistically to parse a sentence incrementally. Surprisal is the claim that rare transitions are difficult: increased processing difficulty is predicted when a parser is required to build a low-probability syntactic structure. Formally, surprisal is defined as the negative log probability of encountering a particular part of speech or word given previous context. We will refer to surprisal as the expectation-based account, following the terminology of Levy (2008)[1].

In many of these studies, evidence has been found for both the memory-based accounts and the expectation-based account. One conclusion that has emerged is that both memory and expectation play a role. For example, in his eye-tracking (ET) study investigating processing difference in English object vs. subject relative clauses, Staub (2010) finds evidence for both expectation-based processing and locality constraints, although these occur in different regions of the target sentence. An example of Staub's design is provided below. In this study, processing difficulty was found on the noun phrase *the fireman* in the ORC (object relative clause) 1b, compared to the SRC (subject relative clause) 1a; this is consistent with the expectation account because the reader would be forced to build a rare object relative in the ORC condition when he/she encounters the noun phrase. However, this study also found greater processing difficulty at the relative clause verb in ORCs than SRCs, which is predicted by memory accounts.

(1) a. The employees that noticed the fireman hurried across the open field.

b. The employees that the fireman noticed hurried across the open field.

As further examples, both Vasishth and Drenhaus (2011) and Levy and Keller (2013) have argued that locality effects may appear when high working memory load is experienced; anti-locality effects may be present when the load is low.

In a recent development, Husain et al. (2014) argue that the strong predictability for a head (predicting an exact lexical item) can neutralize the locality effect; locality may manifest itself only when predictability strength is weak, that is, when only a verb phrase is predicted, and not the exact identity of the verb. In their self-paced reading (SPR) study, Husain et al. (2014) used a 2×2 design, crossing Predictability and Dependency Distance to investigate locality and anti-locality effects. In the strong predictability conditions, Hindi complex predicates (CPs) were used. In these noun-verb sequences, the noun strongly predicted the upcoming light verb, e.g., the noun *khayaal*, "care," strongly predicts the verb *rakhnaa*, "put," in *khayaal rakhnaa*, literally, "care put" ("to take care of"). The weak predictability condition, on the other hand, used the same verb used in the complex predicate, but the noun did not predict the verb. An example is *gitaar rakhnaa*, "guitar put"; "to put (down) a guitar"; here, the verb retains its literal meaning. Thus, when the reader sees *gitaar*, they cannot predict the exact identity of the verb, because many other verbs are possible here (e.g., bought). To summarize, in the strong predictability condition, the noun predicted the exact identity of the verb, while in the weak predictability condition the exact identity of the verb was not predicted with high certainty—although *a* verb was predicted. The second factor, dependency distance, was manipulated by placing one to two adverbials between the nominal predicate/object and the verb in the short condition. The long condition had two to three intervening adverbials. Reading time was measured at the verb. The results showed that CP light verbs were read faster in long vs. short distance conditions, but for the non-CP verb there was a tendency toward a slowdown in long vs. short conditions. Finally, there was weak evidence for an interaction (estimate on the log ms scale: 0.03, Bayesian 95% credible interval $[-0.02, 0.07]$, posterior probability of the effect being greater than 0 was 0.77). That is, there was some indication that with increased distance there was a speedup at the light verb in the CP conditions and a slowdown in the non-CP conditions. Although these results can also be interpreted as showing no interaction, Husain and colleagues suggested that strong predictability of the head could be canceling the locality effect, with the locality effect manifesting itself only when predictability strength was weak.

[1]Another expectation-based account in the literature is the entropy reduction hypothesis or ERH (Hale, 2006); we do not investigate ERH in this paper, but we do discuss a related idea, entropy, in Section 8.

In the present study, we build on the work by Husain et al. (2014) described above. Husain and colleagues' work suggested that the strength of the predictability may modulate whether locality effects occur or not; we investigate the cross-linguistic generality of this claim using Persian, which, like Hindi, also has a complex predicate construction that allows us to manipulate strong and weak predictability. We turn next to a short discussion of the complex predicate construction in Persian as it relates to our experiments.

2. COMPLEX PREDICATES IN PERSIAN

CPs consist of a sequence containing a non-verbal element (e.g., a noun) and a verb, where the meaning of the sequence is non-compositional (Samvelian, 2001). An example is shown in (2).

(2) Maryam be man latme zad
Maryam to me damage hit
'Maryam caused damage to me (Maryam harmed me).'

The verb, often called a light verb, lacks sufficient semantic force to function as an independent predicate (Vahedi-Langrudi, 1996; Karimi-Doostan, 1997, 2005) and can be combined with different types of non-verbal items such as nominal, adjectivals or prepositional phrases (Dabir-Moghaddam, 1997).

In our study, we used separable CPs as defined by Karimi-Doostan (2011). According to Karimi-Doostan, a complex predicate can be separated if it satisfies both of the following two conditions: (1) if the nominal part is a noun to which adjectives, demonstratives, and wh-words, etc. can be attributed, and (2) if this noun has an internal argument structure (referring to an action or event). From this perspective, Persian CPs are categorized in three groups: (1) predicative verbal nouns (e.g., *anja:m da:dan*, perform+to give), (2) predicative nouns (e.g., *latme zadan*, damage+to hit), and (3) non-predicative nouns (e.g., *gush da:dan*, ear+to do). Among these three types, only the second one satisfies both of the conditions.

We began by independently validating our assumption that the CPs we used in our experiments are predictable and separable. We first conducted a norming study (a sentence completion task), to establish that the light verbs (of the separated CPs) are highly predictable when the nominal is provided, as compared to non-CP verbs in simple predicate conditions. We then conducted an acceptability rating study to determine how acceptable Persian CPs are when they get separated.

3. NORMING STUDIES

In order to prepare appropriate stimuli, two norming studies were run. The first study involved offline sentence completion and served to validate (i) whether the identity of the verb in the complex predicate is highly predictable, and (ii) whether the identity of the verb in the control conditions is not predictable.

The second study involved offline acceptability rating; the goal was to choose CPs for our experiments which are separable. That is, we wanted to identify CPs which native speakers would find acceptable even if an intervener occurs between the noun-verb sequence. Instructions for both studies are provided in the supplementary Datasheet 1.

The sentence completion study was carried out to derive the predictions of the expectation account. Previous work on expectation effects suggests that sentence completion data may be useful for this purpose. For example, Levy and Keller (2013) used sentence completion data to complement their corpus analyses for deriving their predictions. In their study, the key issue was whether the intervening material (e.g., a dative marked NP) leads to a prediction of a dative verb. Their Table 4 shows that the intervening material sharpened the expectation for the type of verb predicted. This shows that sentence completion data can be used to determine empirically whether the prediction for a specific verb or a verb type is sharpened by intervening material; in the Levy and Keller case, it makes sense that the intervener sharpens the expectation, but clearly the nature and content of the intervening phrases will be crucial in determining whether expectations are sharpened (Konieczny, 2000; Grodner and Gibson, 2005)[2]. Similarly, Husain et al. (2014) used sentence completion to establish that the identity of the verb in a complex predicate is highly predictable given the preceding context, but the identity of the verb in a simple predicate is not (see their Table 4). A third example is Jäger et al. (2015); they used both corpus data and sentence completion to establish that a sentence starting with a determiner, classifier, and an adverb leads to the prediction of a relative clause continuation in Chinese, and that the conditional probability of a subject relative continuation is higher than that of an object relative continuation (see their Table 2). Given these previous results, we assume that sentence completion data is informative about the predictions of the expectation-based account.

3.1. Sentence Completion Studies

Two groups (32 participants each) of Persian native speakers, who did not take part in any of the other experiments, participated in two sentence completion pre-tests in which they were asked to complete the sentences after they were presented the sentence fragment until the pre-critical word. For example, as shown in example 3, subjects were shown incomplete sentences which they had to complete; in this example, the missing verb is shown in parentheses. The participants were allowed to complete the sentence with as many words as they wanted, but our interest was only in the first word that they would write, which would most likely be a verb. This allowed us to calculate the proportion of continuations in which the exact verb was produced.

(3) a. Ali a:rezouyee bara:ye man (kard) …
Ali wish-INDEF for 1.S (do-PST …
'Ali (made) a wish for me …'

[2]We return to this point in the Section 8, where we discuss the effect of entropy on reading times.

b. Ali a:rezouyee ke besya:r doost-da:sht-am
Ali wish-INDEF that a lot like-1.S-PST
bara:ye man (kard) …
for 1.S (do-PST) …
'Ali (made) a wish that I liked a lot for me …'

The materials were exactly the same as the ones used in the experiments presented below. For the Experiment 1 items, the average prediction accuracy for the exact verb in the strong predictability conditions was 64.46% for the short condition and 59.44% for the long condition; for the Experiment 2 items, it was 65.28 and 62.85% for the short and long conditions respectively. By contrast, the average prediction accuracy for the exact verb in the weak predictability conditions in Experiment 1 was 35.42 and 34.03% for the short and long conditions; and in Experiment 2, it was 36.36 and 30.21% for the short and long conditions. As shown in **Tables 1**, **2**, an analysis using Bayesian generalized linear mixed models with a binomial link function shows a main effect of predictability in both the first experiment and the second experiment.

In the Bayesian models, we used weakly informative priors for the fixed effects (a Student t-distribution with 2 degrees of freedom), and for the random effects (a so-called LKJ prior on the correlation matrix of the random effects' variance-covariance matrix). For an introduction specifically for psycholinguistics, see Sorensen and Vasishth (2015); Nicenboim and Vasishth (2016). One way to interpret whether there is an effect of a particular factor in Bayesian (G)LMMs is to check that the 95% uncertainty interval does not contain zero.

As is clear from the mean percentages for each condition, the light verbs used in the complex predicate conditions were relatively predictable, and the heavy verbs used in the simple predicate conditions were relatively unpredictable. It is also clear from this study that, in our materials, increasing the amount of intervening material does not render the upcoming verb more predictable. The additional information provided by the intervening material for predicting the upcoming verb has been suggested by Konieczny (2000) as one possible explanation for shorter reading times at the verb in long- vs. short-distance conditions. Although this proposal is likely to be correct for some constructions (see discussion in Grodner and Gibson, 2005), in our materials, the sentence completion data do not provide any evidence that the intervening words we used in our design sharpen the expectation for the verb[3].

TABLE 1 | Model results from the Bayesian linear mixed model for the sentence completion study (Experiment 1).

Comparison	Mean	Lower	Upper	$P(b < 0)$
Intercept	−0.2055	−0.8375	0.407	0.744
Distance	−0.1584	−0.4709	0.143	0.8548
Predictability	0.9635	0.3184	1.6186	0.0025
Distance × Predictability	−0.1246	−0.4688	0.216	0.766

Shown are the mean and 95% uncertainty intervals, and the probability of the parameter being less than 0.

TABLE 2 | Model results from the Bayesian linear mixed model for the sentence completion study (Experiment 2).

Comparison	Mean	Lower	Upper	$P(b < 0)$
Intercept	−0.142	−0.7587	0.4698	0.677
Distance	−0.16	−0.4042	0.0843	0.9035
Predictability	1.1188	0.4919	1.7495	2e-04
Distance × Predictability	0.0365	−0.2102	0.2727	0.3682

Shown are the mean and 95% uncertainty intervals, and the probability of the parameter being less than 0.

3.2. Acceptability Rating of Separable vs. Inseparable CPs

Because the noun-verb sequences must be separable for our design to work, we also carried out an acceptability rating pre-test to make sure that the separability of the CPs used in our study is acceptable to native speakers. We tested for the acceptability of different types of noun-verb dependencies by interposing a short prepositional phrase between them. Taking Karimi-Doostan's classification of CPs into account, 36 items from each of the three categories were selected and randomized to test 50 native speakers of Persian (these participants did not take part in any other experiments reported here). They were asked to rate the sentences from 1 (unacceptable) to 7 (completely acceptable). Every participant saw all items. The average acceptability ratings for predicative verbal nouns, predicative nouns and non-predicative nouns were 3.23 (first quartile 1, third quartile 5), 6.08 (first quartile 6, third quartile 7), and 3.12 (first quartile 1, third quartile 5) respectively. That is, items with predicative nouns were the most acceptable. We used all the 36 items of the predicative noun condition in our Experiments 1, 2, and 32 items in Experiments 3, 4 (see the Section 6.1 of Experiment 3 for an explanation).

4. EXPERIMENT 1

4.1. Method

4.1.1. Participants

Forty-two participants aged between 17 and 40 years old (mean 24 years) participated in this experiment in Tehran, Iran. All participants were native speakers of Persian and were unaware of the purpose of the study. This study was carried out in accordance with the Helsinki Declaration, and letters of consent were obtained from all the participants.

4.1.2. Materials

We created 36 experimental sentences with a 2 × 2 factorial design, manipulating predictability strength and distance between the object noun and verb. The short intervener was a prepositional phrase and the long intervener was a relative clause added before the prepositional phrase. In order to mask

[3]In fact, in our sentence completion data, as discussed in the Section 8, entropy increases with distance.

the experiment, we included 100 filler sentences with varying syntactic structures (see supplementary materials). Here is an example of the target sentences:

(4) a. Strong predictability, short distance (PP)
Ali a:rezouyee bara:ye man kard va…
Ali wish-INDEF for 1.S do-PST and…
'Ali made a wish for me and…'

b. Strong predictability, long distance (RC+PP)
Ali a:rezouyee ke besya:r doost-da:sht-am
Ali wish-INDEF that a lot like-1.S-PST
bara:ye man kard va…
for 1.S do-PST and…
'Ali made a wish that I liked a lot for me and…'

c. Weak predictability, short distance (PP)
Ali shokola:ti bara:ye man xarid va…
Ali chocolate-INDEF for 1.S buy-PST and…
'Ali bought a chocolate for me and …'

d. Weak predictability, long distance (RC+PP)
Ali shokola:ti ke besya:r doost-da:sht-am
Ali chocolate-INDEF that a lot like-1.S-PST
bara:ye man xarid va…
for 1.S buy-PST and…
'Ali bought a chocolate that I liked a lot for me and….'

The critical region is the verb (*kard* and *xarid*).

Each sentence (including fillers) was followed by a yes/no comprehension question which targeted different thematic roles in the sentence. Half the questions had a yes answer and half had a no answer. The questions used for the target sentences are provided in the supplementary material.

4.1.3. Procedure

Participants were tested individually using a PC. They were explained the task before they performed the SPR experiment. The participants were instructed to read for comprehension in a normal manner and had a practice session of five sentences. All the sentences were displayed on a single line and were presented in 22 pt Persian Arial font using Linger software (http://tedlab.mit.edu/~dr/Linger/). In order to read each word of a sentence successively in a moving window display, participants had to press the space bar; then the word seen previously was masked and the next word was shown. After each sentence, they were asked to answer a comprehension question to ensure that the participants paid attention to the complete sentence.

4.1.4. Data Analysis

The data analysis was conducted in the R programming environment (R Development Core Team, 2013), using Bayesian hierarchical (so-called linear mixed) models using Stan (Stan Development Team, 2014; Gabry and Goodrich, 2016). Sum contrasts were used to code main effects and interactions. In addition, a nested contrast was defined for a secondary analysis in order to look at the effect of distance in CPs vs. the control conditions separately; these were also coded as sum contrasts. We fit full variance-covariance matrices for participants and items (the so-called maximal model, Barr et al., 2013; Bates et al., 2015). All data and code are available from http://www.ling.uni-potsdam.de/~vasishth/code/SafaviEtAl2016DataCode.zip.

4.2. Predictions

Based on the Husain et al. (2014) results, in Experiment 1, we expected that increasing noun-verb distance would lead to faster reading time at the verb in the strong predictable conditions, but slower reading time in the weak predictable conditions. Thus, we expected to obtain a cross-over interaction.

The memory based accounts (Just and Carpenter, 1992; Gibson, 2000; Lewis and Vasishth, 2005) predict that increasing distance should lead to a slowdown at the verb; these accounts make no predictions about the strength of predictability.

There are two alternative predictions possible for the expectation account, depending on how one operationalizes expectation. First, if sentence completion probabilities are a reasonable proxy for conditional probabilities—and the previous research reported above (Levy and Keller, 2013; Husain et al., 2014; Jäger et al., 2015) suggests that they may be—then we predict (a) no difference in reading time at the verb as a function of distance, and (b) faster reading time at the verb in the strong predictable conditions than the weak predictable conditions. Prediction (a) arises because, in the sentence completion data, we see no effect of distance on the predictability of the upcoming verb, in either the strong or weak predictability conditions; prediction (b) arises due to the difference in predictability of the exact verb that we see in the strong vs. weak predictability conditions (see the results of the sentence completion studies).

An alternative possible prediction of the expectation account is that increasing distance should facilitate processing at the verb. Surprisal predicts facilitation with increasing distance whenever distance causes the number of possible parses to decrease; this decrease in the number of possible parses leads to the probability mass being reassigned among the remaining parses. In our materials, when a participant reads the noun in the noun-verb complex predicate, they are expecting the light verb with high probability (nearly 1). However, in the long distance condition, the next word begins a relative clause; this leads to an expectation that the light verb will appear *after* the relative clause verb. But what appears after the relative clause verb is a PP that modifies the upcoming light verb. For a facilitation to be predicted in this long-distance condition by the surprisal metric, it would have to be the case that the conditional probability of the light verb following the RC and PP would be higher than the conditional probability of the light verb in the short-distance (PP) condition. In order to get a sense of how the conditional probabilities change in the noun-light verb conditions as a function of distance, we extracted all light verb sentences from a Persian corpus (Seraji, 2015) and then counted, for different numbers of

TABLE 3 | The conditional probability of a light verb appearing given the complex predicate noun and *n* intervening phrases between the noun and the light verb.

n Intervening phrases	Probability of verb
0	3826/4003 = 0.95
1	131/133 = 0.98
2	28/31 = 0.90
3	5/5 = 1
4	2/2 = 1
6	1/1 = 1

TABLE 4 | The conditional probability of a light verb appearing given the complex predicate noun and *n* intervening words between the noun and the light verb.

n Intervening words	Probability of verb
0	3826/4003 = 0.96
1	104/104 = 1
2	36/39 = 0.92
3	4/5 = 0.8
4	9/10 = 0.9
5	3/3 = 1
6	3/3 = 1
7	1/1 = 1
8	2/2 = 1
10	2/2 = 1
12	1/1 = 1
13	1/1 = 1
14	1/1 = 1

modifying phrases, the proportion of cases that a verb followed the intervening phrase. For example, in a Persian sentence such as *John in the morning went*, there is one intervening phrase, the PP. As shown in **Table 3**, we find that the conditional probability of the verb appearing next is always high, but goes to 1 with increasing distance. This suggests that in general, increasing distance tends to sharpen the expectation for an upcoming verb. We also did this calculation using the number of intervening words as a metric, rather than the number of intervening phrases. The result, shown in **Table 4**, is substantially the same as in **Table 3**. Of course, these corpus counts don't give us any direct information about the predictions regarding our particular experiment design.

Regarding the strong vs. weak predictability conditions, note that the expectation account of Hale and Levy does not predict that processing should be facilitated when the exact identity of the upcoming verb is predicted (strong predictability condition), compared to the case when just some verb is predicted (weak predictability condition). This is because the surprisal metric is usually calculated using the conditional probability of the part-of-speech (verb) given preceding context, and this will be the same in both the strong and weak predictability conditions. However, it is possible to subsume the difference between strong and weak predictability under the surprisal account by reframing the conditional probabilities in terms of the exact identity of the verb. In this case, the expectation account would predict faster reading times in the strong predictability conditions compared to the weak predictability conditions, regardless of distance.

To summarize, regarding the distance manipulation, the expectation account predicts either no effect or a facilitation at the verb as a function of distance; and regarding the predictability manipulation, the expectation account (appropriately formulated to include the conditional probability of the exact lexical item predicted) would predict a main effect of predictability.

TABLE 5 | Means, 95% uncertainty intervals, and $P(b < 0)$, the probability of the estimate being less than 0, in the question-response accuracy analysis for Experiment 1.

Comparison	Mean	Lower	Upper	$P(b < 0)$
Intercept	3.3271	2.8738	3.8724	0
Distance	0.033	−0.3242	0.4033	0.4212
Predictability	−0.174	−0.6128	0.2239	0.8002
Distance × Predictability	0.158	−0.1147	0.4385	0.1345

TABLE 6 | Means, 95% uncertainty intervals, and $P(b < 0)$, the probability of the estimate being less than 0, in the reading time analysis for Experiment 1.

Comparison	Mean	Lower	Upper	$P(b < 0)$
Intercept	6.2434	6.1644	6.3226	0
Distance	0.0397	0.0174	0.0619	2e-04
Predictability	−0.0328	−0.0566	−0.0096	0.998
Distance × Predictability	−0.0179	−0.0405	0.0046	0.942

4.3. Results

4.3.1. Comprehension Accuracy

Participants answered correctly on average 93% of all comprehension questions (excluding fillers). Accuracy was 91, 94, 95, and 91% respectively for the four conditions in (1). As shown in **Table 5**, a Bayesian generalized linear mixed model of the binary responses showed no evidence for an effect of distance or predictability, or an interaction between predictability and distance.

4.3.2. Reading Time

Reading times (RTs) were analyzed at the verb. As shown in **Table 6** and **Figure 1**, there was a main effect of distance, such that increasing distance led to longer reading times. There was also a main effect of predictability: the complex predicate conditions were read faster overall. A weak interaction was also seen: stronger locality effects were seen in the control conditions than in the complex predicate conditions. A nested analysis shows that the distance effect was driven by the control (weak predictability) condition. The estimates for the strong predictability condition were coef. = 0.0218, [−0.0094, 0.0524], $P(b < 0) = 0.0875$); and the estimates for weak predictability were coef. = 0.0581, 95% uncertainty intervals [0.0261, 0.0912], $P(b < 0) = 2e\text{-}04$.

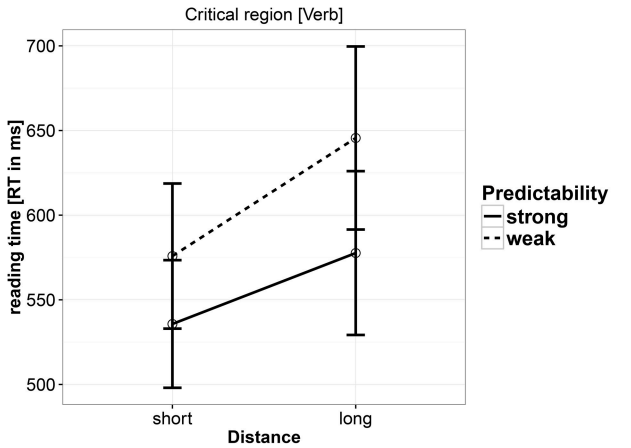


FIGURE 1 | Reading times at the critical verb in Experiment 1.

4.4. Discussion

Experiment 1 found a main effect of predictability such that the strong predictability conditions were read faster than the weak predictability conditions, and a main effect of distance, such that the short conditions were read faster than the long conditions. A nested contrast showed that this effect of distance was driven by the weak predictability conditions, i.e., reading time at the verb in condition c was faster than the reading time in condition d. A weak interaction suggests that the locality effect may be somewhat stronger in the weak predictability condition. The suggestion of an interaction seems to provide only weak support, if any, for the idea that strong predictability can at least attenuate locality effects (Husain et al., 2014). The overall effect of distance is consistent with memory-based accounts, which correctly predict a slowdown at the verb in the long conditions, i.e., a main effect of distance. However, as the nested comparison shows, the main effect of distance is driven only by the weak predictability (non-complex predicate) conditions. Memory-based theories would be unable to explain this because they predict a slowdown in long conditions irrespective of predictability strength. However, note that the absence of an interaction makes this absence of a distance effect in the strong predictability conditions difficult to interpret. The expectation account's prediction regarding distance, that increasing the argument-verb distance would either have no effect or result in a facilitation, was clearly not validated; however, the main effect of predictability is consistent with a version of the expectation account that uses the conditional probability of the exact lexical item (verb) appearing given the preceding context.

Our original motivation for this study was to attempt a replication of the Husain et al. (2014) findings. The results are not entirely inconsistent with those of Husain et al. (2014), but they are also not a strong validation of the expectation-memory cost tradeoff posited in that paper. As in the Husain et al. study, we see a main effect of predictability driven by the complex predicate condition. This effect could be explained in terms of reduced retrieval cost at the verb due to its high expectation. An obvious confounding factor here is that the verbs in the strong vs. weak predictability conditions are not identical; this prevents us from ruling out the possibility that low-level differences in the verbs might be responsible for the facilitation due to prediction strength.

We turn next to Experiment 2, in which we manipulate the type of intervener. Here, in the long distance condition, instead of a relative clause and prepositional phrase (PP) intervener, a long PP intervenes. The motivation was to increase distance without having different types of interveners in the short vs. long conditions, as this might be a fairer comparison.

5. EXPERIMENT 2

5.1. Method

5.1.1. Participants

Forty-three participants, with the same criteria as in Experiment 1, participated in this experiment in Tehran, Iran. This study was carried out in accordance with Helsinki Declaration, and consent forms were obtained from all the participants.

5.1.2. Materials

The stimuli and fillers were the same as in Experiment 1 except for the long conditions (b and d), where the intervener was a longer prepositional phrase (PP) instead of the combination of a relative clause and a PP as in the previous experiment. The PP was lengthened using several different structures, all of which had one or more instance of the ezafe possessive marker (Samvelian, 2007):

1. N-ez N-ez N/pronoun/proper name
2. N-ez adj-ez N/pronoun/proper name
3. N-ez adj-ez N
4. N-ez N-ez adj
5. N adj-ez adj
6. superlative adj N N/pronoun/proper name
7. N-ez pronoun

One set of examples using the first type of PP shown above is as follows:

(5) a. Strong predictability, short distance (PP)

Ali a:rezouyee bara:ye man kard va…
Ali wish-INDEF for 1.S do-PST and…

'Ali made a wish for me and…'

b. Strong predictability, long distance (longer PP)

Ali a:rezouyee bara:ye doost-e xa:har-e man kard va…
Ali wish-INDEF for friend-EZ sister-EZ 1.S do-PST and…

'Ali made a wish for my sister's friend…'

c. Weak predictability, short distance (PP)

Ali shokola:ti bara:ye man xarid va…
Ali chocolate-INDEF for 1.S buy-PST and…
'Ali bought a chocolate for me and…'

d. weak predictability, long distance (longer PP)
Ali shokola:ti bara:ye doost-e xa:har-e
Ali chocolate-INDEF for friend-EZ sister-EZ
man xarid va…
1.S buy-PST and…
'Ali bought a chocolate for my sister's friend and…'

More details about the PPs are provided in the supplementary materials.

5.1.3. Procedure and Data Analysis

The procedure and data analysis methodology was the same as Experiment 1.

5.2. Predictions

In Experiment 2, the distance manipulation involves lengthening the PP. There are two possible predictions of surprisal. One is that surprisal may predict no difference at the verb; this would be because the end of the PP raises a strong expectation for a verb, and this strong expectation for a verb would be the same in both the short and long PP conditions. Another alternative possible prediction of surprisal is that lengthening the PP could lead to a facilitation. This prediction could hold if increasing distance, counted in terms of the number of intervening words, generally increases the predictability of the upcoming verb; this is a possibility given the corpus counts in **Table 4**.

5.3. Results

5.3.1. Comprehension Accuracy

Participants answered 93% of all comprehension questions correctly on average (excluding fillers). The accuracies by condition were 96, 92, 94, and 89% respectively for the four conditions in (2). As shown in **Table 7**, the Bayesian generalized linear mixed models of the responses showed a main effect of distance, such that accuracies were lower in the long conditions. No effect of predictability strength, and no interaction between predictability strength and distance were found.

5.3.2. Reading Time

As shown in **Table 8** and **Figure 2**, the results showed a main effect of distance, with long distance conditions being read slower. There was only a weak effect of predictability, with the strong predictability condition being read faster than the weak predictability condition. No interaction was found between predictability and distance. A nested contrast showed that the distance effect is seen in both strong predictability (coef. = 0.0623, [0.0274, 0.0965], $P(b < 0) = 0$) and weak predictability (coef. = 0.0475, [0.0098, 0.085], $P(b < 0)$=0.0078) conditions.

5.4. Discussion

In this experiment, we replicated the locality effects found in Experiment 1, but we no longer see a weakening of the locality effect that was seen in Experiment 1 (a marginal interaction was found in Experiment 1). Nested contrasts showed that locality effects are equally strong in both the strong and weak predictability conditions. In Experiment 2, we also see an effect of predictability, with the strong predictable verb being read faster. Thus, regarding the distance manipulation, the working-memory account's prediction is validated, and the expectation-based account's prediction is not supported. The main effect of predictability does furnish evidence consistent with the expectation-based account.

A secondary analysis was conducted to compare the strength of the locality effect in the two experiments, and to determine whether an interaction between distance, predictability and experiment was present. The between-participant factor experiment was coded using sum coding: Experiment 1 was coded −1, and Experiment 2 was coded +1 (further details are available in the supplementary materials). The results are shown

TABLE 7 | Means, 95% uncertainty intervals, and $P(b < 0)$, the probability of the estimate being less than 0, in the question-response accuracy analysis for Experiment 2.

Comparison	Mean	Lower	Upper	$P(b < 0)$
Intercept	3.1092	2.7277	3.5353	0
Distance	−0.4246	−0.7556	−0.133	0.9972
Predictability	0.1798	−0.1871	0.5605	0.157
Distance × Predictability	−0.0742	−0.3478	0.1832	0.7098

TABLE 8 | Means, 95% uncertainty intervals, and $P(b < 0)$, the probability of the estimate being less than 0, in the reading time analysis for Experiment 2.

Comparison	Mean	Lower	Upper	$P(b < 0)$
Intercept	6.2676	6.1867	6.3488	0
Distance	0.0547	0.0269	0.0827	0
Predictability	−0.0203	−0.0417	0.0013	0.966
Distance × Predictability	0.0077	−0.016	0.0318	0.2585

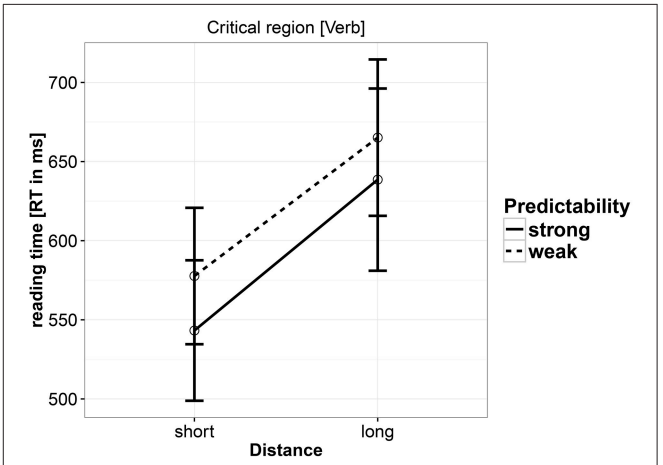


FIGURE 2 | Reading times at the critical verb in Experiment 2.

TABLE 9 | Comparison of Experiments 1 and 2.

	Mean	Lower	Upper	$P(b < 0)$
Intercept	6.2578	6.1974	6.3198	0
Distance	0.0475	0.0307	0.0647	0
Predictability	−0.0266	−0.0442	−0.0078	0.9958
Expt	0.0138	−0.0425	0.0653	0.299
Distance × Predictability	−0.0054	−0.0203	0.0101	0.761
Distance × Expt	0.0073	−0.0075	0.0219	0.1558
Predictability × Expt	0.0063	−0.0069	0.0193	0.171
Pred × Dist × Expt	0.0128	−0.0012	0.0264	0.0357

in **Table 9**. There isn't any convincing evidence for an interaction between distance and experiment; there is only weak evidence for a larger effect of distance in Experiment 2. We cannot therefore argue for a qualitative difference in the distance effects found in Experiments 1 vs. 2.

In Experiment 2, the intervener was a long, uninterrupted prepositional phrase whereas in Experiment 1, the intervener consisted of a short RC followed by a PP. One can speculate as to why Experiment 2 shows equally strong distance effects in both predictability conditions: processing a single long intervening phrase may be harder than processing two different phrases because it may be harder to chunk a single long phrase compared to two shorter phrases; this is predicted by the Sausage Machine proposal of Frazier and Fodor (1978). If this is correct, then the complexity of the intervener may indeed be a relevant factor in determining whether strong expectation can weaken locality effects. It is possible to test this claim by using an intervener that is much easier to process; an example would be an adverb containing no noun phrases.

We were motivated by the recent replication crisis in psychology (Open Science Collaboration, 2015) to attempt to replicate our results using a different method. Furthermore, replications using ET would be very informative because it is possible that SPR overburdens the working-memory system in an unnatural manner. If this is the case, one prediction would be that the ET data would not necessarily show locality effects. We describe these experiments next.

6. EXPERIMENT 3

6.1. Method

6.1.1. Participants

Forty participants, with the same criteria for inclusion as in the previous experiments, participated in the ET study in University of Potsdam, Germany.

6.1.2. Materials

The experimental items were exactly the same as Experiment 1 (SPR), except that the following four items from Experiment 1 were removed: item id 5, *sheka:yat kardan* (complain + to do), item id 9, *sahm bordan* (share + to win), item id 26, *pishraft kardan* (progress + to do), and item id 32, *hes kardan* (feel + to do). The reason for removal was that the results of the sentence-completion studies suggested that these light verbs had lower predictability than the other light verbs in the stimuli. It could be that this lower predictability is due to the existence of some other alternative light verbs with which the nominal part can combine to make other possible CPs. The last two CPs also had a lower acceptability rating (item 26 had 4.7, and item 32 had 3.5). As a consequence, in our ET study, we had thirty-two experimental items and 64 fillers. All items, including fillers are available in the supplementary materials.

6.1.3. Procedure

An ET study was prepared using Experiment-Builder software, and participants' eye-movements were recorded using an EyeLink 1000 tracker, with a connection to a PC. Before the experiment started, the participants were instructed to read the sentences silently at a normal pace and had a practice block consisting of five sentences. After answering the comprehension questions of the practice block, they were provided with feedback indicating whether or not the answer was correct. A 21-inch monitor was placed 60 centimeters from the participants' eyes. In order to reduce head movements, the participants were asked to use the chin-rest. They viewed the sentences with both eyes, but only the right eye was recorded. The items were presented in one line and in 18 points Persian Arial font (from right to left). First, they had to fixate on a dot at the right edge of the screen so that the sentence appeared. After they finished reading, they had to fixate on the dot in the bottom left corner of the screen; once they fixated on the dot, the comprehension question was presented. Unlike the practice items, they were not provided with any feedback. Calibration was performed at the beginning of the experiment, after their 5-min break (which occurred after they had were halfway through the experiment), and whenever it was necessary.

6.1.4. Data Analysis

Raw gaze duration data was obtained using the Data Viewer software[4]. This data was then processed to get different ET measures using the `em2` package (Logačev and Vasishth, 2014). As discussed earlier, Bayesian linear mixed models were used for the analysis. All analyses were carried out using log-transformed data. Zero ms reading times were removed before carrying out the analysis.

6.2. Results

6.2.1. Comprehension Accuracy

On average, participants correctly answered 92% of the target comprehension questions. Mean accuracy by condition was 91 % for condition a, 91% for condition b, 95% for condition c, and 89% for condition d. We found no effects of distance and predictability, and no interaction.

6.2.2. Reading Time

The critical region was the verb, as in Experiments 1 and 2. The same sum contrast coding was used as in Experiments 1 and 2; in addition, nested contrast coding was used to investigate the effect of distance within the two predictability conditions. We present results for first-pass reading time and regression path duration.

[4]http://www.sr-research.com/dv.html

TABLE 10 | Means, 95% uncertainty intervals, and $P(b < 0)$, the probability of the estimate being less than 0, in the reading time analysis for Experiment 3.

ET measure	Comparison	Mean	Lower	Upper	$P(b < 0)$
FPRT	Intercept	5.623	5.5627	5.6833	0
	Distance	0.0504	0.0123	0.0868	0.0062
	Predictability	−0.0522	−0.0844	−0.0189	0.9968
	Distance × Predictability	−0.01	−0.039	0.0196	0.7455
RPD	Intercept	5.7286	5.646	5.8105	0
	Distance	0.0331	−0.0139	0.0814	0.074
	Predictability	−0.0754	−0.1248	−0.0265	0.9992
	Distance × Predictability	−0.0032	−0.0374	0.0316	0.5742

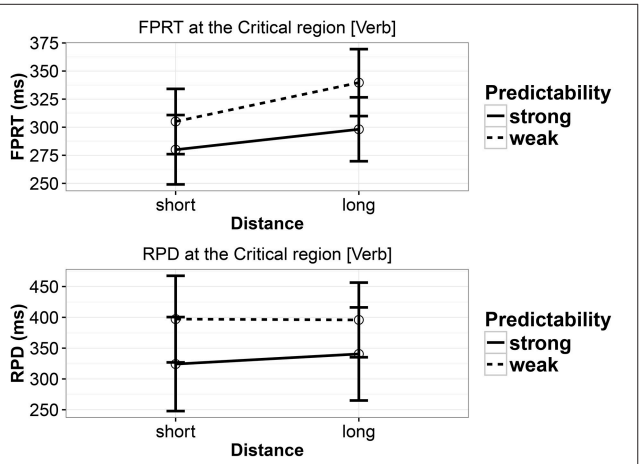


FIGURE 3 | First-pass reading time and regression path duration in Experiment 3 at the critical verb. Error bars show 95% confidence intervals.

The effect of predictability, seen in Experiments 1 and 2, is also present in first-pass reading time (FPRT) and regression path duration (RPD); the strong-predictability conditions had shorter reading times. There was also an effect of distance in FPRT but only a weak effect in RPD; the long-distance conditions had longer reading times. **Figure 3** and **Table 10** show the details of the analyses. A nested contrast showed that in FPRT the distance effect was present in the weak-predictability conditions (coef. = 0.0613, [0.0155,0.1081], and $P(b < 0) = 0.004$); in the strong-predictability conditions the effect was weak (coef. = 0.0423, [−0.0022,0.0865], and $P(b < 0) = 0.0318$). RPD showed only a weak effect of distance within the two predictability levels. For the weak-predictability level, coef. = 0.0359, [−0.0169,0.0884], $P(b < 0) = 0.0891$; and for the strong-predictability level, coef. = 0.0294, [−0.0305,0.0883], $P(b < 0) = 0.1596$.

6.3. Discussion

In the ET Experiment 3, we replicated the locality effects found in the Experiment 1 in first-pass reading time. Nested contrasts showed that the locality effect appeared in weak-predictability conditions, which is similar to the result in Experiment 1. A main effect of predictability was found in FPRT and RPD, replicating the effect in Experiment 1.

Since we failed to find any interaction between predictability and distance, we cannot conclude, as Husain et al. (2014) did, that expectation effects can cancel locality effects. The locality effects are consistent with working memory accounts (Gibson, 2000; Lewis and Vasishth, 2005) and inconsistent with the distance-based predictions of the expectation account (Levy, 2008). As in the SPR experiments, we have evidence consistent with a version of the expectation account that predicts that strong predictability conditions will be read faster than the weak predictability conditions.

In sum, the main result in Experiment 3 is that we have replicated the locality effect and the facilitation due to strong predictability.

7. EXPERIMENT 4

7.1. Method

7.1.1. Participants

Forty participants, with the same criteria as in the previous experiments, participated in the ET study in Golm campus, University of Potsdam, Germany.

7.1.2. Materials

The experimental items were exactly the same as Experiment 2 (SPR), but with 32 items (see the explanation for Experiment 3 regarding the four items that were removed). The experimental items were complemented with 64 filler sentences with varying syntactic structures (see supplementary materials).

7.1.3. Procedure and Data Analysis

The procedure and data analysis were exactly the same as Experiment 3 (ET).

7.2. Results

7.2.1. Comprehension Accuracy

On average, participants answered 90% of comprehension questions correctly. They had 94% response accuracy for condition a, 88% for condition b, 94% for condition c, and 86% for condition d. None of the factors had an effect on accuracy.

7.2.2. ET Measures

The reading times at the critical region are summarized in **Figure 4**. Unlike Experiment 3, in the current experiment, we found effects of distance and predictability in both the measures (see **Table 11**). In other words, in the two measures reported, the long conditions (b and d) were read slower than the short conditions (a and c), and the weak predictability conditions (c and d) were read slower than the strong predictability conditions (a and b). None of the measures showed any interaction between predictability and distance.

Nested comparisons showed that in first-pass reading time, the locality effect was seen in the strong-predictability condition (coef. = 0.0507, [0.0011, 0.1003], $P(b < 0) = 0.022$), but there

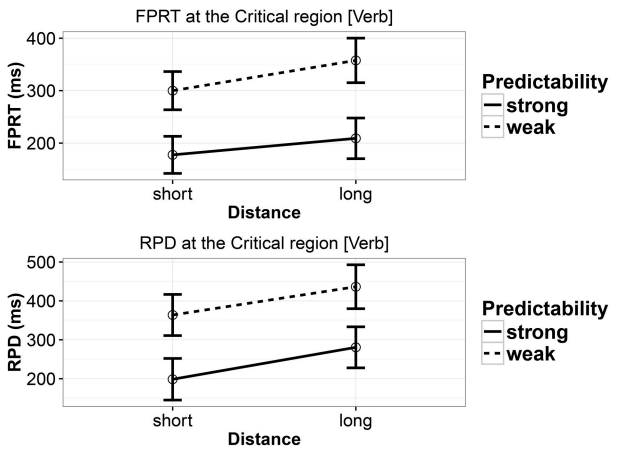


FIGURE 4 | First-pass reading time and regression path duration in Experiment 4 at the critical verb. Error bars show 95% confidence intervals.

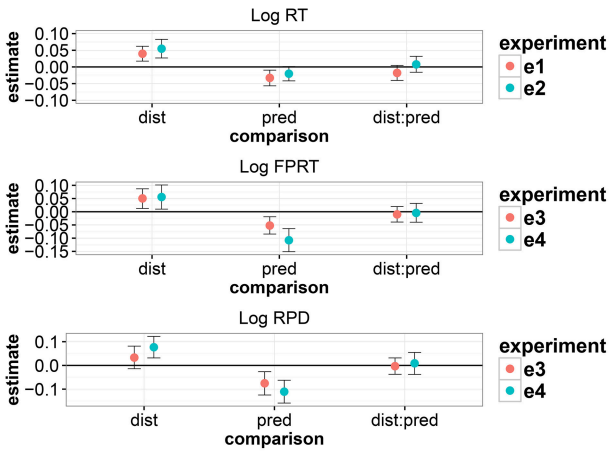


FIGURE 5 | Summary of the magnitudes of effects (derived from the linear mixed models) across the four experiments. The error bars show 95% uncertainty intervals and show the range within which we can be 95% certain that the true parameter lies given the data.

TABLE 11 | Means, 95% uncertainty intervals, and $P(b < 0)$, the probability of the estimate being less than 0, in the reading time analysis for Experiment 4.

ET measure	Comparison	Mean	Lower	Upper	$P(b < 0)$
FPRT	Intercept	5.6731	5.6015	5.7448	0
	Distance	0.0557	0.0099	0.1013	0.0082
	Predictability	−0.1079	−0.1512	−0.0638	1
	Distance × Predictability	−0.0046	−0.0397	0.0315	0.6088
RPD	Intercept	5.7958	5.7113	5.8799	0
	Distance	0.0767	0.0316	0.1225	5e-04
	Predictability	−0.1108	−0.1588	−0.0626	1
	Distance × Predictability	0.0089	−0.0381	0.0547	0.3452

was a weaker tendency toward a locality effect in the weak-predictability condition (coef. = 0.061, [−0.0046, 0.1261], $P(b < 0) = 0.0355$). In regression-path duration, both strong- and weak-predictability conditions showed a locality effect (strong-predictability: coef. = 0.0858, [0.0253, 0.1492], $P(b < 0) = 0.0031$; low-predictability: coef. = 0.0675, [0.0027, 0.1317], $P(b < 0) = 0.0211$.

7.3. Discussion

Experiment 4 replicated the results of Experiment 2: a main effect of distance and a main effect of predictability, with no evidence for an interaction. The effects in FPRT and RPD showed essentially the same patterns as in the first ET study. However, the locality effects were even stronger, in the same way that the second SPR study showed stronger locality effects. Also, these effects are equally strong in both strong and weak predictability conditions, mirroring our finding in the second SPR study.

Overall, regarding the distance manipulation, the results are consistent with memory-based accounts, and inconsistent with the expectation account. The main effect of predictability is consistent with the expectation account, as discussed earlier. In Experiment 4, we don't see any evidence consistent with the Husain et al. (2014) proposal; if anything, the locality effect is *stronger* in the strong-predictability conditions.

8. GENERAL DISCUSSION

As summarized graphically in **Figure 5**, our main finding from the four Persian studies is that the locality effect predicted by memory accounts is upheld, but there is no evidence for the expectation-based account's prediction of facilitation in longer distance conditions. We consistently see a main effect of predictability, which is consistent with expectation accounts. Finally, there is no compelling evidence in the Persian data that strong expectations cancel locality effects.

There is also suggestive evidence that the complexity of intervening material could strengthen the locality effect: when the intervener is an RC followed by a PP, we see a marginal interaction between distance and predictability, but when the intervener is a single long PP, we see no evidence for an interaction between distance and predictability strength, and we tend to see stronger effects.

We consistently found a main effect of predictability in all four experiments: the strong predictability conditions were read faster at the verb than the weak predictability conditions. This is consistent with the expectation-based account. Since the verbs in the strong and weak predictability conditions are not identical, we cannot rule out the possibility that word frequency or other such low-level factors are responsible for these effects. However, it is plausible that the highly predictable verb is processed faster than the less predictable verb. Thus, the main effect of predictability can be seen as evidence for expectation-based accounts, operationalized in terms of the conditional

probabilities of the appearance of the exact verb given the preceding context.

It is possible that we were unable to replicate Husain et al's findings because of the nature of the intervener used in the Persian studies. Unlike, Husain et al. (2014) where the long distance condition had extra adverbials compared to the short condition, in Experiment 1 we have a more complex intervener, a relative clause. Another reason for finding the effects which are different from the study by Husain et al. (2014) could be that in Persian, separating the nominal part of the CP from the light verb occurs relatively rarely, compared to Hindi. There is some support for this in corpus data. Based on the Hindi dependency treebank (Bhatt et al., 2009), the average distance, counted as the number of intervening phrases, between an object and its (heavy) verb is 0.82 (with minimum 0 and maximum 15, and first and third quantiles 0 and 1), and the average distance between a noun and light verb is 0.07 (minimum 0 and maximum 18, with first and third quantiles 0 and 0). In the Persian dependency treebank (Seraji, 2015), the average distance between an object and (heavy) verb is 2.48 (with minimum 0 and maximum 9, and first and third quantiles 1 and 3), while the average distance between a noun and light verb is 0.05 (with minimum 0, and maximum 6, and first and third quantiles 0 and 0). Thus, the adjacency of CPs in Persian is strongly preferred (maximum 6 vs. Hindi's maximum 18), although as validated in the acceptability rating norming study, this separability is apparently acceptable and not considered ungrammatical[5].

8.1. An Alternative Explanation of Locality Effects in Terms of Entropy

Could there be an alternative explanation for the locality effect seen in the four experiments, one that does not invoke greater memory cost in the long-distance conditions? One possibility is that entropy (uncertainty) increases with increasing distance. Entropy is an information-theoretic measure that essentially represents how uncertain we are of the outcome (Shannon, 2001). In the present case, this would translate to our uncertainty about the upcoming verb. If there are n possible ways to continue a sentence, and each of the possible ways has probability p_i, where $i = 1, \ldots, n$, then entropy is defined (Shannon, 2001) as $-\sum_i p_i \times \log_2(p_i)$. The entropy associated with the upcoming verb can be calculated using our offline sentence completion data[6].

8.1.1. Evaluating the Effect of Entropy

In order to evaluate whether entropy could explain the locality data, we computed entropy for each item in each condition for both experiments. The estimated entropies for each condition in the two experiment designs are shown in **Figure 6**. It is important to note here that entropy for each condition in **Figure 6** is based on only nine data points per condition (we only have $9 \times 4 = 36$ items); for different items, there is substantial variability in the entropy patterns by condition. Nevertheless, in the figure we can see that in the items used for Experiments 1 and 3, the entropy is higher in the long-distance conditions. The effect of entropy is less clear for the items used in Experiments 2 and 4, because of the relatively wider confidence intervals. Clearly uncertainty is higher in the RC+PP experiment than in the long PP experiment. A closer look at the high predictability conditions shows that the entropy difference between the long and short distance conditions is larger in the RC+PP intervener items than the entropy difference in the long PP intervener items (it is larger by 0.14, with 95% uncertainty intervals −0.01 and 0.28, probability of the difference in entropy being less than 0 is 0.03). This is suggestive—if weak—evidence that the intervening RC may be responsible for creating a greater degree of uncertainty regarding the upcoming verb. This is a bit surprising because stronger locality effects were seen in the long PP experiments.

In order to investigate whether entropy affects reading times at the verb, we fit a maximal Bayesian linear mixed model with predicate type and distance as sum-coded factors, and entropy (centered) as a continuous predictor; all higher order interactions were also included. The dependent variable was log reading time at the critical verb. As shown in **Table 12**, in Experiment 1, in addition to the effects of predictability and distance, we find an effect of entropy, and an interaction between distance and entropy, such that long distance conditions lead to a greater effect of entropy. None of the other experiments showed any effects of entropy. Thus, although the evidence in favor of entropy is far from overwhelming, a potentially important finding here is that entropy could explain locality effects at least in our Experiment 1. To our knowledge, this is the first demonstration

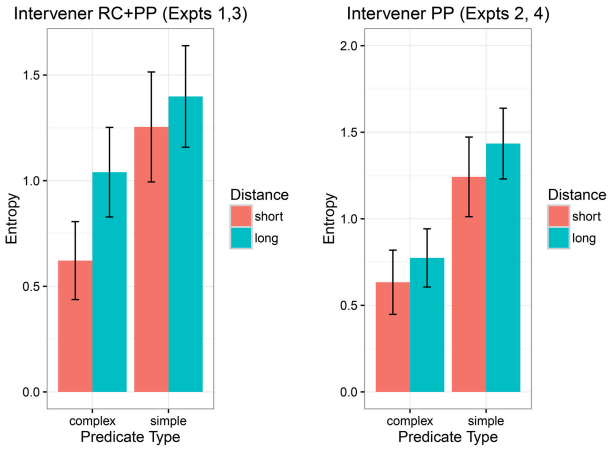


FIGURE 6 | The estimated entropy (with 95% confidence intervals), computed using the sentence completion data, for the two experiment designs.

[5]These intervening phrases have been computed using dependency treebanks. Consequently, phrasal boundaries are approximations. Also, because of annotation differences between the two treebanks, phrase boundary criteria sometimes differ for the two languages. The phrasal counts lead to the same conclusions regardless of whether one counts intervening phrases or words.

[6]See Linzen and Jaeger (2015) for a recent empirical investigation of entropy in sentence comprehension using corpus data instead of sentence completion data. Linzen and Jaeger calculated entropy in several ways, and also evaluated another metric called entropy reduction (ER), which was proposed by Hale, 2006; however, we cannot evaluate ER here because that would require knowing the entropy for the word preceding the verb.

TABLE 12 | Model results from the Bayesian linear mixed model for the effect of entropy (apart from other predictors) on log reading times in Experiment 1.

Comparison	Mean	Lower	Upper	$P(b < 0)$
Intercept	6.23	6.16	6.3	0
Predictability	−0.02	−0.04	0	0.95
Distance	0.03	0.01	0.06	0.01
Entropy	0.05	0.02	0.09	0
Pred:Dist	−0.01	−0.03	0.02	0.71
Pred:Entropy	−0.02	−0.06	0.01	0.88
Dist:Entropy	0.04	0.01	0.07	0.01
Pred:Dist:Ent	0	−0.03	0.04	0.45

Shown are the mean and 95% uncertainty intervals, and the probability of the parameter being less than 0.

that locality effects may arise due to factors other than memory costs.

But why does entropy increase in longer-distance dependencies? A possible explanation suggests itself in terms of memory overload causing forgetting. It is possible that the participants forgot that a noun-verb dependency exists in the long-distance complex predicate condition. One prediction of this forgetting-inducing-entropy account would be that in the sentence completion study, participants would tend to produce more ungrammatical continuations in the long-distance condition than the short-distance condition. This is borne out in Experiment 1: the accuracy in the short condition was 97%, and in the long condition it was 92%. A Bayesian generalized linear mixed model was fit with a full variance-covariance matrix for participants and items[7]. The results of the model fit showed a reduction in grammaticality of sentence completions in the long vs. short conditions; the log odds were −0.9436 [−2.042, −0.1284], with a probability of the log odds being negative being 0.99. The sentence completion study corresponding to Experiment 2 (which showed no effect of entropy on reading times) showed no difference in grammaticality of completions; the short and long conditions had grammatical continuations with the proportions 0.97 and 0.98.

Thus, it is possible that in Experiment 1, the increase in entropy is due to participants forgetting the left context partially. Clearly, a planned experiment is called for to investigate this further. An important point to note here is that the increased entropy in the long-distance condition may be a *consequence* of forgetting, not a cause in itself: entropy itself would not predict any increase in ungrammatical continuations, but the forgetting hypothesis does.

8.1.2. Does Predictability of an Upcoming Verb Increase with Distance?

We showed above that increasing uncertainty about the upcoming verb may explain locality, at least in Experiment 1. One important question that arises, especially in the strong predictability conditions, is the following: does increasing distance nevertheless sharpen the expectation for the verb, as suggested by Konieczny (2000)? In order to address this question, we fit a Bayesian generalized linear mixed model (GLMM) with a logistic link function that investigated the change in probability mass for the target verb as a function of distance in the strong predictability conditions.

For the first sentence completion study (which had the RC+PP intervener in the long condition), in the long-distance condition, the probability of producing the target verb fell: on the log-odds scale, the mean and 95% uncertainty interval were −0.305 [−0.8127, 0.1591] and the posterior probability of the reduction being less than 0 was 0.9. The odds ratio of producing a target verb in the long vs. short condition was 0.74, with 95% uncertainty interval [0.44,1.17]. This means that in the long condition, participants are less likely to produce the target verb, but since the uncertainty interval for the odds ratio includes 1, the reduction in probability of target verb production is possibly unchanged in the short vs. long distance conditions. If anything, there is a weakening of the expectation for the target verb, contrary to the sharpened expectation proposal of Konieczny (2000).

For the second sentence completion study (which had a PP in the long condition), in the long-distance condition, the probability of producing the target verb also fell: the logs odds were −0.17 [−0.5,0.12]; and the posterior probability of the reduction being less than 0 was 0.86. The odds ratio of producing a target verb in the long vs. short condition was 0.84, with 95% uncertainty intervals [0.61, 1.13]. Thus, in the second sentence completion study, there is only weak evidence of a reduction in probability of producing the target verb in the long-distance condition.

To summarize, our sentence completion data for Experiments 1 and 2's strong predictability condition show that increasing distance tends to reduce the proportion of target verbs produced, although the evidence for this reduction is rather weak overall. Our data from Persian therefore seem to go against the suggestion by Konieczny (2000) that increasing distance leads to narrowing down the prediction to the target verb.

Caution is needed in interpreting these results based on the sentence completion data. The biggest issue with the sentence completion data is that it was an offline task; it is difficult to argue that offline completion data can inform us about online processes. It would be much more informative to run an online sentence completion study, forcing participants to make quicker decisions about the sentence completions. Further, most of our

[7]The predictor (short vs. long condition) was coded using sum contrasts, with the long condition coded as 1 and the short condition as −1; the dependent variable was binary and represented whether a target verb was produced by the participant for a particular item-condition combination or not. Participants and items were specified as partially crossed random factors, and a full variance-covariance matrix was fit for both random effects. The priors for the intercept and slope were the Student's t-distribution with 2 degrees of freedom, allowing a range of approximately −10 to 10 on the log odds scale, with 0 the most likely value. The prior on the variance-covariance matrices was defined via the LKJ prior (Stan Development Team, 2013, 2014) on the correlation matrix; see Sorensen and Vasishth (2015) for a tutorial intended for psycholinguists and cognitive scientists. The model was fit using the `stan_lmer` function from the `rstanarm` package (Gabry and Goodrich, 2016).

findings relating to the sentence completion data are *post-hoc* and based on exploratory analyses. It would also be very informative to carry out sentence completion studies for experiments such as those of Konieczny (2000); Grodner and Gibson (2005); Vasishth and Lewis (2006); Bartek et al. (2011); Vasishth and Drenhaus (2011); Levy and Keller (2013) in order to establish whether increasing distance can weaken expectation cross-linguistically.

In future work it may be worth investigating existing locality effects in English, German, and Hindi from the perspective of forgetting inducing entropy. A further possibility worth investigating is whether entropy reduction (Hale, 2006) rather than entropy can explain the locality effects cross-linguistically. In our Persian experiments, it is possible that the entropy at the word preceding the verb is higher than the entropy at the verb, and it is possible that the reduction in entropy is larger in the long-distance condition. Unfortunately, we have no way to test this in the present design, but future studies could compute entropy reduction empirically in the same way that we computed entropy using sentence completion data. Thus, in principle it is possible that entropy reduction could explain locality effects as well. A related issue that would then arise is whether entropy or entropy reduction furnishes a better explanation for locality effects.

A broader issue that the above discussion raises is, can all intervention effects be explained via an appeal to information-theoretic metrics? Levy (2008) had pointed out that information-theoretic metrics cannot explain all the results relating to intervention effects; he was mainly referring to locality effects, which can only be explained through memory-based accounts. In later work, Vasishth and Drenhaus (2011); Levy and Keller (2013); Levy et al. (2013) also find that both memory and expectation-based accounts are needed to explain the range of observed effects. It is because of the inability of information-theoretic metrics to explain locality effects that Levy (2008) argued for "two-factor" accounts. If entropy or some other entropy-based measure turns out be an explanation for locality effects, can we argue for a simpler account that only appeals to information-theoretic metrics? A major empirical problem for such a reductionist account would be the large range and variety of intervention effects (see Engelmann et al., Manuscript submitted). for a review and computational modeling) that can only be explained through memory-based accounts. Other recent results that would be impossible to explain via a reductionist account are the work by Nicenboim et al. (2015) and Nicenboim et al. (2016). Thus, a reductionist account that assumes that *all* effects can be explained by what is predicted next would always falter when it comes to explaining effects that arise not from predictive processes but from retrieval-based processes.

9. CONCLUDING REMARKS

In conclusion, as regards the distance manipulation, the evidence from Persian is in favor of working-memory accounts, although forgetting-causing-entropy is also a candidate explanation. There is not much evidence from Persian that strong-predictability conditions cancel locality effects, as Husain and colleagues had suggested. Interestingly, there is no evidence in these experiments for the prediction of the expectation account regarding the distance manipulation, that increasing argument-verb distance facilitates processing due to increasing conditional probabilities of the upcoming verb. The suggestion in Levy et al. (2013) that "the verb-medial languages tend to exhibit the general patterns predicted by memory-based theories, whereas verb-final languages tend to exhibit the general patterns predicted by expectation-based theories" seems to be difficult to maintain (also see Husain et al., 2015, for locality effects in Hindi). One implication of our findings from Persian is that locality and expectation effects observed across studies seem to be highly conditional on the language and syntactic construction being considered—broad cross-linguistic generalizations may be difficult to make.

AUTHOR CONTRIBUTIONS

MS, SV, and SH designed the experiments; MS prepared the items, recruited participants, and conducted all the studies reported; SH conducted all the corpus analyses and extracted the statistics from the corpora; SV, MS, and SH analyzed the data and prepared the final document.

FUNDING

This work was supported by the IDEALAB program, and the University of Potsdam. We acknowledge the support of the Deutsche Forschungsgemeinschaft (German Research Foundation) and Open Access Publication Fund of Potsdam University.

ACKNOWLEDGMENTS

Thanks to Prof. Dr. Shahla Raghibdoust in Allameh Tabataba'i University who helped the first author to recruit the participants in Iran. We are grateful to Carla Kessler in University of Potsdam for her help in designing the eye-tracking study using Experiment-Builder software. Many thanks to Lena Jäger who carried out a careful sanity check for the eye-tracking results. We would like to thank the anonymous reviewers of this article, and the audience in the 28th CUNY conference of human sentence processing in University of Southern California for their insightful feedback.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpsyg.2016.00403

REFERENCES

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: keep it maximal. *J. Mem. Lang.* 68, 255–278. doi: 10.1016/j.jml.2012.11.001

Bartek, B., Lewis, R. L., Vasishth, S., and Smith, M. R. (2011). In search of on-line locality effects in sentence comprehension. *J. Exp. Psychol. Learn. Mem. Cogn.* 37, 1178. doi: 10.1037/a0024194

Bates, D., Kliegl, R., Vasishth, S., and Baayen, H. (2015). Parsimonious mixed models. *arXiv preprint*. arXiv:1506.04967.

Bhatt, R., Narasimhan, B., Palmer, M., Rambow, O., Sharma, D. M., and Xia, F. (2009). "A multi-representational and multi-layered treebank for Hindi/Urdu," in *Proceedings of the Third Linguistic Annotation Workshop*, (Stroudsburg, PA: Association for Computational Linguistics), 186–189. doi: 10.3115/1698381.1698417

Caplan, D., and Waters, G. (2013). Memory mechanisms supporting syntactic comprehension. *Psychon. Bull. Rev.* 20, 243–268. doi: 10.3758/s13423-012-0369-9

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge, MA: MIT

Dabir-Moghaddam, M. (1997). Compound verbs in Persian. *Stud. Linguist. Sci.* 27, 25–59.

Frazier, L., and Fodor, J. D. (1978). The sausage machine: a new two-stage parsing model. *Cognition* 6, 291–325. doi: 10.1016/0010-0277(78)90002-1

Gabry, J., and Goodrich, B. (2016). *rstanarm: Bayesian Applied Regression Modeling via Stan*. R package version 2.9.0-3.

Gibson, E. (1998). Linguistic complexity: locality of syntactic dependencies. *Cognition* 68, 1–76. doi: 10.1016/S0010-0277(98)00034-1

Gibson, E. (2000). "Dependency locality theory: a distance-based theory of linguistic complexity," in *Image, Language, brain: Papers from the First Mind Articulation Project Symposium*, eds A. Marantz, Y. Miyashita, and W. O'Neil (Cambridge, MA: MIT Press), 95–126.

Grodner, D., and Gibson, E. (2005). Consequences of the serial nature of linguistic input for sentential complexity. *Cogn. Sci.* 29, 261–290. doi: 10.1207/s15516709cog0000_7

Hale, J. (2001). "A probabilistic Earley parser as a psycholinguistic model," in *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies*, (Stroudsburg, PA: Association for Computational Linguistics), 1–8. doi: 10.3115/1073336.1073357

Hale, J. (2006). Uncertainty about the rest of the sentence. *Cogn. Sci.* 30, 643–672. doi: 10.1207/s15516709cog0000_64

Hsiao, F. P.-F., and Gibson, E. (2003). Processing relative clauses in Chinese. *Cognition* 90, 3–27. doi: 10.1016/S0010-0277(03)00124-0

Husain, S., Vasishth, S., and Srinivasan, N. (2014). Strong expectations cancel locality effects: evidence from hindi. *PLoS ONE* 9:e100986. doi: 10.1371/journal.pone.0100986

Husain, S., Vasishth, S., and Srinivasan, N. (2015). Integration and prediction difficulty in Hindi sentence comprehension: evidence from an eye-tracking corpus. *J. Eye Mov. Res.* 8, 1–12. doi: 10.16910/jemr.8.2.3

Jaeger, F. T., Fedorenko, E., Hofmeister, P., and Gibson, E. (2008). "Expectation-based syntactic processing: antilocality outside of head-final languages," in *CUNY Sentence Processing Conference*, (North Carolina).

Jäger, L., Chen, Z., Li, Q., Lin, C.-J. C., and Vasishth, S. (2015). The subject-relative advantage in Chinese: evidence for expectation-based processing. *J. Mem. Lang.* 79–80, 97–120. doi: 10.1016/j.jml.2014.10.005

Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognition* 20, 137–194. doi: 10.1207/s15516709cog2002_1

Just, M., and Carpenter, P. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychol. Rev.*, 99, 122–149. doi: 10.1037/0033-295X.99.1.122

Karimi-Doostan, G. (1997). *Light Verb Constructions in Persian*. Ph.D thesis, University of Essex.

Karimi-Doostan, G. (2005). Light verbs and structural case. *Lingua* 115, 1737–1756. doi: 10.1016/j.lingua.2004.08.002

Karimi-Doostan, G. (2011). Separability of light verb constructions in Persian. *Stud. Linguist.* 65, 70–95. doi: 10.1111/j.1467-9582.2011.01178.x

Konieczny, L. (2000). Locality and parsing complexity. *J. Psycholinguist. Res.* 29, 627–645. doi: 10.1023/A:1026528912821

Kutas, M., and Hillyard, S. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature* 307, 161–163. doi: 10.1038/307161a0

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition* 106, 1126–1177. doi: 10.1016/j.cognition.2007.05.006

Levy, R., Fedorenko, E., and Gibson, E. (2013). The syntactic complexity of Russian relative clauses. *J. Mem. Lang.* 69, 461–495. doi: 10.1016/j.jml.2012.10.005

Levy, R. P., and Keller, F. (2013). Expectation and locality effects in German verb-final structures. *J. Mem. Lang.*, 68, 199–222. doi: 10.1016/j.jml.2012.02.005

Lewis, R. L., and Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cogn. Sci.* 29, 375–419. doi: 10.1207/s15516709cog0000_25

Linzen, T., and Jaeger, T. F. (2015). Uncertainty and expectation in sentence processing: evidence from subcategorization distributions. *Cogn. Sci.* doi: 10.1111/cogs.12274. [Epub ahead of print].

Logačev, P., and Vasishth, S. (2014). *The em2 Package for Computing Eyetracking Measures*. Potsdam: University of Potsdam.

Nicenboim, B., Logačev, P., Gattei, C., and Vasishth, S. (2016). When high-capacity readers slow down and low-capacity readers speed up: working memory differences in unbounded dependencies. *Front. Psychol.* 7:00280. doi: 10.3389/fpsyg.2016.00280

Nicenboim, B., and Vasishth, S. (2016). Statistical methods for linguistic research: foundational Ideas-Part II. *arXiv preprint*. arXiv:1602.00245.

Nicenboim, B., Vasishth, S., Gattei, C., Sigman, M., and Kliegl, R. (2015). Working memory differences in long-distance dependency resolution. *Front. Psychol.* 6:312. doi: 10.3389/fpsyg.2015.00312

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science* 349, aac4716. doi: 10.1126/science.aac4716

R Development Core Team (2013). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Samvelian, P. (2001). *The Syntax of Iranian Languages, Canadian Journal of Linguistics*. Toronto:University of Toronto Press.

Samvelian, P. (2007). A (phrasal) affix analysis of the Persian Ezafe. *J. Linguist.* 43, 605–645. doi: 10.1017/S0022226707004781

Seraji, M. (2015). *Morphosyntactic Corpora and Tools for Persian*. Ph.D thesis, Uppsala University.

Shannon, C. E. (2001). A mathematical theory of communication. *ACM Sigmobile Mob. Comput. Commun. Rev.* 5, 3–55. doi: 10.1145/584091.584093

Sorensen, T., and Vasishth, S. (2015). Bayesian linear mixed models using Stan: a tutorial for psychologists, linguists, and cognitive scientists. *arXiv preprint*. arXiv:1506.06201.

Stan Development Team (2013). Stan: A C++ library for probability and sampling, version 2.1.

Stan Development Team (2014). *Stan Modeling Language Users Guide and Reference Manual, Version 2.2*.

Staub, A. (2010). Eye movements and processing difficulty in object relative clauses. *Cognition* 116, 71–86. doi: 10.1016/j.cognition.2010.04.002

Vahedi-Langrudi, M.-M. (1996). *The Syntax, Semantics and Argument Structure of Complex Predicates in Modern Farsi,* Ottawa, ON: University of Ottawa.

Vasishth, S. (2003). *Working Memory in Sentence Comprehension: Processing Hindi Center Embeddings*. Routledge, NY: Routledge.

Vasishth, S., and Drenhaus, H. (2011). Locality in German. *Dialog. Disc.* 2, 59–82. doi: 10.5087/dad.2011.104

Vasishth, S., and Lewis, R. L. (2006). Argument-head distance and processing complexity: explaining both locality and antilocality effects. *Language* 82, 767–794. doi: 10.1353/lan.2006.0236

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2016 Safavi, Husain and Vasishth. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 16 October 2015
doi: 10.3389/fpsyg.2015.01552

Poor readers' retrieval mechanism: efficient access is not dependent on reading skill

Clinton L. Johns[1]*, *Kazunaga Matsuki*[1,2] and *Julie A. Van Dyke*[1]

[1] Haskins Laboratories, New Haven, CT, USA, [2] Department of Linguistics and Language, McMaster University, Hamilton, ON, Canada

A substantial body of evidence points to a cue-based direct-access retrieval mechanism as a crucial component of skilled adult reading. We report two experiments aimed at examining whether poor readers are able to make use of the same retrieval mechanism. This is significant in light of findings that poor readers have difficulty retrieving linguistic information (e.g., Perfetti, 1985). Our experiments are based on a previous demonstration of direct-access retrieval in language processing, presented in McElree et al. (2003). Experiment 1 replicates the original result using an auditory implementation of the Speed-Accuracy Tradeoff (SAT) method. This finding represents a significant methodological advance, as it opens up the possibility of exploring retrieval speeds in non-reading populations. Experiment 2 provides evidence that poor readers do use a direct-access retrieval mechanism during listening comprehension, despite overall poorer accuracy and slower retrieval speeds relative to skilled readers. The findings are discussed with respect to hypotheses about the source of poor reading comprehension.

Keywords: memory retrieval, sentence processing, speed-accuracy trade-off, reading comprehension, individual differences

OPEN ACCESS

Edited by:
Matthew Wagers, University of California, Santa Cruz, USA

Reviewed by:
Brian Dillon, University of Massachusetts Amherst, USA
Andrea Eyleen Martin-Nieuwland, The University of Edinburgh, UK

***Correspondence:**
Clinton L. Johns
johns@haskins.yale.edu

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: *10 July 2015*
Accepted: *25 September 2015*
Published: *16 October 2015*

Citation:
Johns CL, Matsuki K and Van Dyke JA (2015) Poor readers' retrieval mechanism: efficient access is not dependent on reading skill. Front. Psychol. 6:1552. doi: 10.3389/fpsyg.2015.01552

INTRODUCTION

The ability to comprehend written language is an enormously important skill, as shown by robust correlations between poor reading comprehension and a variety of undesirable consequences, including constrained economic mobility, reduced economic success, and increased risk of poor health outcomes (Kutner et al., 2007; National Institute for Literacy, 2008). Many models of text processing (e.g., Kintsch, 1988, 1998; Myers and O'Brien, 1998; van den Broek et al., 1999; for reviews see Long et al., 2006; McNamara and Magliano, 2009), sentence processing (Gibson, 2000; Van Dyke and Lewis, 2003; for review see Van Gompel, 2013), and reading disability (e.g., Hogaboam and Perfetti, 1975; Shankweiler and Crain, 1986) incorporate the idea that comprehension is constrained by the architecture of the human memory system. Given this, it is important to understand the interaction between memory mechanisms—such as retrieval—and the sentence parsing processes on which successful comprehension depends. Previous research has shown that university students employ a content-addressable, direct-access mechanism to efficiently retrieve information from memory during reading (e.g., McElree and Dosher, 1989; McElree, 2001; McElree et al., 2003; for reviews see McElree and Dyer, 2013; McElree, 2015). In this article, we assess the potential relation between reading skill and memory retrieval. We report two speed-accuracy tradeoff (SAT) experiments that (1) validate an auditory implementation of the technique for the assessment of the dynamics of memory retrieval, and (2) investigate whether

poor readers, like skilled readers, are able to employ a content-addressable direct-access retrieval mechanism during online auditory sentence comprehension. Our goal is to determine whether poor comprehenders possess the architectural primitives that are known to support skilled reading comprehension.

Models of sentence parsing typically offer richly detailed accounts of the linguistic processes that drive parsing operations. Examples of such operations include heuristic routines (e.g., minimal attachment, late closure, main assertion preference, active filler strategy); serial (or parallel) control structures, which may or may not activate (or inhibit) competing interpretations; ranked vs. unranked consideration of extra-syntactic (e.g., semantic, referential, pragmatic, visual) information; and so on. When these processes go awry, errors are often either explicitly or implicitly associated with increased demands on the memory system, which are assumed to yield suboptimal application of these parameters. To illustrate, consider sentences (1) and (2), from a study by Frazier and Rayner (1982):

(1) While Susan was dressing the baby played on the floor.
(2) Since Jay always jogs a mile seems like a short distance to him.

Sentences such as (1) and (2) are thought to tax the memory system because an accurate parse, and consequent comprehension, is only possible by violating initial syntactic commitments (licensed by minimal attachment in (1) and late closure in (2); Frazier, 1979, 1987; for review, see Frazier, 2013). In such cases the parser must construct, assess, abandon, and reconstruct entire syntactic structures, and reassessment is assumed to involve costly search and diagnosis processes (e.g., Fodor and Inoue, 1994, 1998, 2000). Or, alternatively, it could be the case that the parser constructs and actively maintains multiple syntactic structures during online processing, the extra burden of which leads to processing difficulty (e.g., MacDonald et al., 1994). Furthermore, it is not just parsing errors that tax the memory system during language processing; complexity effects, in which more complex syntactic structures are claimed to be more memory-intensive, have been widely studied. A classic example is the difference in processing elicited by unambiguous sentences such as (3), which contains an object-extracted relative clause, and (4), in which the embedded subject-extracted relative clause results in a simpler syntactic structure (from King and Just, 1991):

(3) The banker that the barber praised climbed the mountain.
(4) The banker that praised the barber climbed the mountain.

In (3), it is thought that the initial filler noun phrase (*the banker*) must be actively maintained during the processing of the embedded clause, after which it may be integrated with the matrix verb *climbed* (e.g., via an active filler strategy; Clifton and Frazier, 1989); in contrast, (4) elicits no such active maintenance, and consequently is less demanding of the memory system.

These examples highlight the fact that the centrality of memory operations during parsing is both widely acknowledged and uncontroversial. In spite of this, theories of parsing (and text processing) are frequently vague regarding the memory mechanisms that support their finely-specified linguistic operations. Further, when a memory component has been elaborated, the focus has been almost entirely on the *storage* component, which is conceptualized as a limited-capacity working memory (WM) system (e.g., Just and Carpenter, 1992; Caplan and Waters, 1999; see also Daneman and Carpenter, 1980). In this system, the dynamic allocation of resources between language and memory operations must support incremental parsing operations, maintenance of critical sentential information, and retrieval of information from both WM and long-term memory (LTM). A fundamental tenet of this approach is that information required for interpretation must be maintained in an active, highly accessible state, and when this is difficult—perhaps owing to low capacity, or to increased computational costs, or both—processing suffers. There is no shortage of research whose findings have been interpreted as evidence for a capacity-based memory architecture (e.g., King and Just, 1991; Fedorenko et al., 2006; Nieuwland and Van Berkum, 2006, among many others). Implicit in these approaches is the idea that stored information is accessed via a serial search process (Just and Carpenter, 1992; Gibson, 1998, 2000): thus, the greater the amount of linguistic material intervening between dependent constituents (which must, therefore, be searched), the more difficult a given construction will be.

Despite the abundance of psycholinguistic studies that adopt this conception of memory, there is substantial disagreement about the nature of the unit of "active maintenance" that defines these search processes. Various proposals have characterized it as words (Warner and Glass, 1987), discourse referents (Gibson, 2000), incomplete grammatical dependencies (Abney and Johnson, 1991; Gibson, 1998), syntactic embeddings (Miller and Chomsky, 1963), or representations of entire alternative syntactic structures (Just and Carpenter, 1992; MacDonald et al., 1992). The fact that consensus has been elusive indicates the weakness of this approach. In addition, significant practical concerns exist, such as poor test-retest reliability of metrics designed to gauge WM capacity (Waters and Caplan, 2003), and collinearity with many other cognitive measures (e.g., Van Dyke et al., 2014). Further, the approach has also been questioned on theoretical grounds; for example, innate capacity differences that limit comprehension ability could emerge naturally from individual linguistic experience rather than from a separable memory system (e.g., MacDonald and Christiansen, 2002).

However, the most fundamental objection to assuming that a search-based limited-capacity memory mechanism supports language comprehension derives not from the need to reconcile these kinds of inconsistencies, but from the disparity between the proposed WM architecture and the empirical evidence regarding the memory structures and operations themselves. For example, there is substantial evidence that the amount of information that can be maintained in an active, accessible state is far more constrained than has been assumed by any parsing architecture supported by a fixed-capacity WM system. Memory studies using the SAT method report that only a

single item (i.e., the last item processed) is actively maintained, meaning that only this item would not require retrieval (McElree, 1998, 2001, 2006; McElree and Dosher, 2001)[1]. All other items—that is, items that should be both *within* as well as *outside of* a traditional WM span—are accessed 30–50% more slowly than the active item (Wickelgren et al., 1980; McElree, 1996, 1998). Results such as these clearly indicate that the capacity of active memory is limited to information that is currently in the focus of attention, while information that is outside focal attention is passively represented. Moreover, items that are outside of focal attention are accessed with constant speed, regardless of how recently they occurred in relation to the retrieval probe. This pattern is consistent with the operation of a cue-based, direct-access retrieval mechanism in which all available cues are matched simultaneously, with the degree of featural overlap between the target and the available retrieval cues determining retrieval success (for a review see Clark and Gronlund, 1996). While language processing with such a severely constrained active memory capacity may seem implausible, the feasibility of a processing architecture in which only the most recent item remains in focal attention has been demonstrated in an implemented computational model (Lewis and Vasishth, 2005; Lewis et al., 2006). Within this architecture, it is the direct-access mechanism that provides the computational power to compensate for the severely constrained memory capacity. Indeed, there are now a number of studies, across a broad range of sentence constructions, that provide evidence for direct access in language processing (e.g., McElree, 2000; McElree et al., 2003; Martin and McElree, 2008, 2009, 2011; Van Dyke and McElree, 2011; for reviews see McElree, 2006, 2015).

The paradigmatic evidence for direct-access retrieval in sentence processing was provided by McElree et al. (2003), who asked university students to read sentences containing grammatical dependencies in which the distance between the grammatical head (e.g., *book*) and its dependent (e.g., *ripped*) was manipulated:

(5) The *book ripped*.
(6) The *book* that the editor admired *ripped*.
(7) The *book* that the editor who quit the journal admired *ripped*.

McElree and colleagues found that as the amount of material interpolated between the matrix verb and the sentential subject increased, the probability of accurate retrieval decreased: participants responded very accurately in (5), less accurately in (6), and still less accurately in (7). If "book" were accessed via a serial search mechanism, similar systematic differences should also have been observed in indices of retrieval speed; that is, a serial search mechanism also predicts that participants should be fastest to access *book* in (5), slower to access *book* in (6), and slower still in (7). Instead, McElree and colleagues found that participants resolved the *book-ripped* dependency very quickly in sentences such as (5), and with a slower—but constant—speed in (6) and (7)[2]. Thus, although the memory representations did vary in their availability (perhaps because of decay, or reduced distinctiveness as the number of NPs increased, or both), participants used the cues provided by the verb (e.g., selectional information) to guide direct retrieval of the appropriate NP from memory. Crucially, these results are not compatible with a serial search-based retrieval mechanism, which predicts that items that vary in their availability should not be accessed with equal speed. Hence, this study clearly shows that the collegiate readers were not engaging in a serial, backwards search through information that is no longer active in memory.

In light of this evidence, it seems plausible to suggest that content-addressable, direct-access retrieval is a fundamental property of the human language faculty, and that a cue-based retrieval parser (e.g., Van Dyke and Lewis, 2003; see also Lewis and Vasishth, 2005; Lewis et al., 2006) is the "default" processor for linguistic input. However, there are two potential objections to this proposal. First, all of the studies attesting to this type of retrieval during language processing have employed visually presented stimuli. That is, these studies only provide evidence that a cue-based retrieval parser is active during *reading*; it remains possible that processing *spoken* language initiates qualitatively different memory operations than those observed in reading tasks. The presence of orthographic information could enhance encoding and access during reading in ways that would necessarily be absent during listening comprehension (e.g., Harm and Seidenberg, 2004)—a potential confound that is amplified by extensive evidence that deficient orthographic decoding plays a role in reading difficulty (Shankweiler and Crain, 1986; Bell and Perfetti, 1994; Long et al., 2006). Second, these studies have uniformly tapped university subject pools for their participants, with the result that evidence for the cue-based retrieval parser comes entirely from relatively skilled readers. This raises the possibility that cue-based, direct-access retrieval develops concomitantly with reading skill; that is, more reading or language experience may "tune" the parser to make it more efficient, while less skilled readers may employ less efficient (e.g., search-based) memory operations during language comprehension. Such an account is consistent with some models of WM that suggest that efficient retrieval is predicated on efficient access structures that are derived from acquisition of skill proficiency (e.g., Ericsson and Kintsch, 1995).

[1]There are two known circumstances, both task-specific, in which multiple items may be in focal attention: the task must either promote the "chunking" of information (McElree, 1998), or encourage participants to actively maintain distant information (McElree, 2001, 2006). The connection between these findings and language operations has not been systematically explored. The only attempt of which we are aware is unpublished data suggesting that some phrase types (PP, adverbial) do not displace information in focal attention (Wagers and McElree, 2009; discussed in McElree, 2015). However, as these structures are not examined in our experiments, we adopt the formulation that is most consistent with findings from published language research: that a single word—the most recently processed item—is maintained in focal attention.

[2]McElree et al. (2003) also included a condition containing two embedded object-relative clauses, of comparable length to items such as (7), which also contain two embedded clauses (a subject-relative and an object-relative). The double-object condition yielded both significantly lower response accuracy and a significantly slower rate of retrieval than all other conditions. McElree and colleagues discussed these findings at length, noting that the data pattern remains inconsistent with a search-based explanation; we refer interested readers to McElree et al. (pp. 81–82).

EXPERIMENT 1

Our first experiment examines memory retrieval during auditory language comprehension. This is crucial for assessing whether both auditory and written language processing use direct-access retrieval, as well as for studying retrieval mechanisms in poor readers, whose poor orthography-to-phonology decoding represents an important confound for any study implemented in the visual modality. We created an auditory implementation of the SAT procedure, in which participants listened to, and responded to, a series of sentences that either were or were not grammatically acceptable.

The SAT procedure provides an unambiguous estimate of access speed, which is required to differentiate direct-access retrieval from serial search processes. This contrasts with more commonly used timing measures, such as reaction and reading times, which are not "process pure": in these paradigms, slower RTs may occur as a result of either actual speed differences, differences in the relative likelihood that information will be successfully recovered from memory, or both. In addition, these measures are vulnerable to idiosyncratic response criteria—that is, participants can adopt liberal or conservative response patterns, emphasizing accuracy at the expense of speed, or speed at the expense of accuracy (see McKoon and Ratcliff, 1992; McElree, 1993; Ratcliff and McKoon, 2008). In contrast, the SAT procedure permits the assessment of both by computing response functions that model the entire time course of information accrual (Wickelgren, 1977). The SAT procedure's fine-grained assessment of retrieval dynamics forms the basis for all unambiguous evidence that a fast, content-addressable, direct-access retrieval mechanism with a single-item focal span supports typical online language comprehension processes (e.g., McElree and Griffith, 1995, 1998; McElree, 2000; McElree et al., 2003; Foraker and McElree, 2007; Martin and McElree, 2008, 2009, 2011; Van Dyke and McElree, 2011; for reviews see Foraker and McElree, 2011; McElree, 2015).

Our goal in Experiment 1 was to validate our auditory implementation of the SAT technique by replicating Experiment 2 of McElree et al. (2003) with a comparable population (university students) using auditory versions of the stimuli from that study. Consistent with that study, we predicted that access would be fastest when the critical item was still active in the focus of attention (i.e., the most recently processed word). If the speed of access in the longer conditions, in which retrieval is necessary, is invariant, this supports an account of listening comprehension in which direct-access retrieval is used. However, if retrieval speed in the longer conditions varies systematically according to the distance between the retrieval cue and its target, this would support a search-based retrieval mechanism.

Method

Participants

Informed consent was obtained from five undergraduates at Yale University. The participants were right-handed native English speakers, and were paid for their participation ($20/h). Each participated in one 1-h SAT training session, followed by two 3-h experimental sessions; these sessions were comprised of two 1-h SAT sessions (for a total of four), separated by a 1-h period in which they completed additional cognitive assessments (for a separate study) and rested. Details about the training and experimental sessions are described below.

Materials

Materials were adapted from those used in Experiment 2 of McElree et al. (2003). These constructions permit assessment of the speed and accuracy with which a matrix intransitive verb (e.g., *ripped, laughed*) retrieves its grammatical subject noun (e.g., *book*); examples appear in **Table 1**. Because we planned to test a population with a wide range of comprehension ability in our second experiment, our materials did not include all of the conditions presented in McElree et al. We selected a subset of conditions that linearly increased the surface distance, and the corresponding time, between each sentence's subject NP and matrix verb. For each item, participants were required to determine whether the subject-verb relation was either acceptable or unacceptable (see Procedure and Data Analysis, below). The conditions in both this and the next experiment are:

No Interpolation (T1 and T2): in the shortest conditions, the subject and verb are directly adjacent to each other (no retrieval needed).

Interpolated Object Relative Clause (T3 and T4): distance between subject NP and verb is increased by four words. T3 and T4 are identical to T1 and T2 with the exception of the additional embedded clause.

Interpolated Object and Subject Relative Clauses (T6 and T7): in the longest conditions, the subject and verb are separated by eight words. T6 and T7 are identical to T3 and T4 with the exception of the additional embedded subject relative clause.

Additional processing encouragement (T5 and T8): as in McElree et al. (2003), we included a second type of unacceptable item in each of the longer conditions. These items, exemplified by T5 and T8, are identical to the other items in their corresponding conditions with one exception. In these items, the grammatical inconsistency that determined acceptability was located in the interpolated information. For example, as shown in **Table 1**, although the embedded transitive verb requires a direct object,

TABLE 1 | Constructions used in Experiments 1 and 2 (adapted from McElree et al., 2003).

Construction	Acceptability	Example
No Interpolation	Acceptable	T1. The book ripped.
No Interpolation	Unacceptable	T2. The book laughed.
Object relative	Acceptable	T3. The book that the editor admired ripped.
Object relative	Unacceptable	T4. The book that the editor admired laughed.
Object relative	Unacceptable	T5. The book that the editor amused ripped.
Object relative + subject relative	Acceptable	T6. The book that the editor who quit the journal admired ripped.
Object relative + subject relative	Unacceptable	T7. The book that the editor who quit the journal admired laughed.
Object relative + subject relative	Unacceptable	T8. The book that the editor who quit the journal amused ripped.

book is not an acceptable argument for *amused*. These types of sentences were included to encourage our participants to attend to (and process) the interpolated material.

We selected 48 instances of each of the eight types of sentence (T1–T8, i.e., three acceptable and five unacceptable) from the original materials used in McElree et al. (2003). This yielded a total of 384 experimental items, which we edited slightly in order to make the vocabulary level more appropriate to the participants in our second experiment. From this set of items, we generated four experimental lists of 96 sentences. Each list was comprised of 12 instances of each sentence type. Participants listened to one list during each of the four SAT sessions.

Procedure

All stimuli were randomized within each testing session and presented using the E-Prime experimental package (Schneider et al., 2002). Unlike the original study in which a single-response Speed Accuracy Tradeoff (SR-SAT) paradigm was used, we adopted the multiple-response Speed-Accuracy Tradeoff (MR-SAT) method (Wickelgren et al., 1980; McElree, 1993; see also Bornkessel et al., 2004; Foraker and McElree, 2007; Van Dyke and McElree, 2011). Because more responses are collected per trial, MR-SAT paradigms require fewer items, and consequently fewer experimental sessions to complete an experiment. Each trial began with the words "Listen carefully," which appeared in the center of the screen throughout the trial. The initial appearance of these words was accompanied by an auditory fixation cue (a tone). This cue alerted participants to the imminent auditory presentation of a sentence, which began 500 ms after the offset of the cue. All sentences were prepared using version 2.0.3 of Audacity® recording and editing software (Audacity Team, 2015; http://audacity.sourceforge.net). The sentences were presented at a natural speaking rate (in contrast to previous visual SAT studies, in which sentences were segmented word-by-word or phrase-by-phrase). A sequence of 15 tones (100 ms, 1000 Hz, every 350 ms) was spliced into the sentence recording, beginning 200 ms prior to the onset of the sentence-final critical word. The tones were presented simultaneously with and following the critical word, forming a 5000 ms response period. Participants were instructed to judge whether each sentence was an acceptable English sentence. They were trained to press the response key(s) corresponding to their acceptability judgment in time with the tone sequence. At the onset of the tones, participants began responding by pressing both response buttons, indicating that they did not yet know whether or not the sentence was acceptable. After hearing the sentence-final word, participants indicated whether the sentence was acceptable or unacceptable by choosing either the YES or NO response key, and continuing to press only that button in time with the tones.

During the training session, participants first heard and responded to response tones in isolation, in order to become familiar with the auditory and motor aspects of the SAT procedure; they subsequently heard and responded to practice items similar to those in the experiment. In addition to the initial training, participants also completed a 15-min refresher session at the beginning of the second experimental session in order to refamiliarize themselves with the task. Participants received feedback about their responses in both training sessions, indicating whether their responses were faster or slower than, or out of sync with, the rhythm of the response tones. In addition, they were taught that they could change their response; for example, if at first they decided that a sentence was acceptable (and consequently stopped pressing the NO response key while continuing to press the YES response key), but subsequently changed their mind and deemed it unacceptable, they could switch their response (i.e., stop pressing YES, and resume pressing NO). Participants were taught that they could change their response at any time—and multiple times, if necessary—during the 5000 ms response period.

Data Analysis

SAT data provide indices of both accuracy and speed associated with responses. In studies using the SAT method, a stable SAT function can be calculated for each participant and, as a consequence, each participant is analyzed separately. This approach has two advantages: it reduces the variance associated with each participant's data, and it minimizes distortion associated with averaging across participants. Consistent patterns that emerge across participants are subsequently considered through analyzing both modeling consistency across individuals and modeling of the averaged data.

Accuracy was computed for each time point in the response period using a standard measure of sensitivity (d'). Potential response bias was controlled by calculating d' using z-scores for hits and false alarms [$d' = z(\text{hits}) - z(\text{false alarms})$]. In this experiment, a "hit" is a YES response to an acceptable sentence, and a "false alarm" is a YES response to an unacceptable sentence.

The asymptote, rate, and intercept for each response function were assessed by fitting the d' accuracy scores at each response point (t), with an exponential approach to a limit:

$$d'(t) = \lambda(1 - e^{-\beta(t-\delta)}) \text{ for } t > \delta, \text{ else } 0$$

Thus, d' is the result of the interaction of the two factors that define an SAT function: the asymptote of the function (λ), and the speed with which that asymptote is reached. Speed is jointly determined by two distinct parameters: the intercept of the function (δ), which is the point at which response accuracy rises above chance, and the rate at which response accuracy reaches asymptotic performance (β). Calculated d' scores are then fit to hierarchically nested models ranging from a null model, in which the experimental conditions are fit using a single asymptote, rate, and intercept, to a fully saturated model, in which the conditions are each fit with a unique set of parameters. For data modeling, we used functions from the package *mrsat* (Matsuki et al., in preparation)[3]. The fitting function applied four different optimization algorithms that are implemented in R functions: (1) an iterative hill-climbing algorithm (Reed, 1976) similar to STEPIT (Chandler, 1969), which has been used in the majority of previous SAT studies of language processing and is implemented

[3] Available from GitHub (https://github.com/). Contact matsukk@mcmaster.ca for details.

in the *acp* function; (2) a limited-memory Broyden–Fletcher–Goldfarb–Shanno algorithm with box constraints (Byrd et al., 1995) implemented as a part of the *optim* function; (3) a box-constrained optimization algorithm based on PORT routines developed by Bell Labs (Fox et al., 1978) as implemented in the *nlminb* function; (4) an unconstrained optimization algorithm based on a Newton-type method implemented in the *nlm* function (Dennis and Schnabel, 1983; Schnabel et al., 1985). Each of these algorithms were applied 10 times with randomly chosen starting parameter values on each run, and the resulting set of parameters that provided the best model fits were selected. Fit quality was assessed in two ways. First, we calculated a modified R^2 statistic, in which the number of parameters present in each model is used to adjust the proportion of variance accounted for by each model (Judd and McClelland, 1989). Second, we evaluated the consistency of the parameter estimates across participants.

All SAT response function and statistical analyses were carried out with the R statistical software, version 3.2.1 (R Core Team, 2015). For analyses, we used the package *lme4* (Bates et al., 2015). We used linear mixed-effect regression (*LMER*; Baayen, 2004, 2008; Baayen et al., 2008) to assess the observed empirical data and the fitted parameter estimates for each of the candidate models described in the Results Section. Mixed-effects models included fixed effects of Construction and random intercepts for participants. For evaluation of the main effect of Construction, we report the associated *F*-value, as well as the denominator degrees of freedom and *p*-values that were calculated based on Satterthwaite's approximation using the *lmerTest* package (Kuznetsova et al., 2015). We also report the *t*-values associated with our analyses, adopting the convention whereby any effect whose absolute *t*-value exceeds 2 is considered significant (Gelman and Hill, 2007).

Results

Figure 1 shows the averaged d' data at each response point, as well as smoothed curves depicting the best fitting model (3λ-1β-2δ; see below) as a function of processing time for the three Construction conditions (No Interpolated Material, Interpolated Object Relative, Interpolated Object + Subject Relative). As in McElree et al. (2003), visual inspection of the data suggests that asymptotic accuracy is negatively correlated with the amount of material interpolated between the matrix verb and its subject. This observation is supported by the LMER analysis of the mean of the last four d' values, which is the empirical estimate of asymptotic accuracy. This confirmed a main effect of Construction, $F_{(2,\ 8)} = 40.17$, $p < 0.001$. Pairwise comparisons showed that accuracy was higher when there was no material between subject and verb ($d' = 3.58$) than when there was an intervening object relative clause ($d' = 2.55$), $t = -4.33$, or when there were intervening subject and object relative clauses ($d' = 1.45$), $t = -8.96$. In addition, the asymptotic accuracy of the Interpolated Object Relative condition was significantly higher than that of the Interpolated Object + Subject Relative condition, $t = -4.63$. This pattern of results replicates the pattern reported in McElree et al. (2003) for these conditions.

Initial hierarchical modeling of the data assessed three models, differing only by the number of asymptote parameters assigned to the models. First, we assessed the null model, in which a common asymptote (λ), rate (β), and intercept (δ) was assigned to each condition. The 1λ-1β-1δ model fit produced an adjusted R^2 for the averaged data of 0.585, ranging from 0.292 to 0.782

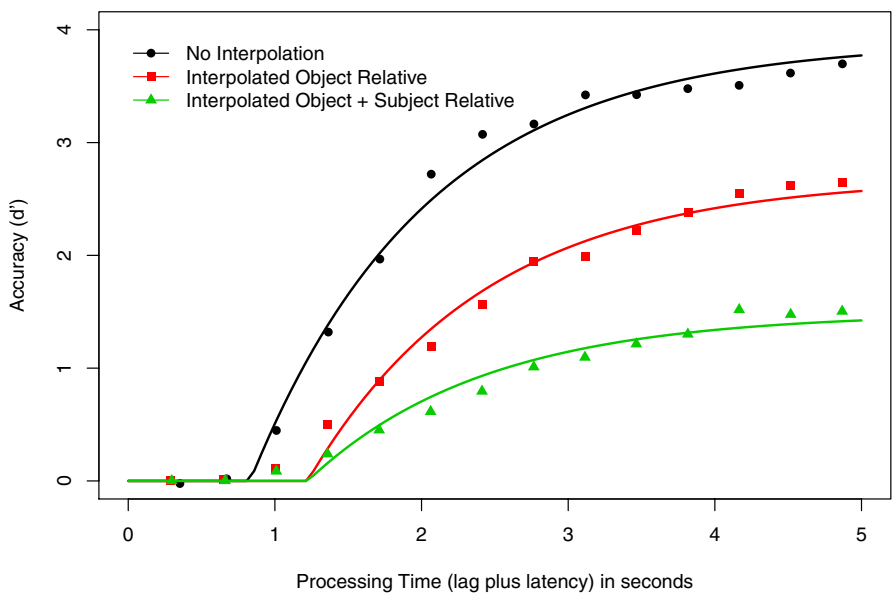


FIGURE 1 | Speed-accuracy tradeoff results for Experiment 1. Average d' accuracy as a function of processing time (in seconds) for the grammaticality judgments of sentences with the following constructions: no embedded material (circles), one object-relative clause (squares), and one object- and one subject-relative clause (triangles). Smooth curves show the 3λ-1β-2δ exponential model (see text of Experiment 1).

across all participants. We next fit a 2λ-1β-1δ model to the data, in which one asymptote was assigned to the No Interpolation condition, and a second was assigned to the conditions with material intervening between the subject and the verb. This model fitting produced an adjusted R^2 for the averaged data of 0.903, ranging from 0.823 to 0.945 across all participants. All participants showed an increase in adjusted R^2 compared with the null model (average adjusted R^2 increase = 0.337; minimum = 0.164; maximum = 0.641). The third fitting assigned a unique asymptote parameter to each Construction condition, a 3λ-1β-1δ model; this produced an adjusted R^2 for the averaged data of 0.980, ranging from 0.955 to 0.993. The addition of an asymptote parameter again showed an increase in adjusted R^2: compared to the 2λ-1β-1δ model, the average adjusted R^2 increase was 0.074 (minimum = 0.029; maximum = 0.153); further, the average adjusted R^2 increase was 0.411 (minimum = 0.192; maximum = 0.700) when this model was compared to the 1λ-1β-1δ model. The λ estimates (in d' units) for the averaged 3λ-1β-1δ model were 4.06 for the No Interpolation condition, 2.58 for the Interpolated Object Relative condition, and 1.43 for the Interpolated Object + Subject Relative condition. An LMER analysis of the λ estimates showed a significant effect of Construction, $F_{(2, 8)} = 104.74$, $p < 0.001$. Pairwise comparisons closely tracked the pattern of the analysis of the empirical d' data above. Specifically, the λ estimates for the No Interpolation condition were higher than both the Interpolated Object Relative ($t = -8.16$) and Interpolated Object + Subject Relative conditions ($t = -14.43$), and the Interpolated Object Relative condition λ estimate was greater than the Interpolated Object + Subject Relative estimate ($t = -6.27$). This finding—that a model with three asymptote parameters better fits the data than do models with two or one asymptote—is consistent with our analysis of the empirical d' data. Thus, subsequent analyses focus on models with three asymptotes.

We next evaluated the effect of Construction on processing speed. Unlike the original study (McElree et al., 2003), the data do not suggest that these analyses should exclude speed differences in either the intercept or the rate; hence, we tested for differences manifesting in the intercept (δ); rate (β); and in both parameters together. We began by fitting a 3λ-1β-2δ model to the data, with potential speed differences assigned to the intercept. As in the asymptote comparisons, one parameter was assigned to the No Interpolation condition, and the second was assigned to the conditions with intervening material. This model produced an adjusted R^2 for the average data of 0.992, ranging from 0.983 to 0.994 for individuals. All participants showed an increase in the adjusted R^2 for the 3λ-1β-2δ over the 3λ-1β-1δ model (average increase = 0.014; minimum = 0.001, maximum = 0.028). We subsequently fit a 3λ-1β-3δ model to the data, but the addition of a third intercept parameter was not warranted: no participants showed an improved adjusted R^2 for this model relative to the 3λ-1β-2δ model (adjusted $R^2 = 0.992$; average increase = −0.002, minimum = −0.003, maximum = −0.001). Parameter estimates for the 3λ-1β-2δ are shown in **Table 2**.

We next evaluated potential speed differences that could result from the rate parameter, and fit a 3λ-2β-1δ model to the data. This model produced an adjusted R^2 of 0.996 for the average data, ranging from 0.990 to 0.993 for individuals; however, data from two participants could not be fit to this model without overestimating the asymptote parameters. For those participants that were successfully fit with this model, the addition of a second rate parameter yielded an improved model fit over the 3λ-1β-1δ model (average adjusted R^2 increase = 0.018; minimum = 0.001; maximum = 0.038). In addition, the 3λ-2β-1δ model fit the data better than the 3λ-1β-2δ model for two of these participants (average adjusted R^2 increase = 0.008; minimum = 0.005; maximum = 0.010). A subsequent fit using a 3λ-3β-1δ model, excluding both participants without 3λ-2β-1δ model parameter estimates, showed that a third rate parameter was not warranted by the data (adjusted $R^2 = 0.996$; average adjusted R^2 increase = 0).

Finally, we considered a 3λ-2β-2δ model, in which speed differences could arise from both rate and intercept; one participant's data could not be modeled, again due to overestimated asymptotes, and was not included. Although this model (adjusted $R^2 = 0.997$) did result in improved adjusted R^2 for the average data relative to the model with two intercept parameters (3λ-1β-2δ; average adjusted R^2 increase = 0.003; minimum = −0.001; maximum = 0.010), there was no adjusted R^2 difference between this model and the 3λ-2β-1δ model (average adjusted R^2 increase = 0; minimum = 0; maximum = 0.002).

LMER analyses of the parameter estimates for the models with two (3λ-2β-1δ, 3λ-2β-2δ; all ts < 1.7) and three [3λ-3β-1δ; $F_{(2, 6)} = 2.41$, $p = 0.171$] rate parameters were non-significant, possibly due to the small number of participants. However, the

TABLE 2 | Experiment 1: adjusted R^2, d's, and parameter estimates for the average data and individual participants for the 3λ-1β-2δ exponential model.

	Adj.	*d*-primes			Asymptotes			Rate	Intercept	
	R^2	d' 1	d' 2	d' 3	λ1	λ2	λ3	β1	δ1	δ2
Avg	**0.992**	**3.58**	**2.55**	**1.45**	**3.897**	**2.690**	**1.489**	**0.826**	**0.831**	**1.223**
S1	0.983	3.83	2.55	1.18	3.975	2.416	1.038	1.358	0.733	1.306
S2	0.994	2.80	1.08	0.46	2.866	1.126	0.425	1.316	1.158	1.026
S3	0.989	3.28	2.58	1.83	4.129	3.271	2.359	0.519	0.948	1.433
S4	0.993	3.62	2.63	1.89	3.988	2.994	2.093	0.775	1.113	1.580
S5	0.985	4.35	3.90	1.89	4.776	3.943	1.947	0.891	0.728	0.973

TABLE 3 | Experiment 1: adjusted R^2, d's, and parameter estimates for the average data and individual participants for the 3λ-2β-1δ exponential model.

	Adj.	*d*-primes			Asymptotes			Rate		Intercept
	R^2	d' 1	d' 2	d' 3	λ1	λ2	λ3	β1	β2	δ1
Avg	**0.996**	**3.58**	**2.55**	**1.45**	**3.666**	**3.256**	**1.685**	**1.156**	**0.439**	**0.886**
S1	0.993	3.83	2.55	1.18	3.820	3.276	1.417	2.328	0.442	0.907
S2	0.993	2.80	1.08	0.46	2.879	1.097	0.412	1.270	1.752	1.144
S5	0.990	4.35	3.90	1.89	4.490	4.427	2.185	1.306	0.603	0.874

LMER analyses of the model with three intercept parameters (3λ-1β-3δ) revealed a significant main effect of Construction, $F_{(2,\ 8)} = 6.31$, $p = 0.023$). *T*-tests indicated that the No Interpolation condition was significantly different than both the Interpolated Object Relative condition ($t = 2.84$) and the Interpolated Object + Subject Relative condition ($t = 3.27$), but the two long conditions were not statistically different ($t < 1$). In addition, the LMER test of the 3λ-1β-2δ model confirmed that the two intercept parameters differed significantly ($t = 2.576$). These analyses are all consistent with the modeling conclusions that adding a third rate or intercept parameter is not warranted. Overall, these analyses indicate that the best model for both individual and average data is the 3λ-1β-2δ model: all participants were fit by this model, and alternative models did not yield consistent improvement. However, because our conclusions do not depend on whether the second speed parameter manifests on either the rate or the intercept, we present the parameter values for both models (see **Tables 2, 3**). The key conclusion is that there is no evidence to support the inclusion of a third speed parameter (either rate or intercept) for any participant or for the average data.

Discussion

Consistent with previous research (e.g., McElree, 2000; McElree et al., 2003), we observed a negative correlation between response accuracy and the amount of material interpolated between the sentences' matrix verbs and the subject nouns. The significant differences in the empirical d' data and in the model asymptotes confirm that as the distance between the subject and verb increases, the probability of accurately resolving the long-distance dependency decreases. Such asymptotic decreases are attributable to either an overall decrease in the quality of the memory representation over time, or to a decrease in the diagnostic distinctiveness of the retrieval cue (i.e., the featural characteristics of the verb) relative to the to-be-retrieved information (see Van Dyke and Johns, 2012). In addition, SAT response functions were best fit by a model in which there were two speed parameters, one reflecting fast access when no retrieval was necessary (i.e., the condition in which no material intervened between a verb and its grammatical head noun, leaving the most recently processed item active), and a second reflecting slower access when the critical item was not in focal attention and required retrieval. Critically, there was no benefit to including a third speed parameter (either on the rate or intercept), which would have supported a search-based retrieval mechanism: verbs retrieved their subjects with the same speed regardless of interpolated material. This pattern of asymptotic and dynamic differences is the characteristic signature of direct-access retrieval, and is apparent in the individual participants' data (see **Table 2**)[4].

In addition, our participants' performance on conditions with grammatical anomalies in an embedded clause (conditions T5 and T8) suggests that they were not simply focusing on the initial noun and final verb in order to make their grammaticality judgments. Averaged correct rejection rates for these conditions for each of the response lags were 49.8, 50.4, 51.3, 54.9, 61.7, 64.3, 70.8, 75.0, 76.3, 78.4, 80.3, 82.3, 83.2, and 84.3%. Correct rejection rates for the corresponding experimental conditions (T4 and T7), in which the ungrammaticality derived from the sentence-final verb, were 49.6, 49.9, 50.7, 54.7, 63.1, 69.2, 75.6, 80.3, 81.9, 83.6, 84.1, 85.6, 85.2, 84.7%. As in McElree and colleagues' original study (McElree et al., 2003), accuracy was higher in the experimental conditions than in the conditions designed to discourage strategic processing. However, unlike the original study, correct rejection rates were not asymptotic at early lags in conditions T5 and T8; rather, the pattern of correct rejections seems to reflect an exponential response function. This difference could arise from any of the ways our study differs from the original, including our use of the multiple-response variant of the SAT technique, our use of auditory presentation of the sentences, or some combination of the two. For example: perhaps the relatively faster presentation of the sentences in an auditory (relative to the previously used visual) modality prevented early decision making. Alternatively, perhaps the need to process (at each response tone) an acceptable verb in light of an earlier anomaly, reduced participants' confidence in rejecting the sentence and/or prolonged repair routines aimed at finding a correct interpretation. However, such explanations are speculative, and ultimately are unrelated to the main issues addressed here. The value of these correct rejection rates is their clear demonstration that our participants processed the interpolated material, rather than simply ignoring it.

[4]One participant, S2, shows speed dynamics contrary to the predicted direction. This is also true of one participant in Experiment 2 (S9). We believe this to be an artifact of the fitting function: modeling the dynamic portion of the SAT function can be difficult if d' scores are very low (McElree, personal communication). However, because these participants' asymptotic accuracy is consistent with both the other participants and the pattern observed in the original study (McElree et al., 2003), and out of consideration for the sample sizes in both experiments, we ultimately elected to include these data in our analyses. If these participants are excluded, our results do not change—and indeed become stronger in the predicted direction.

The results of this experiment replicate McElree and colleagues' demonstration of direct-access retrieval (McElree et al., 2003). These results are significant for three reasons. First, our MR-SAT replication of the original SR-SAT study continues a tradition of validating important findings about the operation of the human memory system across SAT techniques (e.g., McElree and Dosher's SR-SAT replication of Wickelgren and colleagues' MR-SAT findings regarding the focus of attention; Wickelgren et al., 1980; McElree and Dosher, 1989). Second, these results constitute the first evidence that, as in reading comprehension, collegiate comprehenders employ a content-addressable, direct-access retrieval mechanism during listening comprehension. Finally, unlike previous research, this interpretation is not susceptible to any confound related to orthographic processing. Thus, these results suggest that direct-access retrieval is a modality independent cognitive operation. Additionally, they validate the auditory MR-SAT procedure as an appropriate tool for investigating the retrieval mechanism in individual participants regardless of reading skill.

EXPERIMENT 2

The results of our first experiment, in tandem with previous studies, suggest that direct-access retrieval may be the "default" setting during language comprehension, as it has now been observed both during reading and listening comprehension. Experiment 2 assessed the potential for direct-access retrieval in poor readers. Motivation for this work comes from studies indicating that capacity-based explanations are unlikely to account for poor reading comprehension (e.g., Traxler et al., 2012; Van Dyke et al., 2014; for review see Van Dyke and Johns, 2012). Rather, they point to limited capacity parsing architectures that rely on a fast, direct-access retrieval mechanism to restore information into the focus of attention as needed (e.g., Lewis and Vasishth, 2005; Lewis et al., 2006). However, studies establishing the presence of direct-access retrieval during language comprehension have been conducted exclusively with university students, presumably possessing a relatively high degree of comprehension skill. As such, this evidence only suggests that memory capacity is not important for argument integration in adult *skilled readers*. This leaves open the question of whether less-skilled readers are able to employ the same direct-access retrieval mechanism as skilled readers; that is, poor comprehension in these readers may arise because they simply do not have access to a direct-access retrieval mechanism, and must instead rely upon a slower, less efficient mode of retrieval (i.e., search) during comprehension. Numerous findings showing that less-skilled readers are typically slower than skilled readers to retrieve phonologically encoded information during comprehension support this possibility (Perfetti, 1985; Swan and Goswami, 1997a,b; Wolf and Bowers, 1999; Goswami, 2011).

Thus, our goal in Experiment 2 was to use the auditory SAT technique to determine whether less-skilled readers have access to an efficient, direct-access retrieval mechanism *at all*. The question of whether less-skilled readers are able to use direct-access retrieval is particularly important given that the prevailing account of memory limitations during reading comprehension suggests that poor readers' comprehension is inherently compromised—that reading skill is essentially predetermined by fundamental, fixed differences in the memory system. The most obvious example of this approach is the notion of intrinsic, fixed WM capacities, which are thought to determine the facility with which a given comprehender may process linguistic information (e.g., Just and Carpenter, 1992; Caplan and Waters, 1999). According to this account, those with low WM capacities are predestined to be poor comprehenders, while those with higher WM capacities are not.

An alternative possibility is suggested by Ericsson and Kintsch (1995) in their Long-Term WM model. According to this model, skilled performance on any task (e.g., mental calculations, medical diagnosis, playing chess) is predicated on the development of highly efficient, skill-specific access structures, in which retrieval cues in active memory facilitate access to information in LTM. In each case, skilled practitioners enjoy rapid access to critical information, while those less-skilled will retrieve information more slowly and with difficulty. In the context of skilled reading comprehension, the development of proficient decoding, by which readers use orthographic representations to access lexical information, may provide the critical link between active and LTM. That is, because skilled readers have highly efficient mappings between the orthographic, phonological, and semantic characteristics of a word, they may enjoy direct-access retrieval of the lexical information upon which higher-level language processes (syntactic parsing, semantic, and discourse integration) depend. Less-skilled readers, in contrast, may instead be forced to rely on less efficient, search-based retrieval.

Critically, both of these accounts suggest that poor readers simply do not have access to an efficient retrieval mechanism to support reading—either because they do not (and cannot) have one, or because they do not have sufficient expertise to develop one. Thus, the importance of this experiment derives from its assessment of less-skilled readers' memory operations *when they are not reading*. If poor readers show the ability to employ content-addressable direct-access during auditory language processing, then they are not inherently saddled with a less efficient default retrieval mechanism. Furthermore, if less-skilled readers demonstrate the ability to use a direct-access retrieval mechanism, then it also cannot be the case that efficient retrieval is a byproduct of the development of reading expertise.

We used the same materials as in our first experiment. In addition, the participants in this study were not university students; we recruited a community-based sample of non-college bound young persons. Our previous experience with this population led us to expect large skill differences on a range of cognitive measures (e.g., Braze et al., 2007, in press; Shankweiler et al., 2008; Kuperman and Van Dyke, 2011; Magnuson et al., 2011; Johns et al., 2014; Van Dyke et al., 2014; Kukona et al., submitted). Our sample was age-matched to the standard college subject-pool population, which permits comparisons with previous studies of memory operations during language processing. As in those studies, we expected our participants' accuracy to vary according to the length of our experimental sentences (see Materials, Experiment 1), with the

lowest accuracy in the longest conditions. As in Experiment 1, the critical comparisons for assessing the retrieval mechanism derive from the processing speed dynamics (rate and intercept) of their response functions. If poor readers use a search-based mechanism, then retrieval speed should vary as a function of the length of the experimental sentences (i.e., as a function of the amount of material interpolated between the matrix verb and its head noun). However, if poor readers are able to use a direct-access retrieval mechanism, speed should be fast when no retrieval is required (i.e., when there is no intervening material) and invariant across all other conditions, which do require retrieval.

Method

Participants

Informed consent was obtained from 22 young people (ages 16–24) recruited from the local New Haven community. We recruited participants in a number of ways, including presentations at adult education centers, advertisements in local newspapers, flyers placed on adult school campuses, community centers, public transportation hubs, local retail and laundry facilities, and referrals from current and past study participants. All participants were right-handed native English speakers without a diagnosed reading or learning disability, and were paid for their participation ($20/h). Each participated in two 3-h experimental sessions identical to those described in Experiment 1, including initial training and an intersession period in which they completed additional cognitive assessments (for another study) and rested.

We assessed Reading Ability via the Peabody Picture Vocabulary Test (PPVT, 3E; Dunn and Dunn, 1997), which is a measure of receptive (i.e., interpretive, rather than productive) vocabulary. Vocabulary is known to be a limiting factor in the development of reading comprehension (Joshi, 2005; Perin, 2013). It frequently emerges as a unique predictor of reading ability, accounting for variance beyond that captured by other measures such as decoding, or by indices of reading comprehension (e.g., Braze et al., 2007, in press; Fraser and Conti-Ramsden, 2008; Ouellette and Beers, 2010; Tunmer and Chapman, 2012). There are now many psycholinguistic studies in which vocabulary was the critical measure for investigating individual differences in linguistic performance (e.g., Traxler and Tooley, 2007; Prat and Just, 2011; see also Long et al., 2008; Hamilton et al., 2013), including work from our lab using the PPVT (Braze et al., 2007, in press; Van Dyke et al., 2014). The distribution of scaled PPVT scores is shown in **Figure 2**; descriptive statistics and age equivalents are shown in **Table 4**. (Our participants completed the vocabulary assessment together with other skill assessments as part of a different study. We present a summary of these assessments in **Table 4** so as to further

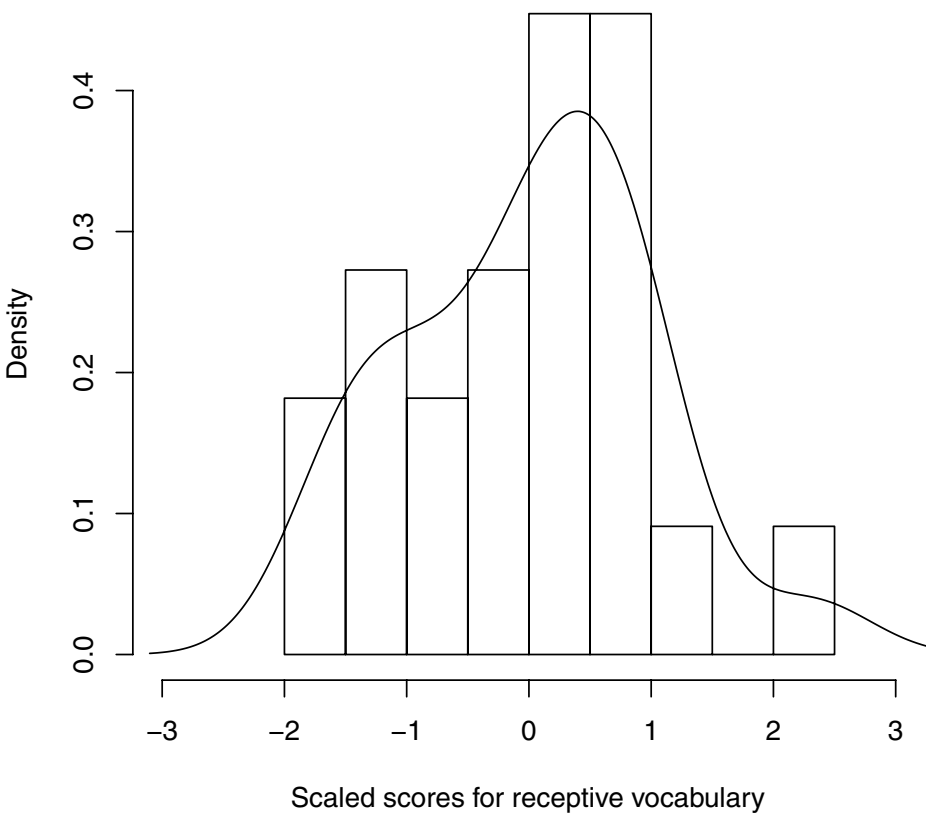


FIGURE 2 | Density histogram and curve of scaled scores for receptive vocabulary (Peabody Picture Vocabulary Test, 3E; Dunn and Dunn, 1997).

TABLE 4 | Range, means, and standard deviations for selected cognitive battery measures.

Measure		Range	*M*	*SD*	Max. possible
1	Receptive vocabulary	74–128	96.41	13.60	204
	Age equivalent score	10–22	16.67	4.66	22
2	Word reading (word attack)	18–32	25.73	3.45	32
	Grade equivalent	3.5–19	8.74	4.33	19
3	Word identification	54–74	65.18	5.70	76
	Grade equivalent	5.1–18	11.38	4.68	19
4	Reading fluency	56–94	75.45	11.86	98
	Grade equivalent	7.7–19	13.19	3.86	19
5	Reading comprehension	30–42	35.77	3.35	47
	Grade equivalent	4.3–19	10.52	4.73	19
6	Oral comprehension	19–30	25.59	2.77	34
	Grade equivalent	4.4–19	11.71	3.51	19
7	Gates-MacGinitie	27–46	36.82	5.58	48
	Grade equivalent	>PHS*	PHS		
8	Working memory capacity	24–57	42.55	8.38	60
9	IQ	63–123	94.18	13.87	–

*Post High School.

Measure 1, Peabody Picture Vocabulary Test-Revised (Dunn and Dunn, 1997); 2–6, Woodcock-Johnson-III Tests of Achievement (Woodcock et al., 2001); 7, Gates-MacGinitie Reading Tests (MacGinitie et al., 2000); 8, listening span (Daneman and Carpenter, 1980); 9, Weschler Abbreviated Scales of Intelligence (Psychological Corporation, 1999).

characterize the cognitive abilities of this sample; however only the vocabulary assessment is used in the current analyses.)

Materials, Procedure, Data Analysis

The materials, procedure, and parameters of the data analysis were identical to Experiment 1, except that the analyses included fixed effects of Reading Ability and the interaction of Reading Ability × Construction.

Results

Figure 3 shows the averaged d' data (data points) and the best fitting 3λ-1β-2δ model (smoothed curves) as a function of processing time for the experimental conditions (No Interpolated Material, Interpolated Object Relative, Interpolated Object + Subject Relative). The LMER analysis of the mean of the last four d' values yielded significant main effects of Construction, $F_{(2, 40)} = 161.00$, $p < 0.001$, and Reading Ability, $F_{(1, 20)} = 56.11$, $p < 0.001$. This effect is depicted in **Figure 4**. However, the interaction of Construction × Reading Ability was not significant, $F_{(2, 40)} = 1.563$, $p = 0.222$. Pairwise comparisons to resolve the main effect of Construction showed that accuracy was higher when there was no material between subject and verb ($d' = 2.41$) than when there was an intervening object relative clause ($d' = 1.32$), $t = -11.57$, or when there were intervening subject and object relative clauses ($d' = 0.73$), $t = -17.66$. In addition, the asymptotic accuracy of the Interpolated Object Relative condition was significantly higher than that of the Interpolated Object + Subject Relative condition, $t = -6.09$. This pattern replicates the empirical d' findings from both our first experiment and McElree et al. (2003).

Hierarchical modeling of the data proceeded as in the previous experiment, first comparing the 1λ-1β-1δ (null), 2λ-1β-1δ, and 3λ-1β-1δ models. The 1λ-1β-1δ model fit produced an adjusted R^2 for the averaged data of 0.540, ranging from 0.299 to 0.895 across all participants. The 2λ-1β-1δ model (in which the additional asymptote parameter was again assigned to the conditions with interpolated material) produced an adjusted R^2 for the averaged data of 0.947, ranging from 0.863 to 0.984 across all participants. All participants showed an increase in adjusted R^2 compared with the null model (average adjusted R^2 increase = 0.409; minimum = 0.03; maximum = 0.665). Finally, the 3λ-1β-1δ model produced an adjusted R^2 for the averaged data of 0.990, ranging from 0.960 to 0.991 for individuals. Compared to the 1λ-1β-1δ model, the average adjusted R^2 increase was 0.455 (minimum = 0.08; maximum = 0.692); compared to the 2λ-1β-1δ model, the average adjusted R^2 increase was 0.046 (minimum = 0; maximum = 0.11). The λ estimates (in d' units) for the averaged 3λ-1β-1δ model were 2.80 for the No Interpolation condition, 1.43 for the Interpolated Object Relative condition, and 0.80 for the Interpolated Object + Subject Relative condition. The LMER analysis of the λ estimates revealed significant main effects of Construction, $F_{(2, 40)} = 196.66$, $p < 0.001$, and Reading Ability, $F_{(1, 40)} = 50.53$, $p < 0.001$, but the interaction was again non-significant, $F_{(2, 40)} = 2.24$, $p = 0.12$. Pairwise comparisons to resolve the significant Construction effect closely tracked the pattern of the analysis of the empirical d' data above. Specifically, the λ estimates for the No Interpolation condition were higher than both the Interpolated Object Relative ($t = -13.34$) and Interpolated Object + Subject Relative conditions ($t = -19.38$), and the Interpolated Object Relative condition λ estimate was greater than the Interpolated Object + Subject Relative condition ($t = -6.04$). This finding—that a model with three asymptote parameters better fits the data than do models with two or one asymptote, and that Reading Ability does not interact with this pattern—is consistent with our analysis of the empirical d' data. Thus, our subsequent analyses again focused on models with three asymptotes.

We next evaluated the potential effects of Construction and Reading Ability on processing speed. It was first necessary to determine the best-fitting model for the average and individual data, so that each participant's rate (β) and intercept (δ) parameters could be examined in light of their scores on our vocabulary assessment. As in our first experiment, the data do not suggest that either the intercept or rate parameters can be excluded from analysis (see **Figure 3**). We first assigned an additional parameter to the intercept, so that the 3λ-1β-2δ model assigned one δ for the No Interpolation condition, and another for the conditions with intervening material. The adjusted R^2 for this model's averaged data was 0.995, ranging from 0.969 to 0.994 for individuals. All participants but four showed an increase in the adjusted R^2 for the 3λ-1β-2δ over the 3λ-1β-1δ model (average increase = 0.006; minimum = −0.001, maximum = 0.02). A subsequent fitting of a 3λ-1β-3δ model to the data showed that the addition of a third intercept parameter was

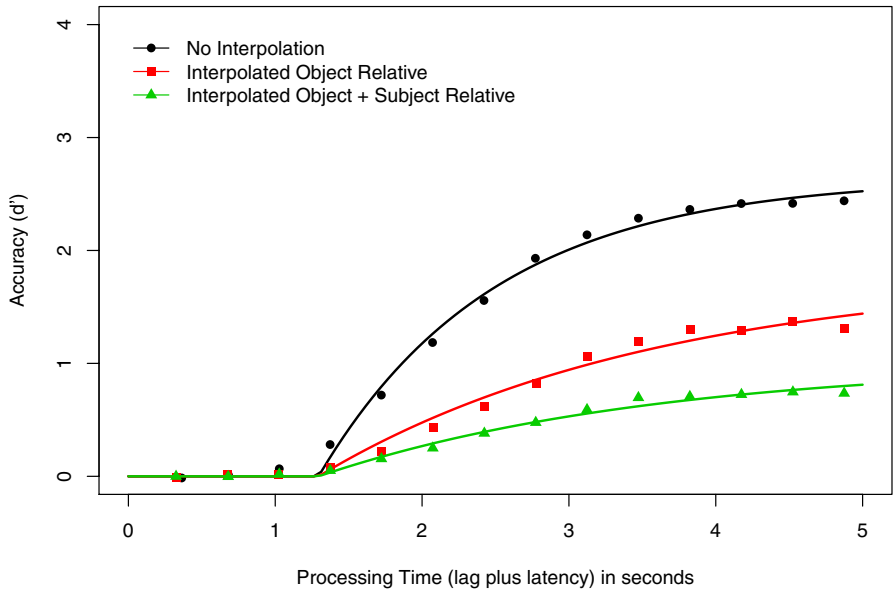


FIGURE 3 | Speed-accuracy tradeoff results for Experiment 2. Average *d′* accuracy as a function of processing time (in seconds) for the grammaticality judgments of sentences with the following constructions: no embedded material (circles), one object-relative clause (squares), and one object- and one subject-relative clause (triangles). Smooth curves show the best fitting 3λ-1β-2δ exponential model.

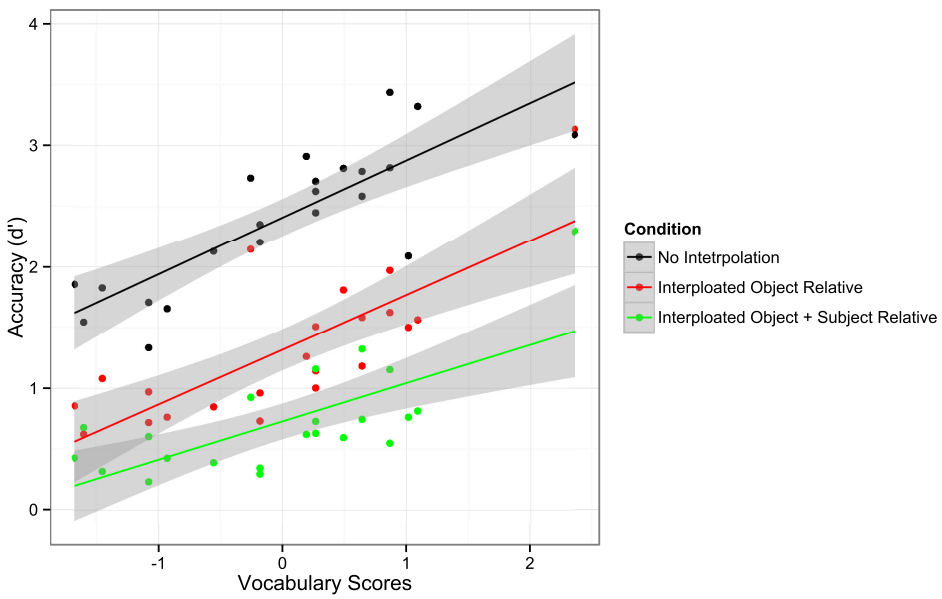


FIGURE 4 | The main effect of Reading Ability on *d′* accuracy in Experiment 2.

not warranted: although eight participants showed an improved adjusted R^2 for this model relative to the 3λ-1β-2δ model, on average the adjusted R^2s were identical (adjusted R^2=0.995; average increase = 0.001; minimum = −0.001, maximum = 0.007).

Next, we evaluated the rate parameter, adding a β so that one parameter was assigned to the No Interpolation condition, and the other to the conditions with interpolated material. This 3λ-2β-1δ model (average adjusted R^2 = 0.995, individual adjusted R^2 = 967 to 0.994) improved model fit over the 3λ-1β-1δ model: all but three participants showed an increase in adjusted R^2 (average adjusted R^2 increase = 0.006; minimum = 0; maximum = 0.034). However, this model was only a minimal improvement over the 3λ-1β-2δ model: although eight participants showed an increased adjusted R^2 (average increase = 0.001; minimum = −0.004; maximum = 0.017), the remaining 14 showed either no improvement or a decrement in fit (from −0.001 to −0.004). Moreover, the adjusted R^2 for the 3λ-2β-1δ

model's average data was identical to the 3λ-1β-2δ model. A subsequent 3λ-3β-1δ model fitting indicated that a third rate parameter was not warranted by the data (adjusted $R^2 = 0.994$; average adjusted R^2 increase = 0.001, minimum = −0.001; maximum = 0.005). In light of this, the absence of a clear difference between the 3λ-2β-1δ and 3λ-1β-2δ models suggests that differences in retrieval speed may derive from the addition of either a second δ or β parameter, determined individually for each participant.

Finally, we considered a 3λ-2β-2δ model, in which speed differences could arise from both rate and intercept. This model (adjusted $R^2 = 0.995$) was a slight improvement for nine participants (and a decrement for one participant) relative to the 3λ-1β-2δ model (average adjusted R^2 increase = 0.001; minimum = −0.001; maximum = 0.016); it was also a slight improvement over the 3λ-2β-1δ model for eight participants (average adjusted R^2 increase = 0.001; minimum = 0; maximum = 0.004). Of those participants showing an increased adjusted R^2 with a 3λ-2β-2δ model, only two showed an increase relative to both of the models with six parameters.

Overall, this pattern of model fits makes two critical points. First, models with three parameters for either the rate or intercept are not appropriate for this data. Second, although it is clear that a model with two speed parameters is appropriate for this data, the additional parameter may manifest on the rate, the intercept, or potentially both indices of retrieval dynamics.

We conducted a series of LMER analyses of the β and δ estimates for the five models considered above. In addition, in order to determine whether our participants' retrieval dynamics varied according to reading skill, Reading Ability was included as a factor (and interaction term) in all comparisons where appropriate. However, across all models, there were no main effects or interactions associated with Reading Ability (3λ-3β-1δ, 3λ-1β-3δ: both Fs < 1.5, lowest p-value = 0.158; 3λ-1β-2δ, 3λ-2β-1δ, 3λ-2β-2δ: all ts < 1.4). Therefore, all subsequent analyses focus only on the Construction factor.

The LMER analysis of the estimates for the model with three rate parameters (3λ-3β-1δ) was non-significant, $F_{(2, 40)} = 1.39$, $p = 0.259$. For the model with three intercept parameters (3λ-1β-3δ), the LMER test revealed a main effect of Construction, $F_{(2, 40)} = 14.21$, $p < 0.001$. Subsequent t-tests revealed that the intercept parameters for the No Interpolation condition were significantly different than both the Interpolated Object Relative ($t = 4.72$) and Interpolated Object + Subject Relative conditions ($t = 4.51$); but the intercepts in the two long conditions (Interpolated Object Relative and Interpolated Object + Subject Relative) were not significantly different ($t < 1$). Both LMER analyses indicate that the addition of a third dynamics parameter is not warranted, and that only models with two dynamics parameters are justified.

We now turn to the models with two dynamics parameter estimates. The more conservative of these models only have six parameters (i.e., 3λ parameters, and 3 parameters divided between the β and δ). T-tests confirm a significant difference between the intercepts in the 3λ-1β-2δ model ($t = 6.41$) and between the rates in the 3λ-2β-1δ model ($t = -2.76$). For the 3λ-2β-2δ model, t-test revealed that the difference between the rate parameter estimates was non-significant ($t = -1.14$); however, the intercept estimates differed significantly ($t = 3.62$). Thus, a conservative interpretation of the current pattern of results suggests that the 3λ-1β-2δ model should be preferred (see **Table 5** for both the average and the individual parameter estimates for this model).

Discussion

The results of this experiment replicate both Experiment 1 and the SR-SAT experiment in McElree et al. (2003). Analyses of both the d' and model asymptote estimates confirm that response accuracy decreased linearly in relation to the amount of material that intervened between sentential NPs and matrix verbs. Thus, as in previous studies, processing the additional interpolated material decreases the likelihood of retrieving the correct constituent and/or mis-parsing the syntactic relations among sentence constituents. These possibilities arise because the additional material either negatively affects the representation of the target constituent, or else the additional material (i.e., the introduction of additional NPs) decreases the diagnostic value of the matrix verbs' retrieval cues (McElree et al., 2003; see also Van Dyke and McElree, 2011; Van Dyke and Johns, 2012). In addition, we also observed individual differences in both d' and asymptotic accuracy based on Reading Ability, such that higher ability was associated with more accurate overall performance. However, there was no interaction of Reading Ability with the amount of interpolated material. Thus, the interpretation of the effect of Reading Ability is straightforward: more skilled readers were able to more accurately resolve the subject-verb dependency than less skilled readers, regardless of distance between subject and verb.

We also observed speed dynamics differences showing that access to the critical item was fastest when there was no interpolated material between noun and verb (i.e., when no retrieval was necessary); and, when intervening material necessitated retrieval, the speed of access did not vary according to how much material intervened between noun and verb. Both the modeling and the inferential statistics indicate that retrieval speed is invariant, regardless of the amount of embedded material. In addition, although we observed differences related to Reading Ability in accuracy measures, we observed no effect (or interaction) of Reading Ability with any index of retrieval dynamics. That is, readers retrieved information that was outside focal attention with equal speed, regardless of Reading Ability.

As in Experiment 1, there is important independent evidence that participants were processing the embedded material. Correct rejection rate at each response lag for the conditions with the anomaly within the interpolated region (T5 and T8) was 49.9, 49.7, 49.6, 51.0, 52.4, 53.5, 54.5, 56.4, 55.8, 56.4, 61.2, 61.2, 61.7, and 60.4%. Correct rejection rates for the corresponding conditions containing a sentence-final ungrammaticality were 49.9, 50.2, 50.3, 51.4, 53.3, 56.9, 60.3, 64.0, 66.5, 68.4, 69.6, 69.8, 71.5, and 71.7%. This pattern is identical to that observed in Experiment 1: overall accuracy is higher in the experimental conditions, and responses to the control conditions appearing to follow an exponential function. One distinction between the two experiments is that these rejection rates—although still clearly

TABLE 5 | Experiment 1: adjusted R^2, d's, and parameter estimates for the average data and individual participants for the 3λ-1β-2δ exponential model.

	Adj.	*d*-primes			Asymptotes			Rate	Intercept	
	R^2	d' 1	d' 2	d' 3	λ1	λ2	λ3	β1	δ1	δ2
Avg	**0.995**	**2.41**	**1.32**	**0.73**	**2.690**	**1.482**	**0.834**	**0.779**	**1.266**	**1.579**
S1	0.978	2.79	1.58	1.32	3.114	1.817	1.534	0.684	1.195	1.287
S2	0.977	3.44	1.97	1.15	3.534	1.759	1.110	1.572	0.895	1.550
S3	0.980	2.20	0.96	0.35	2.363	1.040	0.440	1.058	1.429	1.990
S4	0.980	1.55	0.62	0.68	1.640	0.641	0.677	3.608	0.876	1.214
S5	0.963	1.34	0.72	0.23	1.402	0.729	0.400	3.386	1.563	1.951
S6	0.984	2.71	1.14	0.73	3.097	1.337	0.882	0.771	1.547	1.893
S7	0.979	2.73	2.15	0.93	2.836	2.128	1.010	1.719	1.827	2.282
S8	0.992	1.71	0.97	0.60	2.007	1.203	0.755	0.972	2.258	2.583
S9	0.986	2.09	1.50	0.76	2.057	1.554	0.708	4.848	2.957	2.814
S10	0.984	1.83	1.08	0.32	1.915	1.172	0.400	1.074	1.353	1.938
S11	0.982	3.09	3.13	2.29	3.287	3.276	2.388	0.964	0.982	1.248
S12	0.969	2.59	1.18	0.75	2.776	1.382	0.875	1.223	2.148	2.490
S13	0.993	2.35	0.73	0.30	2.315	0.757	0.400	2.367	1.335	1.602
S14	0.982	3.32	1.56	0.81	3.365	1.524	0.720	2.252	1.717	1.714
S15	0.984	2.45	1.00	1.16	2.812	1.214	1.283	0.772	1.626	1.687
S16	0.991	2.82	1.62	0.55	2.938	1.560	0.567	3.810	1.601	1.684
S17	0.984	1.86	0.86	0.43	1.930	0.890	0.481	1.751	2.250	2.291
S18	0.982	2.13	0.85	0.39	2.126	0.818	0.400	1.690	1.177	1.292
S19	0.988	1.66	0.76	0.43	1.699	0.759	0.482	1.405	1.594	1.882
S20	0.987	2.91	1.26	0.62	3.179	1.375	0.652	1.065	1.601	1.848
S21	0.989	2.81	1.81	0.60	3.104	2.003	0.710	1.126	2.026	2.383
S22	0.988	2.63	1.51	0.63	2.898	1.765	0.668	0.876	1.466	1.747

above chance for all conditions—are lower than those in the first experiment. This is consistent with the overall performance of the participants in this experiment, who had considerably lower *d*'s in every condition than the university students in Experiment 1, and is undoubtedly a function of the broader range of reading ability.

This pattern of results—in which accuracy differs systematically according to the amount of material interpolated between the retrieval cue and the to-be-accessed item, but retrieval speed does not—is once again consistent only with content-addressable, direct-access retrieval. Thus, this experiment provides the first evidence that memory capacity is not important for argument integration in both skilled and less-skilled readers during listening comprehension. In addition, based on these results, the slowing associated with poor reading comprehension (e.g., Perfetti, 1985; Swan and Goswami, 1997a,b; Wolf and Bowers, 1999; Goswami, 2011) cannot be directly attributed to the absence of an efficient mechanism for retrieving critical information from memory. That is, the direct-access retrieval mechanism that is thought to subserve basic memory operations (see Clark and Gronlund, 1996), and which has been observed during language processing in collegiate readers (e.g., McElree et al., 2003; for review see McElree, 2015) was not innately compromised in our sample of less-skilled readers. These results also indicate that direct-access retrieval is not the result of increasingly proficient reading ability, as many of our participants had low word reading and comprehension ability (see **Table 4**). This suggests that a model of retrieval from LTM based on task-specific expertise (e.g., Ericsson and Kintsch, 1995) does not support argument integration during routine language processing. Rather, the pattern of results we observed suggests that individual variation in language processing is driven by the quality of the representation to be retrieved, and not the mechanism by which it is retrieved (Van Dyke and Shankweiler, 2013). This conclusion is bolstered by the use of the auditory SAT procedure: none of our effects can be attributed to either felicitous or impaired processing based on orthographic information (Harm and Seidenberg, 2004).

GENERAL DISCUSSION

These experiments contribute to the growing body of evidence in support of cue-based direct-access retrieval as the memory mechanism supporting argument integration during online sentence processing. Both of our experiments demonstrate the signature pattern of direct-access retrieval: variation in accuracy based on dependency distance, but constant retrieval speed when a distal constituent is required to complete a long-distance dependency. Our results replicate previous findings that suggest that a direct-access retrieval mechanism supports online parsing operations (McElree et al., 2003; see also McElree, 2000; Martin and McElree, 2008, 2009, 2011; Van Dyke and McElree, 2011).

Our results also extend previous findings, as we are the first to report that this type of mechanism supports comprehension of spoken language. As such, these studies suggest that direct-access retrieval is modality independent. Consequently, they further suggest that this retrieval mechanism, long known to subserve basic memory operations outside the domain of linguistic processing, may also be a core property of the human language faculty (see also McElree, 2015).

Our findings with respect to reading ability are consistent with this possibility. The results of Experiment 2 confirm that poor readers do not *de novo* employ a qualitatively different memory mechanism than that used by good comprehenders. Moreover, the use of the SAT methodology allows us to make several nuanced (and, perhaps, surprising) claims with respect to poor reading ability. For example, that we observed no main effects or interactions of Reading Ability on indices of retrieval speed may be unexpected in light of the many previous reports of lower fluency and slower reading rates in poor readers (for reviews see Torgesen et al., 2001; Chard et al., 2002); models of reading frequently attribute such behavior to impaired speed of retrieval (e.g., LaBerge and Samuels, 1974; Stanovich, 2000). However, because standard fluency measures capture both speed and the overall quality of readers' interaction with a text (Adams, 1990; Ashby et al., 2013), they do not take into consideration the speed-accuracy tradeoffs inherent in any timed assessment. Accordingly, it is not possible to clearly distinguish the contributions of representation quality and memory access speed to reading speed measures with traditional assessments.

In contrast, the implication of our results are clear: differences in representational quality, rather than in retrieval speed, contribute more to a comprehender's performance. Specifically, all our effects of Reading Ability were found only on the asymptote, which is understood within the SAT literature as an index of representation quality (e.g., memory strength; see Dosher, 1979; Wickelgren et al., 1980). Indeed, readers are known to differ in their ability to differentiate memory representations along various dimensions, with skilled readers able to make fine-grained distinctions that less skilled readers cannot (Perfetti and Hart, 2002; Perfetti et al., 2005; Landi and Perfetti, 2007; Perfetti, 2007; see also Long and Prat, 2008). Clinical reports showing that dyslexic readers are less able to make linguistically relevant phonetic distinctions compared to age-matched reading-level controls (e.g., Bogliotti et al., 2008; Goswami et al., 2011) are also consistent with this interpretation. Finally, there is also evidence that interventions that specifically attempt to increase reading speed are largely unsuccessful (Torgesen et al., 2001; Berends and Reitsma, 2005; Marinus et al., 2012), unless the intervention seeks to strengthen the representation of specific words or word parts (Mattingly, 1972; National Reading Panel, 2000; Thaler et al., 2004; Conrad and Levy, 2011). Findings such as these support the argument that representational quality is the crucial determinant of whether a given representation will be available for argument integration (e.g., Perfetti and Hart, 2002; Perfetti, 2007; Perfetti et al., 2008; Frishkoff et al., 2011).

Our observation of direct-access retrieval in our poor readers has important implications for the study of, and remediation of, reading difficulty and disability. Although the current study of auditory sentence processing does not demonstrate that poor readers employ direct access during reading, it does demonstrate that direct-access retrieval is not inherently "broken" or unavailable to these readers. This suggests that, like skilled readers, they are eligible to use a parsing architecture characterized by a severely limited active memory and an efficient direct access retrieval mechanism (Lewis et al., 2006). Because all readers, regardless of skill, have the minimal capacity required by such a system—the most recently processed item—inherent differences in WM capacity cannot be the source of comprehension difficulty, at least with respect to basic argument integration. Further support for this position comes from our recent study of a community-based sample of adult readers, in which comprehension of visual sentences was related not to WM capacity but, as in our second experiment, to receptive vocabulary (Van Dyke et al., 2014). Other recent work, in which first-grade children's development of reading comprehension skill was tracked before, during, and after intensive training on WM tasks, is similarly consistent: even when WM performance increased significantly, there was no measurable effect on the children's development of reading comprehension skill (Fuchs et al., 2014; see also Banales et al., 2015).

Poor quality lexical representations have a particularly serious impact on the efficiency of direct-access retrieval, wherein retrieval cues must be able to uniquely identify target representations. If representations do not instantiate important or relevant distinctions, then the mapping between cue and target will be indeterminate, leading to retrieval of incorrect representations. This situation has been studied extensively in the memory domain under the rubric of "cue-overload" (e.g., Watkins and Watkins, 1975) and has also been referred to as retrieval interference (see Van Dyke and Johns, 2012 for a review). Van Dyke and McElree (2006) demonstrated this effect in the language domain using a dual task paradigm (see also Gordon et al., 2002). Participants read sentences such as these:

(8a) It was the boat that the guy who lived by the sea *sailed* over two sunny days.
(8b) It was the boat that the guy who lived by the sea *fixed* over two sunny days.

For each sentence, a memory load was either present or absent; if present, participants received a short list of words to memorize prior to reading the sentence (e.g., TABLE-SINK-TRUCK). The presence of retrieval interference was determined by the main verb. In conditions such as (8a), the verb *sailed* is not overloaded: because the memory list words are not "sailable," the verb's semantic cues are able to uniquely identify the displaced subject NP *boat*. However, in conditions such as (8b), the verb *fixed* is an overloaded retrieval cue: that is, because the semantic cues provided by *fixed* are not uniquely diagnostic of its target in memory, the "fixable" items in the memory list compete with the "fixable" target in the sentence. Van Dyke and McElree found, in university students, that cue-overload increased reading difficulty at the verb—an effect which disappeared when the competing matches in the memory list were absent. (Similar effects in reading paradigms without a dual

task have also been reported; e.g., Gordon et al., 2001; Van Dyke and Lewis, 2003; Van Dyke, 2007.) A subsequent study, using the same paradigm and materials with a community-based sample of participants, found that readers' sensitivity to interference induced by overloaded retrieval cues varied negatively with receptive vocabulary (indexed, as in our SAT experiment, by PPVT; Van Dyke et al., 2014). In that study, low vocabulary scores were uniquely predictive of greater interference effects, including online reading difficulty and impaired performance on offline comprehension questions. Van Dyke and colleagues proposed that such readers—many of whom also had low scores on a range of other linguistic skill measures—were likely to have lexical representations in which important distinctions (on orthographic, phonologic, and/or semantic dimensions) were absent. It is precisely these distinctions that could be crucial for discriminating among similar, competing lexical representations when a retrieval cue is overloaded.

Van Dyke et al. (2014) were the first to report that poor readers were more vulnerable to retrieval interference than skilled readers. However, the association between low verbal ability and effects related to the strength or quality of representations, rather than retrieval speed, is also broadly consistent with a recent SAT study examining individual differences in interference resolution in recognition memory (Öztekin and McElree, 2010). Using an extreme groups design, Öztekin and McElree assessed recognition of words that were either present in a studied list; absent from, but consistent with the semantic categories of, studied list items ("distant negatives"); or absent from the studied list, but nonetheless present in the immediately preceding study list ("recent negatives"). As in the current study, there were no individual differences associated with retrieval speed, which was invariant for all items but the most recently processed list word. Also as reported here, individual differences emerged only on the SAT parameter associated with representation quality: low ability participants had lower asymptotic accuracy. This difference was driven by low ability participants' greater rate of false alarms to the recent negative lure trials. Öztekin and McElree suggested that this greater susceptibility to interference could result from lower-quality representations, or from the impaired ability to distinguish between information based on familiarity and episodic details (i.e., cue-overload). As this study also used the SAT method, we take these results as important corroborating evidence for our own position: namely, that individual differences have their effect on measures of representation quality (or strength), and not on retrieval speed[5]. Taken together, these studies converge on the notion that it is the probability of retrieving the necessary item, determined by qualitative properties of the item's representation, that is a crucial determinant of reading ability—rather than intrinsic capacity differences, or the absence of an efficient retrieval mechanism.

Finally, as this is the first time the SAT method has been used to examine individual differences in language processing, we acknowledge that the suggestion that poor reading ability may be unrelated to slowed retrieval should be treated cautiously. Moreover, although the size of the current sample is in line with other published SAT studies, it would be desirable to replicate our study with an even larger sample to verify our results with respect to speed parameters. However, it is important to note that the main conclusion from this study is actually entirely orthogonal to whether poor reading ability is associated with slower retrieval speed. The crucial finding here is that regardless of reading ability, retrieval speed was unaffected by the amount of interpolated material between the target subject and its verb. The fact that the speed to access the target subject in our longest condition (Interpolated Object + Subject Relative Clause condition) was the same as that for accessing the target in the shorter Interpolated Object Relative Clause condition means that these retrievals occurred without executing a backwards sequential search through the contents of memory. Rather, all participants employed a direct-access retrieval mechanism irrespective of Reading Ability. Thus, even if we had observed a main effect of ability on speed parameters, this would have only attested to the possibility that retrieval was slower overall. This would have said nothing about the presence or absence of a direct-access retrieval mechanism in poor comprehenders.

The experiments reported here validate the auditory SAT procedure as a useful, highly sensitive tool for investigating the architecture of language comprehension across individuals with widely varying linguistic abilities. Because it gauges performance in the auditory modality, the procedure is not susceptible to problems related to inefficient orthographic decoding skills that confound other online assessments. This opens up new possibilities for investigations of memory access during language processing to special populations, such as adolescents with poor reading comprehension, dyslexics, spoken language bilinguals (e.g., heritage language speakers), or functionally illiterate language users. In addition, because longer, multi-sentence and passage-length materials have been difficult to implement in the visual SAT paradigm, our findings suggest the possibility of investigating memory retrieval during the online processing of discourse-level dependencies. Especially considered alongside the potential to investigate the influence of a broader range cognitive abilities on the dynamics and accuracy of memory retrieval during online language comprehension, the results of this study raise many exciting possibilities for future research.

[5]Öztekin and McElree (2010) used a working memory assessment as their only skill measure. There is much evidence for high correlations between this assessment and other language and reading-related measures, including vocabulary, phonological processing, decoding ability, rapid naming, reading fluency, and spoken language ability. Hence, we prefer to interpret this result as referring to a more general verbal skill ability rather than about working memory capacity per se. See Van Dyke et al. (2014) for further discussion of this issue.

ACKNOWLEDGMENTS

This research was supported by the following NIH grants to Haskins Laboratories: R21 HD-058944 (JV, PI), R01 HD-073288 (JV, PI), R01 HD-071988 (David Braze, PI), and P01 HD-01994 (Jay G. Rueckl, PI). The authors are grateful to Brian McElree for access to materials from the McElree et al. (2003)

study; Erica Davis and Joshua Coppola for assistance with data collection; Emma Voytek (née Chepya) for assistance adapting and recording all stimuli; Victor Kuperman for consultations regarding skills testing and statistical analysis; and Morgan L. Bontrager, Andrew A. Jahn, Hannah R. Jones, and Dave Kush for their insightful comments and advice.

REFERENCES

Abney, S. P., and Johnson, M. (1991). Memory requirements and local ambiguities of parsing strategies. *J. Psycholinguist. Res.* 20, 233–250. doi: 10.1007/BF01067217

Adams, M. J. (1990). *Beginning to Read*. Cambridge, MA: MIT Press.

Ashby, J., Dix, H., Bontrager, M., Dey, R., and Archer, A. (2013). Phonemic awareness contributes to text reading fluency: evidence from eye movements. *Sch. Psychol. Rev.* 42, 157–170.

Audacity Team. (2015). *Audacity®: Free Audio Editor and Recorder [Computer program]. Audacity® software is copyright (c) 1999-2015 Audacity Team*. The name Audacity® is a registered trademark of Dominic Mazzoni.

Baayen, R. H. (2004). "Statistics in psycholinguistics: a critique of some current gold standards," in *Mental Lexicon Working Papers 1* (Edmonton, AB), 1–45.

Baayen, R. H. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. New York, NY: Cambridge University Press.

Baayen, R. H., Davidson, D. J., and Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *J. Mem. Lang.* 59, 390–412. doi: 10.1016/j.jml.2007.12.005

Banales, E., Kohnen, S., and McArthur, G. (2015). Can verbal working memory training improve reading? *Cogn. Neuropsychol.* 32, 104–132. doi: 10.1080/02643294.2015.1014331

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2015). *lme4: Linear Mixed-effects Models using Eigen and S4. R package version 1.1-8*. Available online at: http://CRAN.R-project.org/package=lme4

Bell, L. C., and Perfetti, C. A. (1994). Reading skill: some adult comparisons. *J. Educ. Psychol.* 86, 244–255. doi: 10.1037/0022-0663.86.2.244

Berends, I. E., and Reitsma, P. (2005). Lateral and central presentation of words with limited exposure duration as remedial training for reading disabled children. *J. Clin. Exp. Neuropsychol.* 27, 886–896. doi: 10.1080/13803390490919191

Bogliotti, C., Serniclaes, W., Messaoud-Galusi, S., and Sprenger-Charolles, L. (2008). Discrimination of speech sounds by children with dyslexia: comparisons with chronological age and reading level controls. *J. Exp. Child Psychol.* 101, 137–155. doi: 10.1016/j.jecp.2008.03.006

Bornkessel, I., McElree, B., Schlesewsky, M., and Friederici, A. D. (2004). Multi-dimensional contributions to garden path strength: dissociating phrase structure from relational structure. *J. Mem. Lang.* 51, 495–522. doi: 10.1016/j.jml.2004.06.011

Braze, D., Katz, L. Magnuson, J. S., Mencl, W. E., Tabor, W., Van Dyke, J. A., et al. (in press). Vocabulary does not complicate the simple view of reading. *Reading Writing*.

Braze, D., Tabor, W., Shankweiler, D. P., and Mencl, W. E. (2007). Speaking up for vocabulary reading skill differences in young adults. *J. Learn. Disabil.* 40, 226–243. doi: 10.1177/00222194070400030401

Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Comput.* 16, 1190–1208. doi: 10.1137/0916069

Caplan, D., and Waters, G. S. (1999). Verbal working memory and sentence comprehension. *Behav. Brain Sci.* 22, 77–94. doi: 10.1017/S0140525X99001788

Chandler, J. P. (1969). STEPIT-Finds local minima of a smooth function of several parameters. *Behav. Sci.* 14, 81–82.

Chard, D. J., Vaughn, S., and Tyler, B.-J. (2002). A synthesis of research on effective interventions for building reading fluency with elementary students with learning disabilities. *J. Learn. Disabil.* 35, 386–406. doi: 10.1177/00222194020350050101

Clark, S. E., and Gronlund, S. D. (1996). Global matching models of recognition memory: how the models match the data. *Psychon. Bull. Rev.* 3, 37–60. doi: 10.3758/BF03210740

Clifton, C. Jr., and Frazier, L. (1989). "Comprehending sentences with long-distance dependencies," in *Linguistic Structure in Language Processing,* eds G. N. Carlson and M. K. Tanenhaus (Dordrecht: Kluwer Academic Publishers), 273–317.

Conrad, N. J., and Levy, B. A. (2011). Training letter and orthographic pattern recognition in children with slow naming speed. *Read. Writ.* 24, 91–115. doi: 10.1007/s11145-009-9202-x

Daneman, M., and Carpenter, P. A. (1980). Individual differences in working memory and reading. *J. Verbal Learn. Verbal Behav.* 19, 450–466. doi: 10.1016/S0022-5371(80)90312-6

Dennis, J. E., and Schnabel, R. B. (1983). *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Englewood Cliffs, NJ: Prentice-Hall.

Dosher, B. A. (1979). Empirical approaches to information processing: speed-accuracy tradeoff functions or reaction time—A reply. *Acta Psychol.* 43, 347–359. doi: 10.1016/0001-6918(79)90029-5

Dunn, L. M., and Dunn, L. M. (1997). *Peabody Picture Vocabulary Test, 3rd Edn.* Circle Pines, MN: American Guidance Services.

Ericsson, K. A., and Kintsch, W. (1995). Long-term working memory. *Psychol. Rev.* 102, 211–245. doi: 10.1037/0033-295X.102.2.211

Fedorenko, E., Gibson, E., and Rohde, D. (2006). The nature of working memory capacity in sentence comprehension: evidence against domain-specific working memory resources. *J. Mem. Lang.* 54, 541–553. doi: 10.1016/j.jml.2005.12.006

Fodor, J. D., and Inoue, A. (1994). The diagnosis and cure of garden paths. *J. Psycholinguist. Res.* 23, 407–434. doi: 10.1007/BF02143947

Fodor, J. D., and Inoue, A. (1998). "Attach anyway," in *Reanalysis in Sentence Processing,* eds J. D. Fodor and F. Ferriera (Dordrecht: Kluwer), 101–141.

Fodor, J. D., and Inoue, A. (2000). "Garden path re-analysis: attach (anyway) and revision as last resort," in *Cross-linguistic Perspectives on Language Processing,* eds M. D. Vincenzi and V. Lombardo (Dordrecht: Kluwer), 21–61.

Foraker, S., and McElree, B. (2007). The role of prominence in pronoun resolution: active versus passive representations. *J. Mem. Lang.* 56, 357–383. doi: 10.1016/j.jml.2006.07.004

Foraker, S., and McElree, B. (2011). Comprehension of linguistic dependencies: speed—accuracy tradeoff evidence for direct—access retrieval from memory. *Lang. Linguist. Compass* 5, 764–783. doi: 10.1111/j.1749-818X.2011.00313.x

Fox, P. A., Hall, A. P., and Schryer, N. L. (1978). The PORT mathematical subroutine library. *ACM Trans. Math. Softw.* 4, 104–126. doi: 10.1145/355780.355783

Fraser, J., and Conti-Ramsden, G. (2008). Contribution of phonological and broader language skills to literacy. *Int. J. Lang. Commun. Dis.* 43, 552–569. doi: 10.1080/13682820701778069

Frazier, L. (1979). *On Comprehending Sentences: Syntactic Parsing Strategies*. Unpublished doctoral dissertation, University of Connecticut, Storrs.

Frazier, L. (1987). "Sentence processing: a tutorial review," in *Attention and Performance, Vol. XII: The Psychology of Reading,* ed M. Coltheart (Hillsdale, NJ: Erlbaum), 601–681.

Frazier, L. (2013). "Syntax in sentence processing," in *Sentence Processing,* ed R. P. G. Van Gompel (New York, NY: Psychology Press), 21–50.

Frazier, L., and Rayner, K. (1982). Making and correcting errors during sentence comprehension: eye movements in the analysis of structurally ambiguous sentences. *Cogn. Psychol.* 14, 178–210. doi: 10.1016/0010-0285(82)90008-1

Frishkoff, G. A., Perfetti, C. A., and Collins-Thompson, K. (2011). Predicting robust vocabulary growth from measures of incremental learning. *Sci. Stud. Read.* 15, 71–91. doi: 10.1080/10888438.2011.539076

Fuchs, D., Fuchs, L. S., Compton, D. L., Elleman, A., Kearns, D., Peng, P., et al. (2014). "Randomized control trial of the value of cognitive training in a reading comprehension program for poor readers in first grade," in *Talk Presented at the 21st Annual Meeting of the Society for the Scientific Study of Reading 17 July* (Santa Fe, NM).

Gelman, A., and Hill, J. (2007). *Data Analysis using Regression and Multilevel/Hierarchical Models*. New York, NY: Cambridge University Press.

Gibson, E. (1998). Linguistic complexity: locality of syntactic dependencies. *Cognition* 68, 1–76. doi: 10.1016/S0010-0277(98)00034-1

Gibson, E. (2000). "The dependency locality theory: a distance-based theory of linguistic complexity," in *Image, Language, Brain: Papers from the First Mind Articulation Project Symposium,* ed A. Marantz (Cambridge, MA: MIT Press), 94–126.

Gordon, P. C., Hendrick, R., and Johnson, M. (2001). Memory interference during language processing. *J. Exp. Psychol.* 27, 1411–1423. doi: 10.1037/0278-7393.27.6.1411

Gordon, P. C., Hendrick, R., and Levine, W. H. (2002). Memory-load interference in syntactic processing. *Psychol. Sci.* 13, 425–430. doi: 10.1111/1467-9280.00475

Goswami, U. (2011). A temporal sampling framework for developmental dyslexia. *Trends Cogn. Sci.* 15, 3–10. doi: 10.1016/j.tics.2010.10.001

Goswami, U., Fosker, T., Huss, M., Mead, N., and Szücs, D. (2011). Rise time and formant transition duration in the discrimination of speech sounds: the ba-wa distinction in developmental dyslexia. *Dev. Sci.* 14, 34–43. doi: 10.1111/j.1467-7687.2010.00955.x

Hamilton, S. T., Freed, E. M., and Long, D. L. (2013). Modeling reader and text interactions during narrative comprehension: a test of the lexical quality hypothesis. *Discourse Process.* 50, 139–163. doi: 10.1080/0163853X.2012.742001

Harm, M. W., and Seidenberg, M. S. (2004). Computing the meanings of words in reading: cooperative division of labor between visual and phonological processes. *Psychol. Rev.* 111, 662–720. doi: 10.1037/0033-295X.111.3.662

Hogaboam, T. W., and Perfetti, C. A. (1975). Lexical ambiguity and sentence comprehension. *J. Verbal Learn. Verbal Behav.* 14, 265–274. doi: 10.1016/S0022-5371(75)80070-3

Johns, C. L., Braze, D., Molfese, P. J., Van Dyke, J. A., Kush, D., Magnuson, J. S., et al. (2014). "Structural MRI reveals correlations between individual differences in language-related cognitive abilities and thickness of language-relevant cortical areas," in *Poster Presented at the 21st Annual Meeting of the Cognitive Neuroscience Society 4 April* (Boston, MA).

Joshi, R. M. (2005). Vocabulary: a critical component of comprehension. *Read. Writ. Q.* 21, 209–219. doi: 10.1080/10573560590949278

Judd, C. M., and McClelland, G. H. (1989). *Data Analysis: A Model-comparison Approach*. San Diego, CA: Harcourt-Brace Jovanovich.

Just, M. A., and Carpenter, P. A. (1992). A capacity theory of comprehension: individual differences in working memory. *Psychol. Rev.* 99, 122–149. doi: 10.1037/0033-295X.99.1.122

King, J., and Just, M. A. (1991). Individual differences in syntactic processing: the role of working memory. *J. Mem. Lang.* 30, 580–602. doi: 10.1016/0749-596X(91)90027-H

Kintsch, W. (1988). The role of knowledge in discourse comprehension: a construction-integration model. *Psychol. Rev.* 95, 163–182. doi: 10.1037/0033-295X.95.2.163

Kintsch, W. (1998). *Comprehension: A Paradigm for Cognition*. New York, NY: Cambridge University Press.

Kuperman, V., and Van Dyke, J. A. (2011). Effects of individual differences in verbal skills on eye-movement patterns during sentence reading. *J. Mem. Lang.* 65, 42–73. doi: 10.1016/j.jml.2011.03.002

Kutner, M., Greenberg, E., Jin, Y., Boyle, B., Hsu, Y. C., Dunleavy, E., et al. (2007). *Literacy in Everyday Life: Results from the 2003 National Assessment of Adult Literacy*. Washington, DC: Institute of Educational Sciences, U.S. Department of Education.

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (2015). *lmerTest: Tests in Linear Mixed Effects Models. R package version 2.0-25*. Available online at: http://CRAN.R-project.org/package=lmerTest

LaBerge, D., and Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cogn. Psychol.* 6, 293–323. doi: 10.1016/0010-0285(74)90015-2

Landi, N., and Perfetti, C. A. (2007). An electrophysiological investigation of semantic and phonological processing in skilled and less-skilled comprehenders. *Brain Lang.* 102, 30–45. doi: 10.1016/j.bandl.2006.11.001

Lewis, R. L., and Vasishth, S. (2005). An activation−based model of sentence processing as skilled memory retrieval. *Cogn. Sci.* 29, 375–419. doi: 10.1207/s15516709cog0000_25

Lewis, R. L., Vasishth, S., and Van Dyke, J. A. (2006). Computational principles of working memory in sentence comprehension. *Trends Cogn. Sci.* 10, 447–454. doi: 10.1016/j.tics.2006.08.007

Long, D. L., Johns, C. L., and Morris, P. E. (2006). "Comprehension ability in mature readers," in *Handbook of Psycholinguistics,* eds M. J. Traxler and M. A. Gernsbacher (Burlington, MA: Academic Press), 801–834.

Long, D. L., and Prat, C. S. (2008). Individual differences in syntactic ambiguity resolution: readers vary in their use of plausibility information. *Mem. Cogn.* 36, 375–391. doi: 10.3758/MC.36.2.375

Long, D. L., Prat, C., Johns, C., Morris, P., and Jonathan, E. (2008). The importance of knowledge in vivid text memory: an individual-differences investigation of recollection and familiarity. *Psychon. Bul. Rev.* 15, 604–609. doi: 10.3758/PBR.15.3.604

MacDonald, M. C., and Christiansen, M. H. (2002). Reassessing working memory: comment on Just and Carpenter (1992) and Waters and Caplan (1996). *Psychol. Rev.* 109, 35–54. doi: 10.1037/0033-295X.109.1.35

MacDonald, M. C., Just, M. A., and Carpenter, P. A. (1992). Working memory constraints on the processing of syntactic ambiguity. *Cogn. Psychol.* 24, 56–98. doi: 10.1016/0010-0285(92)90003-K

MacDonald, M. C., Pearlmutter, N. J., and Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychol. Rev.* 101, 679–703. doi: 10.1037/0033-295x.101.4.676

MacGinitie, W., MacGinitie, R., Maria, K., and Dreyer, L. (2000). *Gates-MacGinitie Reading Tests (4E)*. Itasca, IL: Riverside.

Magnuson, J. S., Kukona, A., Braze, D., Johns, C. L., Van Dyke, J. A., Tabor, W., et al. (2011). "Phonological instability in young adult poor readers: Time course measures and computational modeling," in *Dyslexia Across Languages: Orthography and the Brain-Gene-Behavior Link*, eds P. McCardle, B. Miller, R. J. Lee, and O. Tzeng (Baltimore: Paul, H. Brookes), 184–201.

Marinus, E., de Jong, P., and van der Leij, A. (2012). Increasing word-reading speed in poor readers: no additional benefits of explicit letter-cluster training. *Sci. Stud. Read.* 16, 166–185. doi: 10.1080/10888438.2011.554471

Martin, A. E., and McElree, B. (2008). A content-addressable pointer mechanism underlies comprehension of verb-phrase ellipsis. *J. Mem. Lang.* 58, 879–906. doi: 10.1016/j.jml.2007.06.010

Martin, A. E., and McElree, B. (2009). Memory operations that support language comprehension: evidence from verb-phrase ellipsis. *J. Exp. Psychol.* 35, 1231–1239. doi: 10.1037/a0016271

Martin, A. E., and McElree, B. (2011). Direct-access retrieval during sentence comprehension: evidence from sluicing. *J. Mem. Lang.* 64, 327–343. doi: 10.1016/j.jml.2010.12.006

Mattingly, I. G. (1972). "Reading, the linguistic process, and linguistic awareness," in *Language by Ear and by Eye: The Relationship between Speech and Reading*, eds J. F. Kavanaugh and I. G. Mattingly (Oxford: Oxford University Press).

McElree, B. (1993). The locus of lexical preference effects in sentence comprehension: a time-course analysis. *J. Mem. Lang.* 32, 536–571. doi: 10.1006/jmla.1993.1028

McElree, B. (1996). Accessing short-term memory with semantic and phonological information: a time-course analysis. *Mem. Cogn.* 24, 173–187. doi: 10.3758/BF03200879

McElree, B. (1998). Attended and non-attended states in working memory: accessing categorized structures. *J. Mem. Lang.* 38, 225–252. doi: 10.1006/jmla.1997.2545

McElree, B. (2000). Sentence comprehension is mediated by content-addressable memory structures. *J. Psycholinguist. Res.* 29, 111–123. doi: 10.1023/A:1005184709695

McElree, B. (2001). Working memory and focal attention. *J. Exp. Psychol.* 27, 817–835. doi: 10.1037/0278-7393.27.3.817

McElree, B. (2006). "Accessing recent events," in *The Psychology of Learning and Motivation,* Vol. 46, ed B. H. Ross (San Diego, CA: Academic Press), 155–200.

McElree, B. (2015). "Memory processes underlying real-time language comprehension," in *Remembering: Attributions, Processes, and Control in Human Memory: Essays in honor of Larry Jacoby,* eds D. S. Lindsay, C. M. Kelley, A. P. Yonelinas, and H. L. III Roedinger (New York, NY: Psychology Press), 133–152.

McElree, B., and Dosher, B. A. (1989). Serial position and set size in short-term memory: the time course of recognition. *J. Exp. Psychol.* 118, 346–373. doi: 10.1037/0096-3445.118.4.346

McElree, B., and Dosher, B. A. (2001). The focus of attention across space and across time. *Behav. Brain Sci.* 24, 129–130. doi: 10.1017/S0140525X01373922

McElree, B., and Dyer, L. (2013). "Beyond capacity: the role of memory processes in building linguistic structure in real time," in *Language down the Garden Path: The Cognitive and Biological Basis for Linguistic Structures,* eds M. Sanz, I. Laka, and M. K. Tanenhaus (Oxford: Oxford University Press), 229–240.

McElree, B., Foraker, S., and Dyer, L. (2003). Memory structures that subserve sentence comprehension. *J. Mem. Lang.* 48, 67–91. doi: 10.1016/S0749-596X(02)00515-6

McElree, B., and Griffith, T. (1995). Syntactic and thematic processing in sentence comprehension: evidence for a temporal dissociation. *J. Exp. Psychol.* 21, 134.

McElree, B., and Griffith, T. (1998). Structural and lexical constraints on filling gaps during sentence comprehension: a time-course analysis. *J. Exp. Psychol.* 24, 432–460. doi: 10.1037/0278-7393.24.2.432

McKoon, G., and Ratcliff, R. (1992). Inference during reading. *Psychol. Rev.* 99, 440–466. doi: 10.1037/0033-295X.99.3.440

McNamara, D. S., and Magliano, J. (2009). Toward a comprehensive model of comprehension. *Psychol. Learn. Motiv.* 51, 297–384. doi: 10.1016/S0079-7421(09)51009-2

Miller, G. A., and Chomsky, N. (1963). "Finitary models of language users," in *Handbook of Mathematical Psychology,* Vol. II, eds R. D. Luce, R. R. Bush, and E. Galanter (New York, NY: Wiley), 419–492.

Myers, J. L., and O'Brien, E. J. (1998). Accessing the discourse representation during reading. *Discourse Process.* 26, 131–157. doi: 10.1080/01638539809545042

National Institute for Literacy. (2008). *Developing Early Literacy: Report of the National Early Literacy Panel.* Washington, DC: National Institute for Literacy.

National Reading Panel. (2000). *Teaching Children to Read: An Evidence-based Assessment of the Scientific Research Literature on Reading and its Implications for Reading Instruction*. Washington, DC: National Institute of Child Health and Human Development.

Nieuwland, M. S., and Van Berkum, J. J. (2006). Individual differences and contextual bias in pronoun resolution: evidence from ERPs. *Brain Res.* 1118, 155–167. doi: 10.1016/j.brainres.2006.08.022

Ouellette, G., and Beers, A. (2010). A not-so-simple view of reading: how oral vocabulary and visual-word recognition complicate the story. *Read. Writ.* 23, 189–208. doi: 10.1007/s11145-008-9159-1

Öztekin, I., and McElree, B. (2010). Relationship between measures of working memory capacity and the time course of short-term memory retrieval and interference resolution. *J. Exp. Psychol.* 36, 383–397. doi: 10.1037/a0018029

Perfetti, C. A. (1985). *Reading Ability*. New York, NY: Oxford University Press.

Perfetti, C. A. (2007). Reading ability: lexical quality to comprehension. *Sci. Stud. Read.* 11, 357–383. doi: 10.1080/10888430701530730

Perfetti, C. A., and Hart, L. (2002). "The lexical quality hypothesis," in *Precursors of Functional Literacy*, eds L. Verhoeven, C. Elbro and P. Reitsma (Amsterdam: John Benjamins Publishing Company), 67–86.

Perfetti, C. A., Wlotko, E. W., and Hart, L. A. (2005). Word learning and individual differences in word learning reflected in event-related potentials. *J. Exp. Psychol.* 31, 1281–1292. doi: 10.1037/0278-7393.31.6.1281

Perfetti, C. A., Yang, C.-L., and Schmalhofer, F. (2008). Comprehension skill and word-to-text processes. *Appl. Cogn. Psychol.* 22, 303–318. doi: 10.1002/acp.1419

Perin, D. (2013). Literacy skills among academically underprepared students. *Community Coll. Rev.* 41, 118–136. doi: 10.1177/0091552113484057

Prat, C. S., and Just, M. A. (2011). Exploring the neural dynamics underpinning individual differences in sentence comprehension. *Cereb. Cortex* 21, 1747–1760. doi: 10.1093/cercor/bhq241

Psychological Corporation. (1999). *Wechsler Abbreviated Scale of Intelligence*. San Antonio, TX: Harcourt Brace and Co.

Ratcliff, R., and McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. *Neural Comput.* 20, 873–922. doi: 10.1162/neco.2008.12-06-420

R Core Team. (2015). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available online at: http://www.R-project.org/

Reed, A. V. (1976). The time course of recognition in human memory. *Mem. Cogn.* 4, 16–30. doi: 10.3758/BF03213250

Schnabel, R. B., Koontz, J. E., and Weiss, B. E. (1985). A modular system of algorithms for unconstrained minimization. *ACM Trans. Math. Softw.* 11, 419–440. doi: 10.1145/6187.6192

Schneider, W., Eschman, A., and Zuccolotto, A. (2002). *E-prime: A User's Guide*. Pittsburgh, PA: Psychology Software Tools.

Shankweiler, D., and Crain, S. (1986). Language mechanisms and reading disorder: a modular approach. *Cognition* 24, 139–168. doi: 10.1016/0010-0277(86)90008-9

Shankweiler, D., Mencl, W. E., Braze, D., Tabor, W., Pugh, K. R., and Fulbright, R. K. (2008). Reading differences and brain: cortical integration of speech and print in sentence processing varies with reader skill. *Dev. Neuropsychol.* 33, 745–775. doi: 10.1080/87565640802418688

Stanovich, K. E. (2000). *Progress in Understanding Reading: Scientific Foundations and New Frontiers*. New York, NY: Guilford Press.

Swan, D., and Goswami, U. (1997a). Picture naming deficits in developmental dyslexia: the phonological representations hypothesis. *Brain Lang.* 56, 334–353.

Swan, D., and Goswami, U. (1997b). Phonological awareness deficits in developmental dyslexia and the phonological representations hypothesis. *J. Exp. Child Psychol.* 66, 18–41.

Thaler, V., Ebner, E. M., Wimmer, H., and Landerl, K. (2004). Training reading fluency in dysfluent readers with high reading accuracy: word specific effects but low transfer to untrained words. *Ann. Dyslexia* 54, 89–111. doi: 10.1007/s11881-004-0005-0

Torgesen, J. K., Rashotte, C. A., and Alexander, A. W. (2001). "Principles of fluency instruction in reading: Relationships with established empirical outcomes," in *Dyslexia, Fluency and the Brain*, ed M. Wolf (Timonium, MD: York Press), 333–355.

Traxler, M. J., Johns, C. L., Long, D. L., Zirnstein, M., Tooley, K. M., and Jonathan, E. (2012). Individual differences in eye-movements during reading: working memory and speed-of-processing effects. *J. Eye Mov. Res.* 5, 1–16.

Traxler, M. J., and Tooley, K. M. (2007). Lexical mediation and context effects in sentence processing. *Brain Res.* 1146, 59–74. doi: 10.1016/j.brainres.2006.10.010

Tunmer, W. E., and Chapman, J. W. (2012). The simple view of reading redux vocabulary knowledge and the independent components hypothesis. *J. Learn. Disabil.* 45, 453–466. doi: 10.1177/0022219411432685

van den Broek, P., Young, M., Tzeng, Y., and Linderholm, T. (1999). "The landscape model of reading: inferences and the online construction of a memory representation," in *The Construction of Mental Representations during Reading,* eds H. van Oostendorp and S. R. Goldman (Mahwah, NJ: Lawrence Erlbaum Associates Publishers), 71–98.

Van Dyke, J. A. (2007). Interference effects from grammatically unavailable constituents during sentence processing. *J. Exp. Psychol.* 33, 407–430. doi: 10.1037/0278-7393.33.2.407

Van Dyke, J. A., and Johns, C. L. (2012). Memory interference as a determinant of language comprehension. *Lang. Linguist. Compass* 6, 193–211. doi: 10.1002/lnc3.330

Van Dyke, J. A., Johns, C. L., and Kukona, A. (2014). Low working memory capacity is only spuriously related to poor reading comprehension. *Cognition* 131, 373–403. doi: 10.1016/j.cognition.2014.01.007

Van Dyke, J. A., and Lewis, R. L. (2003). Distinguishing effects of structure and decay on attachment and repair: a cue-based parsing account of recovery from misanalyzed ambiguities. *J. Mem. Lang.* 49, 285–316. doi: 10.1016/S0749-596X(03)00081-0

Van Dyke, J. A., and McElree, B. (2006). Retrieval interference in sentence comprehension. *J. Mem. Lang.* 55, 157–166. doi: 10.1016/j.jml.2006.03.007

Van Dyke, J. A., and McElree, B. (2011). Cue-dependent interference in comprehension. *J. Mem. Lang.* 65, 247–263. doi: 10.1016/j.jml.2011.05.002

Van Dyke, J. A., and Shankweiler, D. P. (2013). "From verbal efficiency theory to lexical quality: The role of memory processes in reading comprehension," in *Reading: From Words to Multiple Texts,* eds M. A. Britt, S. R. Goldman and J.-F. Rouet (New York, NY: Routledge, Taylor and Francis Group), 115–131.

Van Gompel, R. P. G. (2013). *Sentence Processing*. New York, NY: Psychology Press.

Wagers, M., and McElree, B. (2009). "Focal attention and the timing of memory retrieval in language comprehension," in *Paper Presented at Architectures and Mechanisms for Language Processing Conference 15, September 7–9* (Barcelona).

Warner, J., and Glass, A. L. (1987). Context and distance-to-disambiguation effects in ambiguity resolution: evidence from grammaticality judgments of garden path sentences. *J. Mem. Lang.* 26, 714–738. doi: 10.1016/0749-596X(87)90111-2

Waters, G. S., and Caplan, D. (2003). The reliability and stability of verbal working memory measures. *Behav. Res. Methods Instrum. Comput.* 35, 550–564. doi: 10.3758/BF03195534

Watkins, O. C., and Watkins, M. J. (1975). Buildup of proactive inhibition as a cue-overload effect. *J. Exp. Psychol.* 1, 442–452. doi: 10.1037/0278-7393.1.4.442

Wickelgren, W. A. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta Psychol.* 41, 67–85. doi: 10.1016/0001-6918(77)90012-9

Wickelgren, W. A., Corbett, A. T., and Dosher, B. A. (1980). Priming and retrieval from short-term memory: a speed accuracy trade-off analysis. *J. Verbal Learn. Verbal Behav.* 19, 387–404. doi: 10.1016/S0022-5371(80)90276-5

Wolf, M., and Bowers, P. G. (1999). The double-deficit hypothesis for the developmental dyslexias. *J. Educ. Psychol.* 91, 415–438. doi: 10.1037/0022-0663.91.3.415

Woodcock, R. W., McGrew, K. S., and Mather, N. (2001). *Woodcock-Johnson III Tests of Achievement.* Itasca, IL: Riverside.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2015 Johns, Matsuki and Van Dyke. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 27 October 2015
doi: 10.3389/fpsyg.2015.01629

Processing implicit control: evidence from reading times

Michael McCourt[1]*, *Jeffrey J. Green*[2], *Ellen Lau*[2] and *Alexander Williams*[1, 2]

[1] Department of Philosophy, University of Maryland, College Park, MD, USA, [2] Department of Linguistics, University of Maryland, College Park, MD, USA

Sentences such as "The ship was sunk to collect the insurance" exhibit an unusual form of anaphora, implicit control, where neither anaphor nor antecedent is audible. The non-finite reason clause has an understood subject, PRO, that is anaphoric; here it may be understood as naming the agent of the event of the host clause. Yet since the host is a short passive, this agent is realized by no audible dependent. The putative antecedent to PRO is therefore implicit, which it normally cannot be. What sorts of representations subserve the comprehension of this dependency? Here we present four self-paced reading time studies directed at this question. Previous work showed no processing cost for implicit vs. explicit control, and took this to support the view that PRO is linked syntactically to a silent argument in the passive. We challenge this conclusion by reporting that we also find no processing cost for remote implicit control, as in: "The ship was sunk. The reason was to collect the insurance." Here the dependency crosses two independent sentences, and so cannot, we argue, be mediated by syntax. Our Experiments 1–4 examined the processing of both implicit (short passive) and explicit (active or long passive) control in both local and remote configurations. Experiments 3 and 4 added either "3 days ago" or "just in order" to the local conditions, to control for the distance between the passive and infinitival verbs, and for the predictability of the reason clause, respectively. We replicate the finding that implicit control does not impose an additional processing cost. But critically we show that remote control does not impose a processing cost either. Reading times at the reason clause were never slower when control was remote. In fact they were always faster. Thus, efficient processing of local implicit control cannot show that implicit control is mediated by syntax; nor, in turn, that there is a silent but grammatically active argument in passives.

Keywords: anaphora, implicit control, implicit argument, rationale clause, self-paced reading

OPEN ACCESS

Edited by:
Matthew Wagers, University of California, Santa Cruz, USA

Reviewed by:
Nayoung Kwon, Konkuk University, South Korea
Keir Moulton, Simon Fraser University, Canada

***Correspondence:**
Michael McCourt
mmccour2@umd.edu

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: *01 July 2015*
Accepted: *08 October 2015*
Published: *27 October 2015*

Citation:
McCourt M, Green JJ, Lau E and Williams A (2015) Processing implicit control: evidence from reading times. Front. Psychol. 6:1629. doi: 10.3389/fpsyg.2015.01629

BACKGROUND

Sometimes an aspect of speaker meaning has unclear provenance. Is it semantic or pragmatic? Is it or is it not determined, that is, by the structural identity of the sentence itself? In such cases online measures may help us find the source of the meaning, as the two routes to interpretation may take measurably different paths.

One familiar example comes from verb phrase ellipsis, as in (2). After (1), the speaker of (2) means that the Yankees traded an outfielder. But is this decided by the structural identity of his sentence token?

(1) The Red Sox traded an outfielder.
(2) The Yankees did too.

Many answer *yes* (Sag, 1976; Williams, 1977; Fiengo and May, 1994; Merchant, 2001). They say that this use of (2), unlike others, has the verb phrase *trade an outfielder*, with all the structure of the verb phrase in (1), just silent. Others answer *no* (Dalrymple et al., 1991; Hardt, 1993; Ginzburg and Sag, 2000; Culicover and Jackendoff, 2005). Every use of (2), they say, has an unstructured verb phrase that simply means *P*, where *P* is a free variable over properties. The value of that variable is then decided "by context," not by the sentence itself. On the first account, the string in (2) is ambiguous between infinitely many sentences, each with a different verb phrase and hence a different meaning. On the second, it has a single meaning that is sensitive to context. These two routes to interpretation—semantic vs. pragmatic, disambiguation vs. anaphora, recovery of structure vs. resolution of a variable—might involve different cognitive processes, and might also register differently in some online processing measure. If they do, that measure may provide some evidence for which account is correct, a question that remains contentious. Accordingly, a rich body of literature has pursued this idea (Tanenhaus and Carlson, 1990; Shapiro and Hestvik, 1995; Frazier and Clifton, 2001, 2005; Martin and McElree, 2008; Kertz, 2010; Yoshida et al., 2012; see Phillips and Parker, 2014 for an overview).

We explore another area in this same light, namely *implicit control of reason clauses,* on display when we use (3) to mean (5).

(3) The candidates were interviewed to find the best person for the job.
(4) Someone interviewed the candidates in order to find the best person for the job.
(5) Someone_k interviewed the candidates in order for them_k to find the best person for the job.

Both (3) and (4) have an infinitival *reason clause* with the verb *find*, adjoined to a *target clause* with the verb *interview*. A reason clause, or *rationale* clause (Faraci, 1974; Jones, 1985), offers a teleological explanation of the fact expressed by its target clause. Why were the candidates interviewed, according to this use of (3)? Because then the interviewers might find the best person for the job. The understood subject of a reason clause, called PRO, may be construed anaphorically, as denoting a thing previously mentioned or implied. Anaphora involving PRO is called *control*, though we commit to no analysis with this term. When (3) is used to mean (5), PRO names the interviewer entailed by the verb in the target clause, *interview*. But the interviewer is named by no audible dependent in that clause; (3) is a *short* passive, with no *by*-phrase. So here control is *implicit*. Control is *explicit* when we use (4) to mean (5). Now the interviewer is audibly realized, here as the subject of an active target clause.

On the *standard theory* of implicit control (Roeper, 1987), the relation is not pragmatic, but syntactic and therefore semantic. Specifically, it is encoded in the context-invariant meaning of the two-part sentence that combines the reason clause and its target clause host; and this encoding goes by way of a syntactic dependency, *binding*[1], which effects sameness of reference. Binding links PRO in the reason clause to a postulated silent argument in the passive target clause, providing PRO with an antecedent.

Semantically, the silent argument is linked to the *deep-S role* of the verb: the semantic relation assigned to the subject of an active clause with that verb. For *interview*, this is the role of interviewer. Syntactically, the silent argument has one of two representations, depending on the analysis of the passive. It may be a formal feature of the verb, part of a feature array that syntactically indexes certain semantic properties, perhaps a "Theta Grid" (Stowell, 1981), "Argument Structure" (Grimshaw, 1990; Manning and Sag, 1998), or "Logical Structure" (van Valin, 1990). Or it may be a separate expression that combines with the verb in syntax (Baker et al., 1989; Stanley, 2000). Either way, the silent argument serves here to provide PRO with a formal antecedent. This allows PRO to be bound, and hence for implicit control to be fixed syntactically, and thus in the compositional semantics. In this way implicit control is assimilated to the paradigm cases of control, where PRO must be coreferent with a particular argument in the next clause up. In (6) or (7), for example, it is must be coreferent with the subject of the *promise* or *rob* clauses, respectively.

(6) Lee heard Mo promise PRO to leave.
(7) Lee robbed Mo while PRO distracting her.

This theory has a good motive. Many restrictions on control of reason clauses, or *reason control*, can be described in syntactic terms (Keyser and Roeper, 1984; Roeper, 1987). When reason control is explicit, the antecedent can be the subject but not the object of its clause (Williams, 1974)[2]. Thus, we can use (8) but not (9) to talk about how the sharks have their gills kept clean, since *these sharks* is the subject in (8) but the object in (9). Conversely only (9) implies that the parasites have gills.

(8) These sharks cover themselves with parasites to have their gills kept clean.
(9) Parasites cover these sharks to have their gills kept clean.

The antecedent can also be a *by*-phrase, when the target clause is a long passive. Thus, we can use (10) to convey that the Red Sox hoped to acquire a better pitcher in trading two outfielders.

(10) Two outfielders were traded by the Red Sox to acquire a better pitcher.

But the right conclusion is not that the antecedent must be assigned the deep-S role of the verb in the target clause. This is not

[1]Binding of PRO is normally called *control*. But we use "control" more neutrally, just to denote the resolution of PRO's reference, whether or not this is decided by binding.

[2]Control by objects is possible, however, for infinitival "purpose clauses" (Faraci, 1974; Williams, 1974; Bach, 1982; Jones, 1985). An example is *Maria brought Mary along to translate*, where the translator is Mary, not Maria. Jones (1985) distinguishes these from reason (rationale) clauses in three further ways, following Williams (1974) and Faraci (1974). Only reason clauses have or permit *in order to*. Only reason clauses can be preposed to sentence-initial position. And only purpose clauses can have a gap in their VP, bound by an argument in the main clause, as in *Mary brought a pen to write with*, where *a pen* binds a gap after *with*. But see Whelpton (2002) for concerns about this taxonomy.

necessary (Williams, 1974; Zubizaretta, 1982; Roeper, 1987), as shown by (11), which can be used to mean that Lisa was arrested so that *she* might seem like a radical.

(11) Lisa was arrested just to seem like a radical.

The better conclusion is that explicit control must be by a *subject*, so long as we presume that a *by*-phrase counts as a subject for at least these purposes. Let us use the term *S* for an argument that "counts as a subject" in this sense, so that reason control must be by an S when explicit. Then we can describe implicit control in analogous terms, if we link the deep-S role in a short passive to a silent S argument, called "implicit" because it is grammatically active. This is the standard theory.

Standard theory in hand, we have a syntactic account of some cases where implicit control is impossible. Sentence (12) describes the theft of a ship, and therefore entails a victim from whom the ship was stolen. But we cannot use (12), it seems, to say that the entailed victim was the intended collector of the insurance, even if he hired the crook for this very purpose. On the standard theory, this is because the role of victim is not linked to an implicit S. And this conclusion is well-justified, since the victim role is assigned to the subject in neither actives nor passives with *steal*.

(12) A hired crook stole the ship to collect the insurance.

Middles, such as (13a), receive a more stipulative account. In a middle as in a short passive, the deep-S role of the verb is assigned to no audible dependent; no audible part of (13a) refers to killers, for example. But with middles this role can never antecede a reason clause PRO (Keyser and Roeper, 1984; Roeper, 1987; Mauner and Koenig, 2000). After (13a), for example, we cannot use (13b) to say that the winter survival of the *killers* explains why prey animals kill easily in the autumn.

(13a) In the autumn the prey animals kill easily.
(13b) #to survive the winter without hunger.

To capture this, the standard theory stipulates a difference in argument structure. In a middle, it says, the deep-S role is not linked to an S, unlike in a passive.

These conclusions have broader implications beyond the analysis of reason clauses, as they make it more plausible that an argument may be silent but grammatically active (Stanley, 2002)[3]. But the standard theory leaves several questions unanswered. It suggests no reason why the implicit S in a passive does not always function as a subject, in relation to all types of adjunct clauses (Vinet, 1988; Iwata, 1999; Landau, 2000), not just reason clauses. By hypothesis (14) has a silent S in the role of thief, and yet we cannot use (14) to mean that my wallet was stolen while the *thief* was distracting me, letting this implicit S control the non-finite temporal adjunct.

(14) My wallet was stolen while distracting me.

The standard theory is also silent on why implicit control is not available to the deep-S role of every passive clause. The meaning that is unavailable to (14) is also unavailable to (13) (Williams, 2015). Yet (15a) is a passive, not a middle, and so should have an implicit S in the role of killer.

(15a) In the autumn the prey animals are killed easily
(15b) #to survive the winter without hunger.

Nor can the standard theory accommodate data like (16) (Williams, 1985, 2015; Lasnik, 1988). Sentence (16) can be used to convey that a young girl cut the ribbon so that the organizers of the event might acquire the support of female voters (Williams, 2015). Yet in a clause with *cut*, there is no argument that stands for organizers of the cutting, as distinct from the cutters.

(16) A young girl cut the ribbon just to acquire the support of female voters.

Finally, the standard theory cannot account for what we call *remote control*, to which we turn in a moment.

Given these doubts, we should welcome additional evidence for the standard theory; and some has been offered in the previous psycholinguistic literature. In a series of stop-making-sense and self-paced reading time studies, Mauner et al. (1995) compared implicit with explicit control of reason clauses. They did so by comparing reason clauses following active, full passive, short passive and intransitive target clauses (15–18).

(17) Someone sank the ship to collect the insurance.
(18) The ship was sunk by someone to collect the insurance.
(19) The ship was sunk to collect the insurance.
(20) The ship sank to collect the insurance.

No differences in acceptability judgments or in reading times were observed in the reason clause in conditions (15–17), but significantly slower reading times and more "unacceptable" responses were observed following the intransitive (20). Mauner and colleagues took these results to support the standard theory of implicit control, on the basis of the following reasoning. First, something like the standard theory of explicit control was assumed: in active examples like (17), PRO is locally bound by the surface subject of the target clause. It was then assumed that finding similar processing profiles for two interpretive dependencies—such as implicit vs. explicit control—would provide evidence that the same mechanisms are at work in resolving them both. Since explicit control by the surface subject of an active target clause is supposed to be mediated syntactically, and since no behavioral differences were observed between explicit (15, 16) and implicit control conditions (19), these earlier results were taken to support the standard view that implicit control is syntactic binding of PRO by a silent argument in the short passive.

Although Mauner and colleagues' results have been taken to constitute important evidence in favor of the standard view of implicit control, this interpretation relies on the assumptions outlined above. In the current study, we test these assumptions further by examining the case of *remote* control. Prior studies

[3]There is a strong case for silent arguments with an anaphoric or "definite" (Fillmore, 1986) interpretation (Partee, 1989; Condoravdi and Gawron, 1996). But the silent argument in a short passive would not be anaphoric or definite. Its interpretation would be equivalent to a narrow-scope existential quantifier: "The candidates were interviewed carefully by *someone*." And the case for such arguments is much weaker (Williams, 2015).

considered only *local* control, where the target and reason clauses are syntactically dependent, forming a single sentence. In *remote* control (Higgins, 1973; Sag and Pollard, 1991; Williams, 2015), as in (21), the two clauses are independent, in two separate sentences. But we can still use (21) to mean (5).

(21) The candidates were interviewed. The goal was to find the best person for the job.
(5) Someone$_k$ interviewed the candidates in order for them$_k$ to find the best person for the job.

In remote control, the infinitival clause is the complement to an equative (or specificational) copula, in a sentence that is separate from the target clause. The subject of the target clause is something like *the goal*, *the reason*, or *the purpose*, a description with a relational noun. We understand that, here, this description is used to refer to a relation that is directed at the target fact, taking *the goal* in (21), for example, to refer to the goal of interviewing the candidates.

Crucially, remote control shows exactly the same restrictions as local control (Williams, 2015). Among others, the contrasts in (1–13) are all preserved when control is remote. (8′) and (9′) show that subjects, but not objects, can be implicit controllers in remote configurations; only (9′) implies that parasites have gills.

(8′) These sharks cover themselves in parasites. The goal is to have their gills kept clean.
(9′) Parasites cover these sharks. The goal is to have their gills kept clean.

And, as with (12) above, it is not possible to use (12′) to mean that the hired crook stole the ship so that his employers could collect the payout.

(12′) A hired crook stole the ship. The reason was to collect the insurance.

Yet here these patterns cannot be explained in terms of syntactic binding. Binding cannot cross independent sentences, and the reason clause, when remote, is syntactically separate from its target. Conceivably—though we do not think that this is correct, for reasons we discuss elsewhere (Green and Williams, in preparation)—the copular clause has hidden structure that conceals a local (same-sentence) binder for PRO, one that is itself anaphoric to an S in the target clause[4]. But even if it did, the anaphoric relation between this local binder and its antecedent in the target clause would still be intersentential. Hence, whatever it is that underlies the interpretive dependency between PRO and the implied interviewer in (21), it cannot be syntactic binding.

The anaphora in (21) must therefore be pragmatic. PRO in a remote reason clause—or, on the alternative that we reject, its hidden local binder—must function not like a bound pronoun, but like a free pronoun or definite description. In turn, the limits on its interpretation cannot be explained directly in terms of structure in the target clause. Rather, its domain of reference must be highly restricted, in terms that only correlate, partially and indirectly, with subjecthood in the target clause[5]. Examples such as (12) suggest that a notion of *responsibility* may be relevant: perhaps PRO in a reason clause, as a matter of the meaning of the construction, ranges only over parties viewed as explanatorily *responsible* for the fact it is meant to explain, a class that may but need not include the individual in the deep-S role to the event of the verb (often, its agent)[6]. However, the grammatical analysis of remote control is beyond the scope of the current work. Here, the key observation is just that the restrictions on local and remote control appear to be identical. Since remote control must be pragmatically mediated, this weakens the motive for a semantic, hence syntactic, account of local control, and at the same time provides a new means of examining the extent to which reading time measures provide support for the syntactic account. If the standard theory is correct, then different mechanisms must be at work in resolution of local (one-sentence) and remote (two-sentence) control: syntactic binding and something like free pronoun interpretation, respectively. On the other hand, if what we now call the *pragmatic theory* is correct, then something like free pronoun interpretation supports resolution not only of remote control, but also of local control.

In the current study, we investigate these alternative hypotheses by examining processing measures in a series of self-paced reading time studies comparing remote and local reason clauses, with and without explicit antecedents. The predictions are the following. Since the standard theory proposes different mechanisms for resolving local and remote control, and the pragmatic theory proposes the same mechanism, the standard theory predicts differences in the processing of local and remote control, while the pragmatic theory does not. These differences might be realized in several ways. First, following the logic in Mauner et al. (1995), implicitness may be costly in forming pragmatic dependencies (because a referent must be inferred), but not costly in forming syntactic dependencies (because binding to the syntactic argument position proceeds in exactly the same way whether it is audible or not). Given this assumption, the standard theory would predict in the current experiments an interaction between implicitness and distance: an effect of implicitness should be present in the pragmatically mediated remote conditions but not in the syntactically mediated local conditions. Second, pragmatically mediated or syntactically mediated dependencies are likely to differ in processing cost

[4]For example, the subject of the copular clause might contain a silenced genitive pronoun, [*his*] *reason*, or a silent relative clause elided under identity with the target clause, *the reason* [*that the candidates were interviewed*]. Then the clause would to be restructured at an unpronounced level of syntax, so that the silent binder is in a local relation to PRO. On the special grammar of copular clauses, see the review in Mikkelsen (2011).

[5]A referential dependency that is intersentential cannot be syntactic, but it may still be sensitive to structural properties of the antecedent. VP Ellipsis is sensitive to voice, for example, and pronominal reference may be sensitive to gender class. Neither of these cases is itself a good model for remote control. But it remains coherent to claim both that PRO in a reason clause is a free anaphor, and that its resolution is sensitive to something like subjecthood. We return to this in passing in the General Discussion.

[6]Constructional restrictions on the domain of a pronoun are not in the standard semantic toolkit, and therefore much more would need to be said. For thoughts see Landau (2000) and Williams (2015), which refine a suggestion in Farkas (1988), with roots in Williams (1974). Also see Whelpton (2002) in opposition, and the discussion in Sag and Pollard (1991).

independent of implicitness because they involve reference to different kinds of memory representations. Therefore, in the current experiments, the standard theory could also predict a main effect of distance, but the direction of this difference depends on the linking hypothesis assumed. If syntactic binding is more costly to resolve than free pronoun interpretation, then the effect of implicitness should be larger in remote control. If free pronoun interpretation is more costly, then the effect of implicitness should be larger in local control. As previous psycholinguistic work does not provide clear predictions about which should be more costly (see Frazier and Clifton, 2000, for discussion), either could be taken to be consistent with the standard theory, although it could also be the case that binding and free pronoun interpretation do not differ fundamentally in processing cost (Cunnings et al., 2014).

The pragmatic theory does not predict any difference in the costs of implicitness in remote vs. local control. If we observed no such differences, this could be due to the fact that local and remote control are both mediated by the same kind of pragmatic mechanism. However, such a conclusion would be too strong here. It might be that reading times in particular are not a sensitive enough measure to detect differences between local and remote control that other measures might detect. Or it could be that processing cost is more generally not a reliable diagnostic of whether a dependency is semantically or pragmatically mediated. However, it is important to remember that Mauner et al.'s (1995) finding of no processing cost for local implicit vs. explicit control is one of the key pieces of evidence currently taken to support the standard theory, and that reading time was one of the online measures used in that study. Skepticism about the ability of self-paced reading to detect differences in processing of local and remote control would thus undermine earlier arguments in favor of the standard theory. These relied on the premise that, in fact, behavioral measures could reflect differences in processing as a function of whether a dependency was semantically or pragmatically mediated. Thus, if we observe no differences in processing of local vs. remote reason clauses, we can at the very least conclude that these earlier results do not in fact provide evidence for the standard account.

EXPERIMENTS

Experiment 1

The goal of Experiment 1 was to test whether differences obtain in processing of local vs. remote control, or between implicit and explicit control within local vs. remote configurations. Experiment 1 manipulated explicitness with passive sentences that varied in the presence or absence of a *by*-phrase that explicitly named the agent of the event described by the passive. Observation of differences in processing of remote and local control, or between implicit and explicit control in local vs. remote configurations, would provide support for the standard theory. Should we observe no such differences, this would either raise a challenge for the standard theory, or undermine previous arguments in its favor, as discussed above.

Methods and Materials

Participants

Participants were 38 native speakers of English from the University of Maryland community. Participants gave informed consent, and received credit in an introductory linguistics course or were compensated $5 for their participation in the experiment. All participants were naïve to the purpose of the experiment. The self-paced reading task lasted for approximately 30 min.

Materials

Sentences were created by combining a finite passive clause with a non-finite reason clause. In a 2 × 2 design, stimuli varied in whether the target clause contained an overt antecedent for the understood subject of the reason clause (*explicitness*), and in whether the target and reason clauses were syntactically independent (*distance*). In conditions labeled *implicit*, the target clause was a short passive and therefore lacked an overt antecedent for PRO. In conditions labeled *explicit*, the target clause was a passive with a *by*-phrase describing the agent of the event, which served as the antecedent of PRO. The dependency was *local* when the reason clause was syntactically an adjunct of the target clause. The dependency was *remote* when the two clauses were syntactically independent, the reason clause being hosted by a copular clause in a separate sentence. An example set of materials is provided in **Table 1**.

In order to control the position of the reason clause across explicit and implicit conditions, we substituted a temporal adjunct, such as *for several hours*, in the implicit conditions in place of the *by*-phrase. Our materials were also crafted to strongly favor interpreting PRO as the satisfier(s) of the deep-S role in the target clause. To this end, we controlled several properties across item sets. First, the reason clause always expressed a property that can be satisfied by people, but not by facts or events. While people can find the best employees for a job, for example, facts or events cannot. This eliminated the possibility, otherwise readily available (Williams, 1985; Lasnik, 1988), of resolving PRO to the fact or event named by the target clause itself. This can happen in (22), which can be used to say that the candidates were interviewed because interviewing them might make a good impression.

(22) Candidates were interviewed to make a good impression.

Second, our passive target clauses mostly had subjects that were semantically implausible as subjects for the reason clause, lowering the chance that they would be taken to antecede PRO. Third, in general our passive target clauses also resisted being read as "adjectival passives," as in *the shoes are polished*. This matters, since adjectival passives do not readily support implicit control by the deep-S role of their verb root. And finally, our target clauses never contained first-person, second-person or impersonal pronouns. This lowered the likelihood, however small, that PRO was read as logophoric or impersonal, like English impersonal *one*, denoting a group that shares the interlocutors' perspective.

Twenty-four sets of four items in these conditions were distributed across four lists in a Latin square design. 96 filler sentences were also included, such that each participant read

TABLE 1 | Experiment 1 materials.

	Regions of analysis											
	Pre-target	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Post-target**
ex. loc:	The candidates were interviewed	by	the	committee				to	find	the	best	person for the job.
im. loc:	The candidates were interviewed	three	weeks	ago				to	find	the	best	person for the job.
ex. rem:	The candidates were interviewed	by	the	committee.	The	reason	was	to	find	the	best	person for the job.
im. rem:	The candidates were interviewed	three	weeks	ago.	The	reason	was	to	find	the	best	person for the job.

a total of 120 sentences. Approximately half of the fillers were one-sentence fillers. The other half were two-sentence fillers, roughly matching the 1:1 ratio of one- to two-sentence items in the main experimental stimuli. Some of the fillers involved adjectival passives and prepositional *to* in order to reduce the likelihood of within-task effects. Such constructions are syntactically very similar to our experimental items, but are differentiated semantically. Their inclusion was intended as a distraction, to make it less likely that readers would gain familiarity during the task with handling reason clauses.

Each sentence was followed by a comprehension question. Comprehension questions varied in whether they targeted information in the target clause, in the reason clause, or concerning the relation between target and reason clause. This reduced the likelihood of participants developing superficial reading strategies during the task.

Procedure

Sentences were displayed on a desktop PC in a moving-window self-paced reading display using the Linger software package (Doug Rohde, MIT). Each sentence initially appeared on a black screen masked by white dashes, with spacing and punctuation intact. Participants revealed the first word by pressing the space-bar on a keyboard. Subsequent words appeared in place of their respective dashes non-cumulatively as participants pressed the space-bar. The order of presentation of target and filler items was randomized for each participant. Participants were instructed to read the sentences carefully and for understanding but at their normal pace. Before the beginning of the experiment, participants were able to gain familiarity with the task with four practice items. Each sentence was followed by a yes/no comprehension question. Incorrect answers to comprehension questions elicited onscreen feedback. The entire procedure took approximately 30 min.

Data analysis

The minimum comprehension question accuracy required for inclusion of a participant's data in the analysis was 80%. Data from two participants were excluded due to comprehension question inaccuracy, resulting in a final dataset of 36 participants.

Statistical analysis was performed in regions 1–10, where region 1 was the first region in which conditions differed (the region beginning the *by*-phrase or temporal adjunct prior to the reason clause), and region 7 was the *to* region that began the critical reason clause. We used mixed-effects linear regressions to assess the reliability of the effects associated with the experimental factors. The effects of *explicitness* and *distance* on reading times were tested using linear mixed effects models in R (Bates et al., 2014; R Core Team, 2014), and p-values were obtained using lmerTest (Kuznetsova et al., 2014), which uses Satterwaithe approximations to calculate degrees of freedom. Note that regions 4–6 (*The reason was*) were only included in the remote conditions, and therefore tests in these regions necessarily examined only the effect of explicitness.

Reading times above 2000 ms were excluded. This resulted in loss of 0.16% of the data. Reading times were then converted to a log scale for statistical analysis. The fixed effects in the model were the factors *explicitness* (explicit vs. implicit), *distance* (local vs. remote), and their interaction. In addition to these fixed-effects, participants, and items were crossed, starting with random intercepts and slopes, and removing one level of complexity until the model converged with correlations of less than 0.9 in random effects in all regions and experiments described here, following the recommendations of Baayen et al. (2008) and Barr et al. (2013). This resulted in a model including random intercepts for subjects and items, but no random slopes.

Results

Mean comprehension question accuracy for experimental stimuli across participants and items was 92% (range across conditions: 91–93%), suggesting that participants were successful in comprehending the main experimental stimuli.

Logged reading times are plotted in **Figure 1**, with significant effects summarized in **Table 2**. Unexpectedly, in self-paced reading times, the most prominent effect we observe was a slowdown for the explicit local condition relative to the other four conditions. Results of the linear mixed effects models are presented in **Table 2**. In regions 1–3, immediately following the short passive, we found no significant main effects of either distance (whether the reason clause is local or remote) or explicitness (whether the target clause is a short or long passive) and no interactions, although the explicit local condition was already numerically slowest by region 3. There was also no significant difference between the remote explicit and implicit conditions in regions 4–6 (*The reason was*), although reading times for the explicit condition were numerically longer. However, at the infinitival *to*, region (9), we observed significant main effects of explicitness ($\beta = -0.14$, $t = -6.0$, $p < 0.001$) and distance ($\beta = -0.09$, $t = -3.7$, $p < 0.001$) and a significant interaction between explicitness and distance ($\beta = 0.09$, $t = 2.8$, $p < 0.01$). Follow-up pairwise comparisons indicated that this

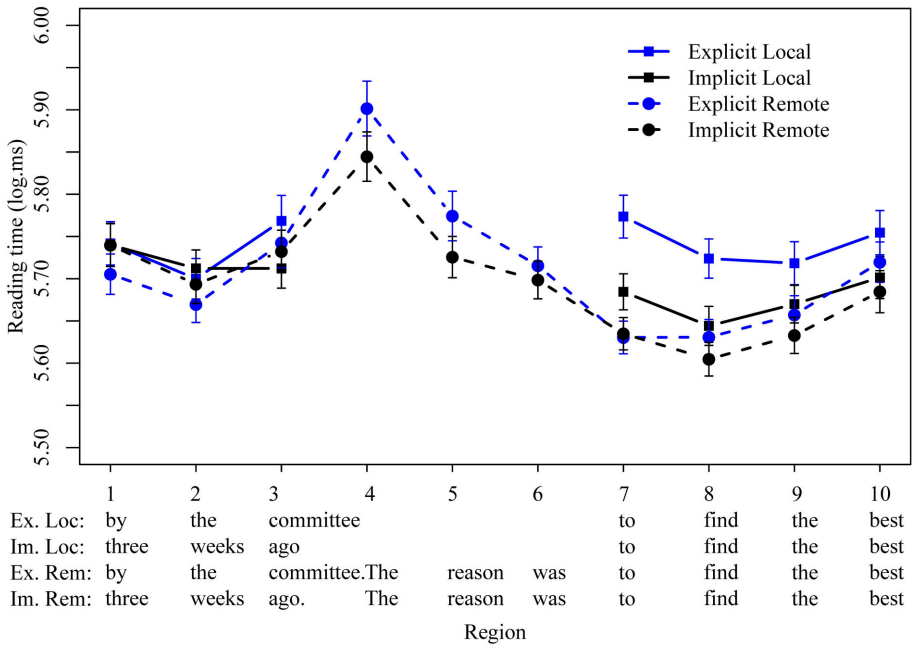


FIGURE 1 | Mean logged reading times (ms) for all four conditions in regions of interest in Experiment 1.

TABLE 2 | Experiment 1 results.

Region	Factors	Estimate	S.E.	*t*-value	*p*-value
	Explicitness	−0.089	0.024	−3.73	<0.001
7	Distance	−0.143	0.024	−5.99	<0.001
To	Explicitness × Distance	0.094	0.034	2.77	0.006
	Explicitness	−0.08	0.024	−3.31	<0.001
8	Distance	−0.093	0.024	−3.39	<0.001
Find	Explicitness × Distance	0.054	0.034	1.57	0.116
	Explicitness	−0.048	0.025	−1.92	0.055
9	Distance	−0.061	0.025	−2.43	0.015
the	Explicitness × Distance	0.024	0.036	0.68	0.498

interaction was driven by a significant effect of explicitness in local conditions ($\beta = -0.09$, $t = -3.3$, $p < 0.001$) but not in remote conditions ($\beta = 0.004$, $t = 0.2$, $p > 0.2$). In short, there appears to be a strong slowing effect of the *by*-phrase on reading times at the infinitival in the reason clause, but only in one-sentence conditions. In region 8 we observed the same pattern numerically, but here the interaction did not reach significance (main effect of distance: $\beta = -0.09$, $t = -3.9$, $p < 0.001$; main effect of explicitness: $\beta = -0.08$, $t = -3.3$, $p < 0.001$; interaction between explicitness and distance: $\beta = 0.05$, $t = 1.6$, $p = 0.1$). These differences diminished following the main verb in regions 9 and 10, although we observed a main effect of distance in region 9 ($\beta = -0.06$, $t = -2.4$, $p < 0.05$) and a marginal main effect of explicitness in region 10 ($\beta = 0.02$, $t = 0.46$, $p > 0.2$).

Discussion

The goal of Experiment 1 was to investigate differences in the cost of resolving implicit vs. explicit control of reason clauses, in local and remote configurations. In this experiment we used a comparison between short and long passives to manipulate whether the antecedent to PRO was implicit or explicit, respectively. The main effect we observed is unexpected according to either the standard theory or the pragmatic theory: the local explicit condition was slower at the beginning of the reason clause than the other three conditions. The logic in Mauner et al. (1995) would predict the longest reading times for the remote implicit condition, which requires the costly operation of inferring an antecedent for PRO. A more generic version of the standard theory would predict only a main effect of distance. What, then, can explain our results? Why were the longest reading times observed when the antecedent was both explicit, and closer to the position at which the dependency is resolved?

Because this pattern contradicts the predictions of all theories about implicit arguments that we are aware of, we conclude that the slowdown for the explicit local condition is most likely due to an independent factor. In particular, we suggest that this slowdown is an index of continued processing difficulty elicited by the preceding *by*-phrase. Normally *by*-phrases carry narrow focus. That is, we would normally read our long passive example—*The candidates were interviewed by the committee*—with prosodic prominence on *committee*, contrasting the committee with other possible interviewers that might be presently relevant. It has been observed that linguistically focused items elicit slower reading times (Lowder and Gordon, 2015). Furthermore, this effect may be particularly pronounced in the

current experiment, because the interpretation of the reason-clause is sensitive to focus (Dretske, 1972): the reason why the candidates were *interviewed*, for example, may not be the reason why they were interviewed *by the committee*. As for why the effect of the *by*-phrase does not obtain in the remote conditions, this is plausibly explained by the availability of more time for processing the focused *by*-phrase prior to the focus-sensitive reason clause during the intervening *The reason was* segment in the remote conditions. Indeed, reading times for the explicit condition were numerically larger during this region.

If this is the correct explanation for the unexpected effects we observe in Experiment 1, then long passives are not ideal as a baseline condition of explicit control, if we want to isolate the specific costs of implicit control. Therefore, in our subsequent experiments, we instead use active transitive clauses for this purpose.

Experiment 2

Experiment 1 tested the effect of explicitness in local vs. remote control configurations by comparing short and long passives, but we found that the *by*-phrases in long passives introduce independent reading time costs. Experiment 2 therefore manipulated explicitness by comparing short passives with active transitive clauses instead.

Methods and Materials

Participants

Participants were 36 native speakers of English from the University of Maryland community. Participants gave informed consent, and received credit in an introductory linguistics course or were compensated $5 for their participation in the experiment. All participants were naïve to the purpose of the experiment. The self-paced reading task lasted for approximately 30 min.

Materials

Twenty-four sets of four target sentences again varied in a 2 × 2 design with the factors *explicitness* and *distance*. However, explicitness is now manipulated by comparing control by short passives with control by active transitive clauses, as shown in **Table 3**. The same fillers and comprehension questions were used as in the earlier experiments. An example set is provided in **Table 3**.

Procedure

The procedure for Experiment 2 was the same as the procedure for Experiment 1.

Data analysis

Comprehension question accuracy was above 80% for all participants. Statistical analysis was performed in regions 1–9, where region 1 was two regions prior to the critical word *to* in the local conditions and region 6 was the *to* region that began the critical reason clause. Analysis procedures were the same as described for Experiment 1. The exclusion of reading times above 2000 ms resulted in a loss of 0.2% of the total data. Note that regions 3–5 (*The reason was*) were only included in the remote conditions, and therefore tests in these regions necessarily examined only the effect of explicitness.

Results

The mean comprehension question accuracy for experimental items across participants and items was 96% (range across conditions: 95–97%), suggesting that participants were successful in comprehending the main experimental stimuli.

Logged reading times are plotted in **Figure 2**, with significant effects summarized in **Table 4**. In the regions preceding the reason clause (1–5), the only significant effect observed was slower reading times for the explicit remote condition than the implicit remote condition for the first word of the second sentence, *The* ($\beta = 0.08$, $t = 2.1$, $p < 0.05$). Although the reason for this difference is not clear, we note that the conditions come back together in the next region and are very tightly matched prior to the beginning of the critical reason clause.

At the infinitival to (region 6), a main effect of distance was observed, with slower reading times for local conditions ($\beta = 0.06$, $t = 2.2$, $p < 0.05$), and we also observed an interaction of explicitness and distance ($\beta = 0.09$, $t = 2.4$, $p < 0.05$). However, this interaction was not in the direction predicted by the standard theory; rather than implicitness requiring costly inference in the pragmatically-mediated remote condition, we observed a cost of implicitness in the local conditions. Pairwise comparisons show that the implicit local condition was significantly slower than the explicit local condition ($\beta = 0.08$, $t = 2.4$, $p < 0.05$), but no differences in reading times were observed for implicit and explicit remote conditions ($p > 0.2$). A similar pattern was observed at the subsequent verb (region 7), with a main effect of distance ($\beta = 0.06$, $t = 2.1$, $p < 0.05$) and a marginal interaction between explicitness and distance ($\beta = 0.06$, $t = 1.7$, $p = 0.09$). At the region following the verb (region 8), no main effects were observed, but we continued to observe an interaction between explicitness and distance, in the same direction ($\beta = 0.08$, $t = 2.1$, $p < 0.05$). No other significant main or interaction effects were observed in the regions of analysis.

TABLE 3 | Experiment 2 materials.

	Regions of analysis										
	Pre-target	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Post-target**
ex. loc:	The committee interviewed	the	candidates				to	find	the	best	person for the job.
im. loc:	The candidates	were	interviewed				to	find	the	best	person for the job.
ex. rem:	The committee interviewed	the	candidates.	The	reason	was	to	find	the	best	person for the job.
im. rem:	The candidates	were	interviewed.	The	reason	was	to	find	the	best	person for the job.

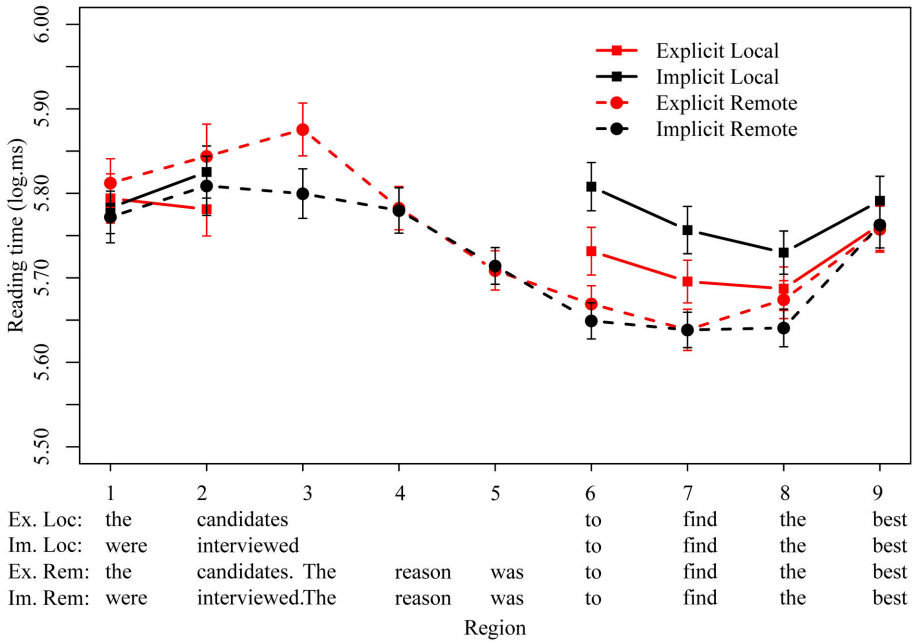


FIGURE 2 | Mean logged reading times (ms) for all four conditions in regions of interest in Experiment 2.

TABLE 4 | Experiment 2 results.

Region	Factors	Estimate	S.E.	*t*-value	*p*-value
	Explicitness	−0.02	0.028	−0.71	0.477
6	Distance	0.062	0.028	2.21	0.027
to	Explicitness × Distance	0.096	0.04	2.42	0.016
	Explicitness	−0.0001	0.027	−0.004	0.997
7	Distance	0.057	0.027	2.13	0.033
find	Explicitness × Distance	0.065	0.038	1.71	0.088
	Explicitness	−0.033	0.025	−1.32	0.186
8	Distance	0.013	0.025	0.52	0.607
the	Explicitness × Distance	0.076	0.036	2.13	0.034

Discussion

The standard theory requires different mechanisms for resolving local and remote control (binding vs. contextual interpretation) and the pragmatic theory proposes the same mechanism (contextual interpretation). Hence, the standard theory predicts differences in the processing of local and remote control, while the pragmatic theory does not. As noted above, these differences might take several forms.

First, Mauner et al. (1995) suggest that syntactic resolution of PRO should have the same processing cost whether the antecedent is explicit or implicit, but that pragmatic resolution of PRO should require costly inference when the antecedent is implicit. According to these assumptions, if local control reflects a syntactically-mediated dependency and remote control reflects a pragmatically-mediated dependency, an interaction between distance and explicitness should be observed such that explicitness has an effect on processing in remote control but not in local control. In Experiment 2 we observed a significant interaction between distance and explicitness at the reason clause, but in the opposite direction: the implicit condition appeared to be costly in the *local* cases and not the remote cases. This pattern is not predicted by either the standard theory or the pragmatic theory, and it also differs from Mauner et al.'s earlier results in which no cost of explicitness was observed for local control of reason clauses.

We hypothesize that the slowdown in the implicit local condition may not reflect the cost of implicitness *per se*, but may rather have been due to the time course of processes elicited by the current materials. We assume that constructing the syntactic and thematic representation associated with the passive may take time (Chow et al., 2015). If this process is not complete by the time the reason clause is encountered, which may have been the case in the local conditions, resolution of PRO will not be immediately possible, causing temporary processing difficulty. However, in the remote conditions, the extra intervening material (*The reason was*) may have acted as a "buffer," providing enough time for the passive sentence to be fully processed by the time the reason clause was encountered. Experiments 3 and 4 include such a buffer in both local and remote conditions and show that this eliminates the cost of implicit control in the local conditions.

Second, the standard theory assumes that local and remote control are mediated by different mechanisms (contextual interpretation and syntactic binding, respectively), and this difference in representational encoding could be reflected online in behavioral measures such as reading time as differences between local and remote configurations that are

independent of explicitness—in other words, a main effect of distance.

In Experiment 2, we observed a significant main effect of distance at the infinitival and the verb in the reason clause, with faster reading times in remote conditions. That is, readers appear to be faster to process a reason clause that is syntactically independent of its target clause as compared to a reason clause whose target clause is a syntactic co-dependent within the same sentence. We refer to this as *the remote speed-up effect* of Experiment 2. These results are thus consistent with the predictions of the standard theory: contextual interpretation of PRO in a reason clause is less costly than syntactic binding.

However, there are also several alternative explanations of the remote speed-up effect, which we explore in Experiments 3 and 4. First, in remote conditions the presence of *the reason was* provided readers with extra time to process the target clause. If target clause was not fully interpreted when the reason clause appeared, processing difficulty would naturally ensue. Second, this phrase also provided readers with a cue that an infinitival reason clause may be on its way. This might facilitate resolution of PRO, and lead to faster reading times in the remote conditions. Experiment 3 tested the first possibility by adding a temporal modifier to the target clause in local conditions to better match the time course with the remote conditions. Experiment 4 tested the second possibility by adding a cue to the reason clause in the local condition (*just in order*) to parallel *the reason was* in the remote conditions.

Experiment 3

Experiment 3 used the same design as Experiment 2 but added an intervening temporal modifier to the local conditions in order to match the distance between the target clause and the reason clause across local and remote conditions. If the overall slowdown for local conditions and the cost of implicitness for local conditions in Experiment 2 were because the target clause had not been fully processed by the time the reason clause was encountered, these differences should be eliminated in Experiment 3.

Methods and Materials

Participants

Participants were 39 native speakers of English from the University of Maryland community, none of whom had taken part in the previous experiments. They either received credit in an introductory linguistics course, or were compensated $10. All participants were naïve to the purpose of the experiment. The self-paced reading task lasted for approximately 30 min.

Materials

Twenty-four sets of four target sentences again varied in a 2 × 2 design with the factors *explicitness* and *distance*. However, to match the amount of time available for processing between the verbs in the local and remote conditions, buffer material was included in the local target clauses, usually a temporal modifier like 3 weeks ago, as shown in **Table 5**. The same fillers and comprehension questions were used as in the earlier experiments. An example set of materials is provided in **Table 5**.

Procedure

The procedure for Experiment 3 was the same as the procedure for Experiments 1 and 2.

Data analysis

The minimum comprehension question accuracy required for inclusion of a participant's data in the analysis was 80%. Data from three participants were excluded due to comprehension question inaccuracy, resulting in a final dataset of 36 participants.

Statistical analysis was the same as described above for Experiment 2, except that because of the additional material in the local conditions, all nine regions of interest could now be analyzed with the full model including both distance and explicitness. Reading times above 2000 ms were again excluded, resulting in a loss of 0.23% of the total data.

Results

Mean comprehension question accuracy for experimental stimuli across participants and items was 91% (range across conditions: 89–92%), suggesting that participants were successful in comprehending the main experimental stimuli.

Logged reading times are plotted in **Figure 3**, with significant effects summarized in **Table 6**. The reason clause began in region 6, and we observed no significant main or interaction effects in preceding regions 1–4. However, region 5, just prior to the infinitival (*was* in the remote conditions and the last word of the temporal modifier in the remote conditions) showed a strong main effect of distance ($\beta = 0.1$, $t = 3.7$, $p < 0.001$) due to faster reading times in remote conditions, and a marginal main effect of explicitness ($\beta = 0.05$, $t = 1.9$, $p = 0.06$) that appeared to be due to slower reading times in the explicit local condition.

TABLE 5 | Experiment 3 materials.

	Regions of the analysis										
	Pre-target	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Post-target**
ex. loc:	The committee interviewed	the	candidates	three	weeks	ago	to	find	the	best	person for the job.
im. loc:	The candidates	were	interviewed	three	weeks	ago	to	find	the	best	person for the job.
ex. rem:	The committee interviewed	the	candidates.	The	reason	was	to	find	the	best	person for the job.
im. rem:	The candidates	were	interviewed.	The	reason	was	to	find	the	best	person for the job.

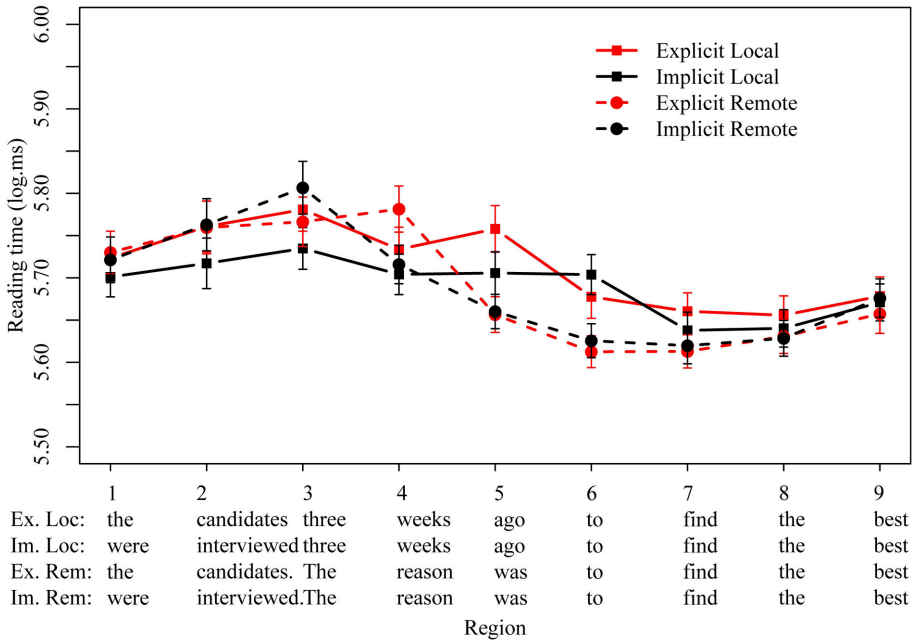


FIGURE 3 | Mean logged reading times (ms) for all four conditions in regions of interest in Experiment 3.

TABLE 6 | Experiment 3 results.

Region	Factors	Estimate	S.E.	*t*-value	*p*-value
	Explicitness	−0.054	0.028	−1.9	0.058
5	Distance	−0.104	0.028	−3.67	0.0003
ago/was	Explicitness × Distance	0.057	0.04	1.44	0.151
	Explicitness	0.026	0.026	1	0.318
6	Distance	0.065	0.026	−2.5	0.013
to	Explicitness × Distance	−0.013	0.037	−0.35	0.724
	Explicitness	−0.023	0.025	−0.95	0.344
7	Distance	−0.048	0.025	−1.95	0.051
find	Explicitness × Distance	0.03	0.035	0.86	0.388

At the infinitival *to* and the subsequent verb (region 6–7), we see no sign of any main effects or interactions involving explicitness, but we again observe faster reading times in the remote conditions (region 6: $\beta = 0.07$, $t = 2.5$, $p < 0.05$, region 7: $\beta = 0.05$, $t = 2.0$, $p = 0.05$). Importantly, we observe no effect of implicitness within the local conditions, in contrast with Experiment 2. That is, in both local and remote conditions, readers are just as fast to process reason clauses whether they follow a short passive or an active target clause.

Discussion

Experiment 3 was designed to better understand two differences between local and remote conditions that were observed in Experiment 2. First, Experiment 2 showed an interaction between distance and explicitness at the reason clause, such that the local conditions showed a cost of implicitness but the remote conditions did not. Because this pattern was predicted by neither the standard theory nor the pragmatic theory, we suspected that it reflected the fact that the passive had not been fully processed by the time the reason clause appeared in the local condition. The results of Experiment 3 are consistent with this hypothesis, as when a temporal modifier was added as a "buffer" between the target clause and the reason clause, no cost of implicitness was observed at the reason clause in the local conditions. This result is thus in keeping with earlier findings concerning local control (Mauner et al., 1995).

Importantly, neither Experiment 2 nor Experiment 3 showed any evidence of one possible pattern of processing differences between local and remote control that might have been predicted by the standard theory. If, as suggested by Mauner et al. (1995), resolving control when the antecedent is implicit requires costly inference when control is pragmatically mediated but not when it is syntactically mediated, then the standard theory would predict a cost of implicitness for remote control and not for local control. This prediction is not borne out by the current results, which show no sign of processing cost for implicitness in the remote control conditions.

In Experiment 2 we also observed overall faster reading times at the reason clause in remote control compared to local control. This pattern would also be consistent with the standard theory if the processes involved in resolving the pragmatic dependency are faster or less effortful than the processes involved in resolving the syntactic dependency. However, this pattern could also have simply been due to the fact that the extra intervening material between the target and reason clause in the remote conditions

(*the reason was*) might have provided more time to fully process the target clause.

The results of Experiment 3 appear to argue against this alternative explanation, because when we control for timing between the local and remote conditions, we continue to observe a strong effect of distance in the reason clause, once again with faster reading times in remote as compared to local reason clauses. However, we also noted another alternative explanation for the facilitated processing of remote control observed here, which is that the content of the intervening material in *the reason was* provided a predictive semantic cue for the upcoming infinitival reason clause. The temporal modifier included in the local conditions in Experiment 3 provided additional processing time, but did not include this kind of semantic cue. In support of this explanation, in Experiment 3 significantly longer reading times were also observed for local relative to remote conditions in the region immediately prior to the reason clause. This early effect cannot be driven by control *per se*, but could be explained if the predictability of *the reason was* sped up reaction times in the remote condition relative to the less predictable temporal modifier (e.g., 3 weeks ago) in the local condition. Experiment 4 was designed to address this remaining discrepancy by making the material immediately preceding the reason clause equally predictable across conditions.

Experiment 4

Experiment 4 used the same design as Experiments 2 and 3 but used the phrase *just in order to* in the local conditions such that both local and remote conditions contained a semantic cue that could be used to predict or prepare for the upcoming reason clause. If the faster reading times in the remote conditions observed in Experiments 2 and 3 were due to the presence of the semantic cue *The reason was*, this difference in processing time should be eliminated in Experiment 4.

Methods

Participants

Participants were 38 native speakers of English from the University of Maryland community, none of whom had taken part in the previous experiments. They received credit in an introductory linguistics course for participating. All participants were naïve to the purpose of the experiment until after participating, when an explanation was provided. The self-paced reading task lasted for approximately 30 min. The task was performed as part of a 1 h session involving an unrelated experiment.

Materials

Twenty-four sets of four target sentences again varied in a 2 × 2 design with the factors *explicitness* and *distance*. However, we included *just in order* in the local conditions in Experiment 4 to match not only the time course, but also the predictiveness of the upcoming reason clause in remote and local conditions. The same fillers and comprehension questions were used as in the earlier experiments. An example set of materials is provided in **Table 7**.

Procedure

The procedure for Experiment 4 was the same as that described above.

Data analysis

The minimum comprehension question accuracy required for inclusion of a participant's data in the analysis was 80%. Data from one participant were excluded due to comprehension question inaccuracy. Data from one participant who was not a native speaker of English were also excluded.

Statistical analysis was the same as described above for Experiment 3. Reading times above 2000 ms were excluded, resulting in a loss of 0.28% of the total data.

Results

Mean comprehension question accuracy for experimental stimuli across participants and items was 93% (range 92–95%), suggesting that participants were successful in comprehending the main experimental stimuli.

Logged reading times are plotted in **Figure 4**, with significant effects summarized in **Table 8**. In the regions preceding the reason clause (1–5) we observed no significant differences except for a main effect of distance due to slower reading times in the remote conditions in region 3 ($\beta = 0.1$, $t = 3.1$, $p < 0.01$), which corresponded to the sentence-initial determiner *the* in remote conditions and *just* in the same region in local conditions. This difference may be associated with the presence of the sentence boundary in the remote conditions.

No other regions showed a significant effect of distance. In particular, at the infinitival *to* and the verb in the reason clause (regions 6–7), we again observed no sign of any main effects or interactions between distance and implicitness. Reading times were numerically faster in remote vs. local conditions in regions

TABLE 7 | Experiment 4 materials.

	Regions of the analysis										
	Pre-target	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Post-target**
ex. loc:	The committee interviewed	the	candidates	just	in	order	to	find	the	best	person for the job.
im. loc:	The candidates	were	interviewed	just	in	order	to	find	the	best	person for the job.
ex. rem:	The committee interviewed	the	candidates.	The	reason	was	to	find	the	best	person for the job.
im. rem:	The candidates	were	interviewed.	The	reason	was	to	find	the	best	person for the job.

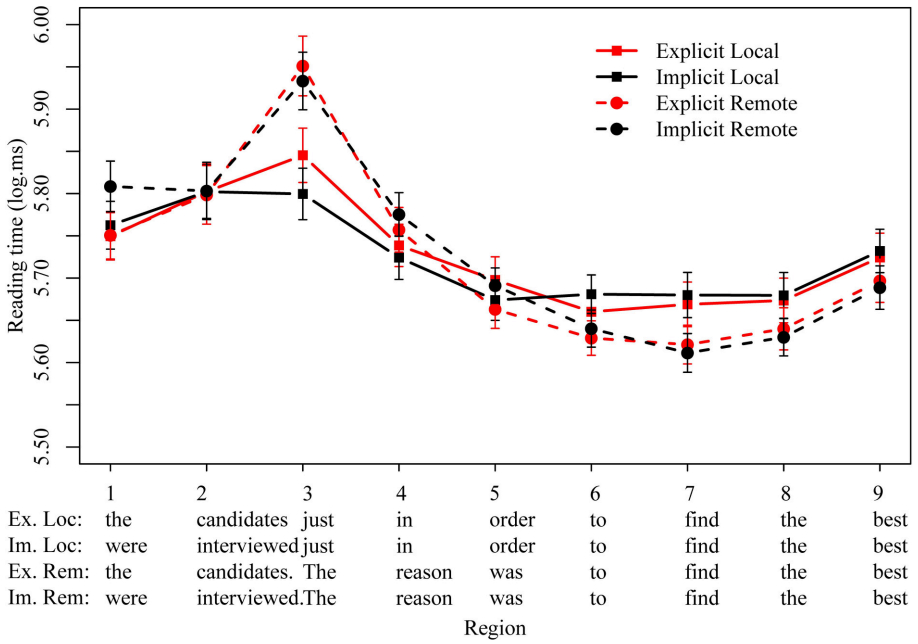


FIGURE 4 | Mean logged reading times (ms) for all four conditions in regions of interest in Experiment 4.

TABLE 8 | Experiment 4 results.

Region	Factors	Estimate	S.E.	*t*-value	*p*-value
	Explicitness	−0.053	0.034	−1.57	0.117
3	Distance	0.105	0.034	3.12	0.002
just/The	Explicitness × Distance	0.027	0.048	0.58	0.56
	Explicitness	0.011	0.026	0.42	0.674
7	Distance	−0.048	0.026	−1.81	0.07
find	Explicitness × Distance	−0.021	0.037	−0.56	0.573

6–9 (*to find the best*). However, in Experiment 4, these differences were small and unreliable, eliciting only a marginal effect of distance in region 7 ($\beta = 0.04$, $t = 1.8$, $p \leq 0.07$; all other regions $p = 0.2$)[7].

Discussion

Experiment 4 examined whether processing differences would be observed in the resolution of reason clauses in local and remote control configurations when both local and remote conditions included a semantic cue for the upcoming reason clause. As the remote conditions necessarily include such a cue in the phrase *the reason was*, in Experiment 4 we included the phrase *just in order* in the local conditions to balance both the distance between the target and the reason clause as well as the presence of a semantic cue for the upcoming reason clause across conditions. Controlling for both timing and predictiveness in this way, we observe no reliable differences in reading times in the reason clause as an effect of whether its target is local or remote, although we continue to observe a trend in this direction. This suggests that much of the "remote speed-up" effect that was observed in Experiments 2 and 3 was due to differences in the extent to which the upcoming reason clause was cued by the prior context, rather than differences between the local and remote configurations in the difficulty of resolving the reason clause.

Because the standard theory proposes different mechanisms for resolving local and remote control (binding vs. contextual interpretation), it predicts differences in the processing of these two configurations; but the pragmatic theory, which proposes the same mechanism (contextual interpretation), does not. In Experiment 4 we observe no such differences in the effect of explicitness on processing of remote vs. local control as reflected in reading times, and no effect of distance in the reason clause. Therefore, to the extent that differences in representation should be reflected as differences in constructing such a representation in online processing, the absence of such differences in Experiment 4 raise a potential challenge for the standard theorist. As we discuss in more detail below, several responses are possible on behalf of the standard theory; it could be that reading times in particular are not a sensitive enough measure to detect differences between local and remote control, or that processing cost more generally does not index whether a dependency is semantically or pragmatically mediated. To the extent that either of these responses are adopted, however, they undermine some earlier arguments in favor of the standard theory.

[7]In Experiment 3, we observed a slow-down in reading times at the first word in the second sentence ("The") in remote conditions, while the same effect did not obtain in Experiment 4. Hence, as a reviewer pointed out, it is precisely when we do not see sentence-boundary effects that we find no advantage for remote control, as in Experiment 4. Whether this is due to a relationship between sentence-boundary effects and implicit control, or some incidental difference between the two experiments, we leave for future investigation.

GENERAL DISCUSSION

In four self-paced reading time experiments, we examined the processing of infinitival reason clauses, in contexts that favor anaphoric construals of PRO, their understood subject. In our materials, the likely referent for PRO is always the individual who satisfies the deep-S role for the preceding target clause; for example the interviewer when the target clause has *interview* as its verb. We compared explicit control, where this role is linked to an audible noun phrase, to implicit control, where it is not. In implicit conditions, the target clause is a short passive. Already Mauner et al. (1995) made this comparison in the local configuration, where the reason and target clauses are syntactically dependent. Ours is the first study to do this for the remote configuration as well, where the two are in separate sentences. What we found, in summary, is this. First, reading times at the reason clause were not longer when the antecedent was implicit relative to when it was explicit, once we controlled the length and content of what intervenes between target and reason clauses across conditions, as in Experiment 4. For local configurations, this agrees with the findings of Mauner and colleagues. That study too found no significant differences in relevant regions between implicit and explicit control, on measures that did distinguish both from cases where, offline, control is judged unacceptable. Our new finding is that no such differences are observed in the inter-sentential remote configuration either, where one might have thought that a costly pragmatic inferencing operation would be required. Second, we also did not observe significant main effects of distance between local and remote control when both the length and the content of intervening material were matched.

Our results bear on a question of grammatical representation. When the understood referent of PRO in a reason clause is an individual mentioned or implied by the target clause, what sort of relation does PRO have to that clause? Is it syntactically linked to an argument there? Or is this a kind of discourse anaphora, with PRO here ranging over a specially restricted domain? On the standard theory (Roeper, 1987), the same relation underlies both explicit and implicit control. This much is consistent with our results, and with those in Mauner et al. (1995), none of which show any relevant effects of the difference. However, the standard theory also takes the common relation to be syntactic, a binding relation between PRO and an argument in the target clause. Such a syntactic link is possible in the local configuration, since the reason clause is adjoined to the target clause. But it is not possible in the remote configuration, since the two clauses are independent. Therefore, if reason control is syntactic when local, as the standard theory says, it must have a different analysis when remote; and if it has the same analysis either way, it cannot be syntactic, and must in both cases be mediated by discourse. Thus, given the standard theory of reason control, we expect a main effect of distance, local vs. remote, on some online measure, while on a uniformly pragmatic theory we do not. On our reading-time measure we found no such effect, not once we controlled for both timing and predictiveness across conditions, as in our Experiment 4. Thus, our results fail to confirm the standard theory.

More importantly, the current results subvert the earlier argument for the standard theory from processing measures. In past work, both the self-paced-reading-time and the stop-making-sense task showed no relevant difference between implicit and explicit control in local configurations (Mauner et al., 1995), while processing costs were observed in baseline conditions (intransitives and middles) in which control of reason clauses appears unacceptable. These data were taken as evidence that both implicit and explicit control were syntactic dependencies. We agree that a similar processing profile may suggest that these are dependencies of the same sort. But the current work illustrates that these prior data cannot be taken to argue that both are syntactic dependencies, since remote control cannot be syntactic, and there too our measures do not distinguish implicit from explicit control.

While these results thus remove a previous argument in favor of the standard theory, they challenge the standard theory directly only if we think that self-paced reading times are sensitive to the difference between syntactic vs. pragmatic anaphora. But as discussed in Section Background, they may not be. Indeed, perhaps these two routes to interpretation are not reliably distinguished by processing cost (see Cunnings et al., 2014, for discussion), or any existing measure of processing. In the latter case there could be no processing evidence for the analysis of control. However, either observation weakens not only our own conclusions, but also the earlier defense of the standard theory. That defense was primarily based on behavioral processing measures (stop-making-sense task and reading times) that were not independently demonstrated to distinguish binding from free anaphora. Hence, our results either provide direct evidence against the standard theory, or undermine earlier arguments in its favor, depending on the evidentiary status of reading times.

Our experiments also highlight the importance of several design factors. Experiment 1 suggested that narrow focus in target clause can increase reading times in the reason clause, a construction whose semantics is sensitive to focus (Dretske, 1972). We believe this makes the long-passive a poor baseline for comparison with implicit control, since normally the *by*-phrase carries narrow focus. Experiments 2 and 3 illustrated the importance of controlling both the length and content of what comes between the target and reason clauses. Reading times for the infinitival verb in the reason clause are slower when it immediately follows the target clause than when it is separated from the target clause by a buffer, either a temporal adjunct in local configurations, or *the reason was* in remote configurations. This may reflect the time it takes to process the passive target clause (Chow et al., 2015). Even with a temporal buffer, reading times were still faster at the infinitive when the buffer is predictive of a reason clause (*the reason was*) than when it is not (*3 days ago*). Yet reading times at the infinitive did not differ significantly between remote and local control in Experiment 4, where we matched the buffers for both length and content, pairing *the reason was* in remote conditions with *just in order* in local conditions.

To finish, let us turn briefly to the issue of implicit arguments. Our results undermine earlier arguments in favor of the standard theory. Although they do not prove an alternative pragmatic

theory correct, they suggest that further investigation of this kind of account would be worthwhile. As we note in the introduction, in a pragmatic theory many of the constraints on control of reason clauses could be captured by a domain condition that is not syntactic but conceptual. Anaphoric uses of PRO in a reason clause denote the individual(s) viewed as responsible for the fact that the reason clause is meant to explain (Farkas, 1988; Landau, 2000; Williams, 2015). This condition is manifest in cases like (10), where PRO finds its antecedent in the surface subject of a passive: the referent of PRO must be viewed as responsible for what happened (Williams, 1974; Zubizaretta, 1982; Roeper, 1987). To be adequate, such a theory would need to say, for example, that the referent of a direct object is never viewed as responsible for the fact expressed by its clause. If it does, a silent argument in passives would play no role in explaining implicit control. Having weakened some of the previous motivation for the standard theory, we suggest that future research ought to explore such pragmatic alternatives in greater detail.

AUTHOR CONTRIBUTIONS

AW and MM designed the experiments. MM and JG implemented the experiments. JG and EL analyzed the data. AW, EL, and MM wrote the paper. All four authors participated in editing and revising the manuscript.

ACKNOWLEDGMENTS

This project was supported in part by NSF IGERT grant #0801465. We wish to thank Aleksandra Fazlipour for her assistance with data collection and preliminary statistical analysis. We also wish to thank many people who have offered helpful comments on earlier versions of this paper. In particular, we thank members of the Cognitive and Neuroscience of Language (CNL) lab at the University of Maryland, participants at the 27th and 28th CUNY Human Sentence Processing conferences and at XPRAG 2015, and audiences at Rutgers University.

REFERENCES

Baayen, R. H., Davidson, D. J., and Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *J. Mem. Lang.* 59, 390–412. doi: 10.1016/j.jml.2007.12.005

Bach, E. (1982). "Purpose clauses and control," in *The Nature of Syntactic Representation*, eds P. Jacobson and G. K. Pullum (Dordrecht: D. Reidel), 35–37.

Baker, M., Johnson, K., and Roberts, I. (1989). Passive arguments raised. *Linguist. Inq.* 20, 219–251.

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: keep it maximal. *J. Mem. Lang.* 68, 255–278. doi: 10.1016/j.jml.2012.11.001

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2014). *_lme4: Linear Mixed-effects Models using Eigen and S4_. R Package Version 1.1-7*. Available online at: http://CRAN.R-project.org/package=lme4

Chow, W.-Y., Smith, C., Lau, E., and Phillips, C. (2015). A 'bag-of-arguments' mechanism for initial verb predictions. *Lang. Cogn. Neurosci.* 30. doi: 10.1080/23273798.2015.1066832

Condoravdi, C., and Gawron, J.-M. (1996). "The context-dependency of implicit arguments," in *Quantifiers, Deduction, and Context*, eds M. Kanazawa, C. Piñón, and H. de Swart (Stanford, CA: CSLI), 1–32.

Culicover, P., and Jackendoff, R. (2005). *Simpler Syntax*. Oxford: Oxford University Press.

Cunnings, I., Patterson, C., and Felser, C. (2014). Variable binding and coreference in sentence comprehension: evidence from eye movements. *J. Mem. Lang.* 7, 39–56. doi: 10.1016/j.jml.2013.10.001

Dalrymple, M., Sheiber, S., and Pereira, F. C. N. (1991). Ellipsis and higher-order unification. *Linguist. Philos.* 14, 399–452. doi: 10.1007/BF00630923

Dretske, F. (1972). Contrastive statements. *Philos. Rev.* 81, 411–437. doi: 10.2307/2183886

Faraci, R. (1974). *Aspects of the Grammar of Infinitives and For-Phrases*. Doctoral dissertation, Massachusetts Institute of Technology.

Farkas, D. F. (1988). On obligatory control. *Linguist. Philos.* 11, 27–58. doi: 10.1007/BF00635756

Fiengo, R., and May, R. (1994). *Indices and Identity*. Cambridge, MA: MIT Press.

Fillmore, C. (1986). Pragmatically controlled zero anaphora. *Berkeley Linguist. Soc.* 12, 95–107.

Frazier, L., and Clifton, C. Jr. (2000). On bound variable interpretations: the LF-only hypothesis. *J. Psycholinguist. Res.* 29, 125–140. doi: 10.1023/A:1005136826534

Frazier, L., and Clifton, C. E. (2001). Processing coordinates and ellipsis: copy α. *Syntax* 4, 1. doi: 10.1111/1467-9612.00034

Frazier, L., and Clifton, C. (2005). The syntax-discourse divide: processing ellipsis. *Syntax* 8, 121–174. doi: 10.1111/j.1467-9612.2005.00077.x

Ginzburg, J., and Sag, I. (2000). *Interrogative Investigations: the Form, Meaning and Use of English Interrogatives*. Stanford, CA: CSLI Publications.

Grimshaw, J. (1990). *Argument Structure*. Cambridge: MIT Press.

Hardt, D. (1993). *Verb Phrase Ellipsis: Form Meaning, and Processing*. Ph.D. dissertation, University of Pennsylvania.

Higgins, F. R. (1973). *The Pseudo-Cleft Construction in English*. Unpublished doctoral dissertation, Massachusetts Institute of Technology.

Iwata, S. (1999). On the status of an implicit argument in middles. *J. Linguist.* 35, 527–553. doi: 10.1017/S0022226799007914

Jones, C. F. (1985). *Syntax and Thematics of Infinitival Adjuncts*. Doctoral dissertation, Amherst, MA, University of Massachusetts.

Kertz, L. (2010). *Ellipsis Reconsidered*. Ph.D. dissertation, University of California, San Diego.

Keyser, S. J., and Roeper, T. (1984). On the middle and ergative construction in english. *Linguist. Inq.* 15, 381–416.

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (2014). *Lmertest: Tests in Linear Mixed Effects Models. R Package Version 2.0-20*. Available online at: http://CRAN.R-project.org/package=lmerTest

Landau, I. (2000). *Elements of Control: Structure and Meaning in Infinitival Constructions*. Berlin: Springer.

Lasnik, H. (1988). Subjects and the theta criterion. *Nat. Lang. Linguist. Theory* 6, 1–18. doi: 10.1007/BF01791588

Lowder, M. W., and Gordon, P. C. (2015). Focus takes time: structural effects on reading. *Psychon. Bull. Rev.* doi: 10.3758/s13423-015-0843-2. [Epub ahead of print].

Manning, C., and Sag, I. (1998). Argument structure, valence, and binding. *Nord. J. Linguist.* 21, 107–144.

Martin, A., and McElree, B. (2008). A content-addressable pointer mechanism underlies comprehension of verb-phrase ellipsis. *J. Mem. Lang.* 58, 879–906. doi: 10.1016/j.jml.2007.06.010

Mauner, G., and Koenig, J.-P. (2000). Linguistic vs. conceptual sources of implicit agents in sentence comprehension. *J. Mem. Lang.* 43, 110–134. doi: 10.1006/jmla.1999.2703

Mauner, G., Tanenhaus, M. K., and Carlson, G. N. (1995). Implicit arguments in sentence processing. *J. Mem. Lang.* 34, 357–382. doi: 10.1006/jmla.1995.1016

Merchant, J. (2001). *The Syntax of Silence*. Oxford: Oxford University Press.

Mikkelsen, L. (2011). "Copular clauses," in *Semantics: An International Handbook of Natural Language Meaning,* Vol. 2, eds C. Maienborn, K. von Heusinger, and P. Portner (Berlin: Mouton de Gruyter), 1805–1829.

Partee, B. (1989). "Binding implicit variables in quantified contexts," in *Papers from the 25th Regional Meeting of the Chicago Linguistic Society*, eds C. Wiltshire, R. Grazyte, and B. Music (Chicago, IL: Linguistic Society), 342–356.

Phillips, C., and Parker, D. (2014). The psycholinguistics of ellipsis. *Lingua* 151, 78–95. doi: 10.1016/j.lingua.0.2013.10.003

R Core Team. (2014). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available online at: http://www.R-project.org/

Roeper, T. (1987). Implicit arguments and the head-complement relation. *Linguist. Inq.* 18, 267–310.

Sag, I. (1976). *Deletion and Logical Form*. Doctoral dissertation, Massachusetts Institute of Technology.

Sag, I., and Pollard, C. (1991). An integrated theory of complement control. *Lang. (Baltim.)* 67, 63–113. doi: 10.2307/415539

Shapiro, L., and Hestvik, A. (1995). On-line comprehension of VP-ellipsis: syntactic reconstruction and semantic influence. *J. Psycholinguist. Res.* 24, 517–532.

Stanley, J. (2000). Context and logical form. *Linguist. Philos.* 23, 391–434. doi: 10.1023/A:1005599312747

Stanley, J. (2002). "Nominal restriction," in *Logical Form and Language*, eds G. Preyer and G. Peter (Oxford: Clarendon Press), 365–388.

Stowell, T. (1981). *Origins of Phrase Structure*. Ph.D. dissertation, Massachusetts Institute of Technology.

Tanenhaus, M. K., and Carlson, G. N. (1990). Comprehension of deep and surface verb phrase anaphors. *Lang. Cogn. Process.* 5, 257–352. doi: 10.1080/01690969008407064

van Valin, R. D. (1990). Semantic parameters of split intransitivity. *Language* 66, 221–260. doi: 10.2307/414886

Vinet, M.-T. (1988). "Implicit arguments and control in middles and passives," in *Advances in Romance linguistics*, eds D. Birsong and J. P. Montreuil (Dordrecht: Foris), 427–437.

Whelpton, M. (2002). Locality and control with infinitives of result. *Nat. Lang. Seman.* 10, 167–210. doi: 10.1023/A:1022107429786

Williams, A. (2015). *Arguments in Syntax and Semantics.* Cambridge: Cambridge University Press.

Williams, E. (1974). *Rule Ordering In Syntax*. Doctoral dissertation, Massachusetts Institute of Technology.

Williams, E. (1977). Discourse and Logical Form. *Linguist. Inq.* 8, 101–139.

Williams, E. (1985). PRO and subject of NP. *Nat. Lang. Linguist. Theory* 3, 297–315. doi: 10.1007/BF00154265

Yoshida, M., Dickey, M., and Sturt, P. (2012). Predictive processing of syntactic structure: sluicing and ellipsis in real-time sentence processing. *Lang. Cogn. Process.* 28, 272–302. doi: 10.1080/01690965.2011.622905

Zubizaretta, M. L. (1982). *On the Relationship of the Lexicon to Syntax*. Doctoral dissertation, Massachusetts Institute of Technology.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2015 McCourt, Green, Lau and Williams. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in
PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 07 October 2013
doi: 10.3389/fpsyg.2013.00708

Dependency-dependent interference: NPI interference, agreement attraction, and global pragmatic inferences

Ming Xiang*, Julian Grove and Anastasia Giannakidou

Linguistics Department, University of Chicago, Chicago, IL, USA

Edited by:
Colin Phillips, University of Maryland, USA

Reviewed by:
Ian Cunnings, University of Reading, UK
Edward M. Husband, University of Oxford, UK

****Correspondence:***
Ming Xiang, Language Processing Lab, Linguistics Department, University of Chicago, 1010 E. 59th Street, #304, Chicago, 60615 IL, USA
e-mail: mxiang@uchicago.edu

Previous psycholinguistics studies have shown that when forming a long distance dependency in online processing, the parser sometimes accepts a sentence even though the required grammatical constraints are only partially met. A mechanistic account of how such errors arise sheds light on both the underlying linguistic representations involved and the processing mechanisms that put such representations together. In the current study, we contrast the negative polarity items (NPI) interference effect, as shown by the acceptance of an ungrammatical sentence like "*The bills that democratic senators have voted for will* ***ever*** *become law*," with the well-known phenomenon of agreement attraction ("*The key to the cabinets are . . .* "). On the surface, these two types of errors look alike and thereby can be explained as being driven by the same source: similarity based memory interference. However, we argue that the linguistic representations involved in NPI licensing are substantially different from those of subject-verb agreement, and therefore the interference effects in each domain potentially arise from distinct sources. In particular, we show that NPI interference at least partially arises from pragmatic inferences. In a self-paced reading study with an acceptability judgment task, we showed NPI interference was modulated by participants' general pragmatic communicative skills, as quantified by the Autism-Spectrum Quotient (AQ, Baron-Cohen et al., 2001), especially in offline tasks. Participants with more autistic traits were actually less prone to the NPI interference effect than those with fewer autistic traits. This result contrasted with agreement attraction conditions, which were not influenced by individual pragmatic skill differences. We also show that different NPI licensors seem to have distinct interference profiles. We discuss two kinds of interference effects for NPI licensing: memory-retrieval based and pragmatically triggered.

Keywords: memory interference, pragmatic inference, individual differences, autistic traits, NPI licensing

INTRODUCTION

During the processing of long distance dependencies, sometimes an element in a sentence that should be irrelevant for constructing a dependency interferes—a phenomenon that has been dubbed the "interference effect." For instance, agreement attraction errors, such as **the key to the cabinets are . . .*, involve an agreement dependency between the singular subject *the key* and the plural copula verb *are* which is ungrammatical because of a number mismatch. However, the intervening noun "*cabinets*" interferes, facilitating the processing of the ungrammatical sentence[1]. Facilitation effects from an interfering element have been shown by various processing measures: such sentences are relatively common in spontaneous production; they can be elicited in controlled laboratory experiments; they are judged to be relatively acceptable; and online reading times on the otherwise problematic verb are generally reduced compared to number mismatched verbs without interference (Bock and Miller, 1991; Bock and Eberhard, 1993; Pearlmutter et al., 1999; Eberhard et al., 2005; Wagers et al., 2009; Dillon et al., 2013).

Such interference effects have been explained as instances of memory interference triggered during cue based memory retrieval. During incremental parsing of a long distance dependency, the tail of the dependency initiates the retrieval of the head in memory. This retrieval is prone to interference when the intermediately preceding material shares certain features with the set of retrieval cues that the parser employs (McElree et al., 2003; Lewis and Vasishth, 2005; Lewis et al., 2006; Wagers et al., 2009). Memory interference can be driven by partially matched morpho-syntactic features, as has been repeatedly shown by agreement attraction errors like the example above, and examples like "*The new executive who oversaw the middle managers were dishonest*" (example from Dillon et al., 2013). In such cases, memory retrieval is initiated in order to search for a plural subject at the plural verb, e.g., "were." Because the search mechanism is content addressable (McElree et al., 2003), it may target any item in memory during the search process, leading to erroneous acceptance of interfering material which bears feature similarity to the correct retrieval target.

[1]Cue similarity does not always lead to facilitation effect. For instance, Van Dyke and Lewis (2003) and Van Dyke and McElree (2006) have shown that cue overload leads to increased processing difficulty. In this paper, we only focus on the facilitatory interference effect.

A large body of the research on interference effects has focused on deriving a thorough mechanistic account of the errors people make in interference situations, as such an account helps in constructing a precise parsing algorithm for long distance dependencies. But the pursuit of a domain-general parsing mechanism has somewhat overshadowed the question of whether or not different kinds of linguistic dependencies should be handled by the same parsing algorithm, and hence whether or not interference arises in the same way across different dependency types. One reasonable hypothesis is that the precise nature of a particular type of linguistic dependency is relevant to explaining differences in how dependencies are processed. In the current study, we tackle the question of whether or not there are dependency-dependent interference effects by looking at a case of interference that seems very much like agreement attraction on the surface, but that may at least partially arise from a different underlying mechanism. The particular type of interference we will discuss appears in the licensing of negative polarity items (NPIs), as in the sentence "**The documentaries that no network TV stations have played during prime time have ever been very controversial*," where the presence of *ever* is illicit. We argue that although such interference may superficially look the same as the subject-verb agreement errors discussed above, there are actually multiple different sources that contribute to NPI interference. In particular, in addition to a memory-retrieval based interference that is similar to agreement interference, there is also a separate rout of pragmatic inferences made at the message level during semantic integration. A sufficient account of NPI interference needs to take into account the close interaction between grammar (e.g., syntax and semantics) and pragmatic inference in sentence processing.

NPI LICENSING AND INTERFERENCE

NPIs are lexical items that need to be licensed in an environment that possesses a particular logical-semantic property. Negation is a cross-linguistically attested licensor for NPIs (as noticed in Klima, 1964). Licensing typically requires the NPI to be in the semantic and syntactic scope (i.e., c-command domain) of negation (Ladusaw, 1979, 1980; Giannakidou, 1998, 2011 for an overview). As shown in (1), the NPIs *any* and *ever* are grammatically licensed when they appear within the scope of negation (1a, b), but they are ungrammatical when there is no negation present (1c, d), or when negation is present, but doesn't c-command the NPI (1e, f).

(1) a. John *didn't* talk *to anybody*.
b. John *hasn't ever* talked to Bill.
*c. John has *ever* talked to Bill.
*d. John talked to *anybody*.
*e. *Anybody didn't* talk to John.
*f. The debate that *nobody* cared about will *ever* end.

Because of their apparent sensitivity to the presence of negation, *any* and *ever* are labeled "negative" polarity items (NPIs)[2], but it must be noted that their distribution, and that of similar NPIs crosslinguistically, is quite broad and includes a vast range of negative and non-negative licensors, including conditionals, modal verbs, generic sentences, questions, the scope of universal quantifiers, comparatives, disjunctions (see Giannakidou, 2011 for detailed overview). Given this broad distribution and the potential differences among NPI classes in English and crosslinguistically, what semantic property unifies licensors as a natural class has been a matter of intense study—and researchers generally agree that NPIs appear in non-veridical environments. Non-veridical environments are (a) negative environments with negation and negative quantifiers (Baker, 1970a; Linebarger, 1980), (b) downward entailing environments (Ladusaw, 1980; Zwarts, 1986, 1996; Hoeksema, 1994; von Fintel, 1999, inter alia), and (c) other non-veridical environments that may not be negative or downward entailing (e.g., modal expressions, questions, imperatives, generic statements; Zwarts, 1995; Giannakidou, 1998, 2006, 2011; Bernardi, 2002). We cannot provide a detailed survey here; but as background for the specific data we address, we discuss licensing by negation and downward entailment (DE) in the next section[3].

In the domain of NPI licensing, an "interference effect" is said to result when an unlicensed NPI becomes more acceptable if a licensor is inserted into the preceding context—but crucially, is *not* in the right structural (c-commanding) position (Drenhaus et al., 2005; Vasishth et al., 2005, 2008). In the example below (examples taken from Drenhaus et al., 2005), the expected contrast obtains between (2a) and (2b); however, there is also a significant difference between the ungrammatical sentences (2b) and (2c). (2c) is judged as "more acceptable" than (2b), even though the licensor *no* doesn't c-command the NPI *ever*, so that it should be unlicensed.

(2) a. **No** man who had a beard was *ever* happy.
*b. A man who had a beard was *ever* happy.
*c. A man who had **no** beard was *ever* happy.

In the (c) example, negation is present but not in a position c-commanding *ever,* as is required for licensing. The NPI *ever* therefore remains unlicensed. In online measures, NPI interference effects have appeared as facilitatory effects (e.g., shorter RTs or smaller ERP amplitudes) on the problematic NPI in the interference condition, as compared to the NPI in the condition with no licensors anywhere in the sentence (Drenhaus et al., 2005; Vasishth et al., 2008; Xiang et al., 2009; Parker et al., 2013).

The interference effect above is on the surface very similar to the memory interference phenomenon introduced earlier

[2]Some NPIs, such as any, seem to obtain so-called *free choice* readings in modal environments and with imperatives, such as *You may talk to any student*, and *Pick any card*!. We won't discuss the free choice use in this paper. The particular NPI *ever* studied in our experiment, does not have a free choice use and it is typically blocked in modal contexts: * *You may ever go to Paris.*

[3]A scalar component has also been posited for some NPIs (Kadmon and Landman, 1993; Krifka, 1995; Lahiri, 1998; Chierchia, 2006), but scalarity doesn't characterize *all* NPIs as a class. There are many non-scalar NPIs (see Giannakidou, 1998, 2011, for a recent overview (Lin, 1996; Giannakidou and Yoon, 2012). And scalar NPIs such as *any* and *ever* do not have only scalar uses (Duffley and Larivée, 2010), and are not morphologically marked as scalar either, i.e., they do not contain scalar markers such as *even* and the like. Whether scalarity plays a role in the interference effect is an open question for future research.

for subject-verb agreement. One account of NPI interference is indeed couched upon retrieval interference due to feature similarity between the retrieval cues and the previously processed linguistic information. Vasishth et al. (2008) argued that the parser uses lexical semantic cues such as [+negative] and syntactic cues such as [+c-command] to retrieve a proper licensor for *ever* from previously processed material in memory. For (2b), no such match is found, and the sentence is determined to be unacceptable. For (2c), however, the quantifier *no* in the embedded subject position partially matches the search criteria: although it doesn't match the syntactic cue [+c-command], it does satisfy the cue [+negative]. During retrieval of a licensor, this partial feature match may boost the activation level of the memory representation of the embedded quantifier *no*, causing it to be more likely to be retrieved once its activation level goes beyond a certain threshold.

Although such an account is plausible, as well as parsimonious, we think it falls short of providing a complete account of NPI interference, because it misses some important distinctions between NPI licensing, on the one hand, and syntactic dependencies such as subject-verb agreement and those involved in relative clauses and cleft constructions, on the other hand. Specifically, while the latter dependencies types involve syntactic relations between lexical items (e.g., a subject and an agreeing verb, or a head noun and a verb in a relative clause), NPI licensing involves not only syntactic conditions (e.g., the c-command requirement on a proper licensor, but also see our remarks in the general discussion about this syntactic condition), but also logical-semantic (e.g., negation, DE, non-veridicality), and pragmatic conditions. Crucially different from subject-verb agreement, pragmatic inferences derived from global semantic interpretation (which traditionally have been considered outside of the grammar proper), can be used to license NPIs (Linebarger, 1980, 1987; Giannakidou, 1998, 2006). We will discuss the pragmatic licensing mechanism in more detail in the next section.

Closely related to the fact that NPI licensing involves multiple mechanisms, there are many different types of licensors other than just the negative determiner *no*, which has been the focus of most of the studies on NPI interference. Interference under the licensor *no* may look superficially similar to subject-verb agreement, because one can identify a [+negative] feature on the licensor, which, when served as a memory retrieval cue, may lead to interference. Whether or not this is indeed the underlying mechanism, or only one of the mechanisms involved, is an empirical question we will address in this paper, but a cue-driven process is at least a logical possibility here. Importantly, when we look at a larger set of licensors, postulating a lexical [+negative] feature becomes untenable for many of them: for instance, with a universal quantifier *every*, focus *only*, conditional *if*, emotive factives like *surprised*, *amazed*, etc. We focus on *only* here, since it can be used as a determiner and therefore constitutes a minimal pair with *no*. We assume that *only* licenses NPIs through a negative exceptive component, since a sentence of the form "[Only NP] VP" entails "[Nobody other than NP] VP" (see the discussion in the section below). But there is little reason to believe that *only* itself contains in its lexical entry a grammatical/syntactic [+negative] feature. Klima (1964) gave syntactic diagnostics for syntactically negative expressions, which include phonologically and morphologically negative expression such as *no*, *none*, *never*, but also negative expressions that are not overtly marked in morphology or phonology as negative, such as *few*, *scarcely*, *hardly*, *seldom*, *rarely*, etc. For example, all of these expressions can be followed by a conjunct with a *neither*-tag, but not by *so*-tag; they may also co-occur in a conjunct with *either*, but not with *too*; etc. We provide some examples below, showing that *only* is not a negative expression under these syntactic diagnostics. Nor are the other non-negative licensors we mentioned above, as the reader may verify.

(3) a. Publishers will usually reject suggestions, and ***no/few/*only/*a few*** writers will accept them, ***either***.
b. Publishers will ***not/hardly/seldom/rarely/*only/*usually*** accept suggestions, and ***neither*** will the writers.

In the absence of a lexically coded [+negative] feature that can trigger a similarity-based interference effect (as with *only*), the question arises whether or not we will still see interference, and, if we do see such an effect, what would account for it. We will address these questions in the current study by examining both the licensors *only* and *no*.

FLEXIBILITY OF LICENSING WITH ENGLISH NPIs: THE ROLE OF PRAGMATIC INFERENCING

It has long been observed that some NPIs become licensed even when no grammatical lexical licensors are present on the surface that contain the required logical-semantic property for licensing. The following examples are largely taken from Linebarger (1987):

(4) a. John kept writing novels *long after* he had *any* reason to believe they would sell.
b. *Exactly four* people in the world would have *ever* read that dissertation: Bill, Mary, Tom, and Ed.
c. Mary was *surprised* there was *any* food left.
d. I am *sorry* that I ever met him.
e. *Only* the students who have *ever* read anything about phrenology attended the lectures. [=117 in Ladusaw (1980)]

In all these examples, there aren't any explicit lexical items that can serve as grammatical licensors, in the sense that they possess the logical property necessary for licensing. Surely, the items *long after*, *exactly four*, *amazed*/*surprised*, and *only* are responsible for the appearance of *any* and *ever*, but they are not logically negative, nor DE, nor non-veridical. Consider the property of negation/DE. DE expressions, as is traditionally stated, allow logical inferences in their scope from a set to a subset. Consider the following entailment relations with negation [examples adapted from Linebarger (1987)].

(5) a. John didn't eat a green vegetable for dinner.
b. John didn't eat kale for dinner.

Kale is a subset of green vegetables. If John didn't eat a green vegetable for dinner (5a), it logically follows that John didn't eat

kale for dinner (5b). The superset-to-subset logical inference is the hallmark of negative and DE expressions. Ladusaw proposed that NPIs appear in the scope of negative and other DE expressions (such as negative quantifier *few*, or the restrictor of universal quantifier *every*). However, none of the examples in (4) contains a DE expression. We show below on this point for "*only*" and "*long after*" [see Linebarger (1987); Atlas (1993); Horn (1996); von Fintel (1999) and Giannakidou (2006) for more discussion that *only* and emotive factive verbs such as *sorry* and *surprise* are not DE in its strict sense].

(6) a. Only Bill went to Greece;
b. Only Bill went to Athens.
(7) a. Bill married Samantha long after he travelled to Europe.
b. Bill married Samantha long after he travelled to France.

We see here that the subset inference from (a) to (b) sentences in (6) and (7) is not licensed with the critical expressions "only" and "long after," and they are therefore *not* logically DE[4].

Faced with many examples of this kind, in which a grammatical licensing mechanism relying on the logical properties of a licensor does not seem to suffice, several researchers have advanced proposals to distinguish a pragmatic licensing mechanism from a grammatical (syntactic/semantic) one. Giannakidou (1998, 2006), for example, talks about two modes of licensing, one semantic, relying on a (c-commanding) grammatical licenser ("direct" licensing), and another, "global pragmatic" licensing ("indirect" licensing) that relies on the availability of a negative inference. In the regular case, NPIs are licensed directly by an expression that bears the required logical-semantic property. However, in the absence of such a grammatical licensor, either the use of an NPI leads to ungrammaticality, or the context enables comprehenders to derive a negative inference pragmatically, which in turn licenses NPIs (see also Baker, 1970a,b; Linebarger, 1987). Such pragmatic inferences have been called "implicatures," and we will refer to them as such from now on[5]. Linebarger (1987) and Giannakidou (2006) have considered *only* as a candidate for pragmatic licensing. The basic intuition there is that the exclusive component in the meaning of *only* is responsible for licensing NPIs (e.g., *Only John ate kale* entails that *Nobody other than John ate kale*). In our recent work (Xiang et al., 2012, 2013), based on ERP evidence, we argued that *only* is a semantic licensor that licenses NPIs through negation in the asserted content (see Atlas, 1993; Horn, 1996, for the semantics of *only*). Although they are different in their specific details, none of these proposals treat *only* as a negative expression that contains a lexically coded [+negative] feature, keeping in line with our discussion in the last section.

It is also crucial to note that, although licensing through global pragmatic reasoning is a possible mode of licensing for many NPIs, not all negative implicatures can be used to make NPIs acceptable (Linebarger, 1987; Horn, 1989, 2002; Giannakidou, 2006). For instance, "*almost*"—though clearly inviting a negative inference (*John almost finished the book* implies that *he did not finish it*)—does *not* license NPIs: **John almost finished anything*[6]. Although the boundary between inferences that can and cannot license NPIs is still an open question, we follow the suggestion in Linebarger (1987) and assume that in order for a derived negative inference to be able to license NPIs, it should be prominent in the sense that the derived proposition warrants the truth of the original proposition. Consider our earlier example with *long after*:

(8) a. John kept writing novels *long after* he had *any* reason to believe they would sell.
b. John kept writing novels even though he didn't have any reason to believe they would sell.

The NPI *any* in (8a) is licensed under the derived negative implicature (8b). There is a very strong inference to (8b) from (8a), and in fact the two are almost semantically equivalent. Most important, if (8b) is true, then it is also true that *John kept writing novels long after he had any reason to believe they would sell*. It seems, then that a "useful" negative inference is one that is semantically close enough to the original proposition. How to formally quantify the notion of "semantic closeness" is an open question and is beyond the scope of the current discussion. What is crucial for current purposes is, first that negative implicatures provide a possible licensing mechanism, at least in English; and second, not all negative implicatures can license NPIs. It is possible that the difference between the "useful" and "useless" inferences is a categorical one, but it is also possible that the two simply occupy different ends of a continuum of pragmatic inferences, on which one finds different degrees of "licensing strength." We will leave this question open. We turn below to the empirical focus of the current paper: the case of NPI interference. We will argue that the interference observed in NPI licensing is at least partially driven by the over-application of the pragmatic licensing mechanism. That is, in cases of NPI interference, comprehenders resort to the pragmatic strategy, i.e., they attempt to use a pragmatic inference, which, however, cannot properly license NPIs. The effect is that such an illicit interference will occasionally boost the acceptability of unlicensed NPIs. The availability of such pragmatic inferences, as we will show, is modulated by individual subjects' pragmatic skills.

INTERFERENCE DRIVEN BY PRAGMATIC INFERENCE

Xiang et al. (2009) argued that the NPI interference effect stems from over-application of a flexible, inference-based licensing

[4]There is ongoing discussion on these expressions. For *only* and *emotive factives*, von Fintel (1999) and Gajewski (2005) analyzed them as Strawson DE, rather than regular DE expressions. Limited by space, we won't discuss further about this possibility, but see theoretical and experimental evidence for problems of this approach in Linebarger (1987), Giannakidou (2006) and Xiang et al. (2012, 2013). We also won't go into further discussion about *long after*, since it is not the focus of this paper, but see Condoravdi (2010) that proposed a DE analysis, and also Krifka (2010) for a discussion.

[5]Although we adopt the term "implicature" here, it should be clear from our discussion that we are aware this concept is still very vague. Not all negative implicatures can license NPIs. The exact grammatical constraints and mechanisms that rule in some "implicatures" but rule out others remain as an open question.

[6]See Horn (2002) for the idea and the negative component in "almost" is "assertorically inert," in contrast to "barely," and hence does not license NPIs.

mechanism that is already in place in the grammar. One possibility, as suggested in Xiang et al. (2009), is that, while parsing a statement like "*the bills that no democratic senators have voted for will* P" ("P" stands for an upcoming predicate), people generate a negative inference about a contrasting set of referents "*the bills that democratic senators HAVE voted for will NOT have the same property P*" on some proportion of trials. Note that such an inference is not logically valid, nor can it be derived from any proper grammatical device. But the particular construction involved in NPI interference effect, i.e., relative clauses, may be responsible for triggering such negative inferences. It is known that restrictive modifiers generally invite inferences about a contrastive set of referents pragmatically (Altmann and Steedman, 1988; Tanenhaus et al., 1995; Sedivy et al., 1999). It has been shown that people are very sensitive to the pragmatic cues of restrictive modifiers: restrictive modifiers perform a discourse function to distinguish the set of referents that possess the property described by the modifier and the set that do not. Such discourse principles are active in parsing because interlocutors engaged in a discourse interaction adhere to the general communicative principle that the exchange of information should be as informative as it needs to be (Grice, maxim of quantity, 1975). To our knowledge, almost all studies on NPI interference so far in the literature have used restrictive relative clauses to host an "intruding" licensor. It is plausible then to argue that the choice of this particular structure facilitates the triggering of negative inferences about a contrasting set.

Although pragmatic inferences driven by communicative pressure are very common in natural language communication, they in general are not qualified to actually license NPIs. If we adopt our rudimentary notion of "semantic closeness" in the last section, the negative inferences made in the interference scenarios are not "close" enough to the original propositions. Consider again the interference example "*The bills that no democratic senators have voted for will become law*." The potential negative inference "*The bills that democratic senators have voted for will NOT become law*" does not have similar enough truth-conditions to the original proposition. Not being semantically close, the negative inference is too weak to render NPIs totally acceptable. But since pragmatic inferences may in principle license NPIs in English, comprehenders may overapply this mechanism and use it in some proportion of the ungrammatical trials, so that negative inferences that are normally not useful for NPI licensing have a facilitating effect on acceptability.

If the interference effect with NPIs is due to over-application of pragmatic inferences, in which subjects extract a negative implicature from the given context, we predict that NPI interference effects should be modulated by individual participants' ability to extract pragmatic inferences from context. Different individuals may possess varying abilities to carry out complex pragmatic reasoning, and we hypothesize that participants who are better at pragmatic reasoning will be more prone to an NPI interference effects, since it is more likely for these participants to successfully construct negative inferences from context, making them more vulnerable to over-applying the pragmatic licensing mechanism. On the other hand, participants who are less skilled in pragmatic inference will generate fewer inferences, and these participants will be more likely to avoid the interference effect.

Furthermore, in the current study, we compare NPI interference with a purely syntactic dependency: subject-verb agreement. We predict that, if the correlation between pragmatic skills and interference in NPI licensing is driven by over-application of a pragmatic-licensing mechanism that is specific to NPIs, no similar correlation should hold between the magnitude of the agreement interference (attraction) effect and individual pragmatic differences, despite the superficial similarity between NPI interference and agreement attraction errors. We test these predictions in the current study. Individual pragmatic skills of our participants were assessed and quantified by the autism-spectrum quotient (AQ, Baron-Cohen et al., 2001), which we turn to now.

THE AUTISM-SPECTRUM QUOTIENT

Pragmatic language problems are among some of the defining characteristics of children and adults with autism (Bishop, 1989; Tager-Flusberg et al., 2005). For example, their linguistic behavior may often consist in inappropriate comments; and they may have difficulty comprehending jokes, sarcasm, and indirect requests (Happé, 1993; Ozonoff and Miller, 1996; Wang et al., 2006). However, it is increasingly recognized that autistic traits are likely to be present on a continuum among the general population, and people who are diagnosed as autistic simply represent one end of this continuum. This raises the possibility that even among the neurotypical population, there exist individual pragmatic differences associated with individual autistic traits. The AQ (Baron-Cohen et al., 2001) assesses the extent of autistic traits that neurotypical individuals possess. There are a total of 50 questions, divided into 5 subscales, each with 10 statements, to which the subject must reply with one of the choices: "Definitely agree," "Slightly agree," "Slightly disagree," or "Definitely disagree." The 5 subsets of questions are designed to tap into five different cognitive functions that have been found to be important when characterizing autistic behavior. The five subscales and a corresponding example item are: social skills (e.g., "I prefer to do things with others rather than on my own."); communication (e.g., "Other people frequently tell me that what I've said is impolite, even though I think it is polite."); attention to detail (e.g., "I often notice small sounds when others do not."); imagination (e.g., "If I try to imagine something, I find it very easy to create a picture in my mind."); attention switching (e.g., "I prefer to do things the same way over and over again"). Half of the questions are designed to elicit an answer of "definitely agree" or "slightly agree"; and the other half, "definitely disagree" or "slightly disagree." Baron-Cohen et al. (2001) provide scoring guidelines. Higher scores indicate more association with autistic traits.

There is an increasing number of studies that document the correlation between AQ (or AQ subscale) scores and processing in certain specific linguistic domains among the neurotypical population (Stewart and Ota, 2008; Nieuwland et al., 2010; Yu, 2010). Particularly relevant for current purposes, the communication and social skills subscales have been linked to pragmatic language comprehension; in particular, the processing of scalar implicatures (Nieuwland et al., 2010; Sikos et al., 2013) and

perspective taking (Grodner et al., 2012). For example, Nieuwland et al. (2010) showed that when computing scalar implicatures (e.g., *some* implies *not all*), participants' ability to generate scalar implicatures online was significantly correlated with their communication subscale scores (CS scores). In particular, participants with better communication skills (i.e., lower CS scores) were more likely to access the scalar implicature interpretation of a sentence like "*some elephants have trunks*," and consequently detect the anomaly of under-informativity.

The growing body of work that shows a correlation between AQ scores and pragmatic language skills makes the AQ a suitable tool for the current study to probe the underlying differences between NPI and subject-verb agreement dependencies. Admittedly, such a correlation only provides a classificatory diagnostic, rather than an explanation of the mechanisms underlying pragmatic reasoning, since it is not yet clear how the communicative and social skills measured in the AQ are recruited in language comprehension. Although the exact nature of the link between extra-linguistic skills and linguistic pragmatic reasoning is not well-understood, the link itself is nevertheless supported by empirical evidence, suggesting that the same cognitive mechanisms may be shared between the two types of tasks. Thus, the AQ provides us with a way to operationalize individual differences in pragmatic reasoning.

CURRENT EXPERIMENT

METHOD

Materials

There are two types of target items in this study: NPI and subject-verb agreement. **Table 1** gives an example of each type. For the NPI materials, there are three basic types of conditions. In the ***Licensed*** conditions (9a and 9b), the NPI *ever* is licensed by a grammatical licensor. In the ***Interference*** conditions (9c and 9d), *ever* is not licensed properly: even though there are licensors (*no* and *only* again) in the same context, they are not in a syntactically c-commanding position. Finally, the ***Plain Unlicensed*** conditions contain unlicensed NPIs with no potential licensors in the preceding context.

Table 1 | Example stimuli.

(9) NPI	
a/c Licensed	***No***/*only* documentaries that the network TV stations have played during prime time have ***ever*** been very controversial.
b/d Interference	The documentaries that ***no***/***only*** network TV stations have played during prime time have ***ever*** been very controversial.
e Plain Unlicensed	The documentaries that the network TV stations have played during prime time have ***ever*** been very popular.
(10) AGREEMENT	
a Grammatical	The ***receptionist*** who the ***boss*** depends on never ***fails*** to do a stellar job.
b Interference	The ***receptionist*** who the ***bosses*** depend on never ***fail*** to do a stellar job.
c Plain Ungrammatical	The ***receptionist*** who the ***boss*** depends on never ***fail*** to do a stellar job.

For the Licensed and the Interference conditions, we looked at the two licensors *no* and *only* in this study to test the generality of previously observed interference effects. *Only* is different from *no* in at least two ways: first, as discussed earlier, *only* does not contain a [+negative] feature; second, *only* is much less frequent than *no* as a licensor (Xiang et al., 2009). This raises the question whether or not interference will arise for *only*, and, if so, whether or not the same account should apply to both licensors.

The set of agreement items (10a–c) was created using the same design. In the ***Grammatical*** condition, the main verb agrees with the matrix subject (*the receptionist*) in its singular number. In addition, the embedded subject (*the boss*) is also singular, creating no interference. In the ***Interference*** condition, the matrix verb fails to agree with the matrix subject, since the matrix subject is singular whereas the verb is plural. However, the intervening embedded subject also carries plural number, and hence may be incorrectly accepted as being in an agreement relation with the main verb. Finally in the ***Plain Ungrammatical*** condition, the main verb fails to agree correctly with the singular matrix subject, but the embedded subject is also singular, mismatching the main verb.

There were 60 sets of the NPI items, 40 sets of the agreement items, as well as 38 extra fillers. The items were distributed into multiple lists using a Latin square design, such that no participant was presented with more than one condition from the same item set.

Participants and procedure

Ninty-two native English speakers (mean age = 20, *sd* = 3.2, 52 female, 40 male) from the University of Chicago campus and surrounding area participated in the study for $10 payment or course credit. Each participant finished a self-paced reading task and also completed an AQ questionnaire (see below). The self-paced reading task was presented using the Linger software (Doug Rohde, MIT). Participants read through each sentence word by word at their own pace. After the last word of each sentence, a question appeared that said: "Is this acceptable?" After participants pressed one of the two answer keys (Y or N) on the keyboard, they went on to the next trial. Practice trials were provided before the experimental session to familiarize participants with the task. Each subject also completed an AQ questionnaire either before or after they completed the self-paced reading task (in a random order).

DATA ANALYSIS AND RESULTS

Among the 92 participants, one did not finish the AQ questionnaire, and his data was not included in any of the analyses below. Three additional participants were excluded from the analysis due to very low overall accuracy across the whole experimental session (<50% correct). For the rest of the 88 participants, we analyzed their acceptance rate results and their online reading times at the critical word *ever*. The grand average results are presented in **Table 2**.

For the data analysis, we will present results from mixed effects logistic regression models on the acceptance rate data and results

Table 2 | Average acceptance rate and RTs on the critical word, presented separately for the NPI and the agreement items (with *sd* in the parenthesis).

	NPI Licensing				
	***No*-Licensed**	***No*-interference**	***Only*-Licensed**	***Only*-interference**	**Plain unlicensed**
Acceptance rate	0.87 (0.16)	0.25 (0.30)	0.81 (0.17)	0.27 (0.32)	0.16 (0.27)
RT at the CW (ms)	421 (114)	439 (120)	420 (120)	453 (146)	473 (155)
	Agreement				
	Grammatical	**Interference**	**Plain ungrammatical**		
Acceptance rate	0.92 (0.18)	0.28 (0.23)	0.12 (0.17)		
RT at the CW (ms)	463 (125)	509 (169)	547 (202)		

from mixed effects linear regression models on the reading times (Baayen et al., 2008). The models were constructed using the *lmer* function in the lme4 package in R (Bates et al., 2012). Separate analyses were carried out for each subset of the target materials (i.e., NPIs and agreement materials). All models reported here are maximal models that have converged (Barr et al., 2013). For the mixed effects models, main interest of comparisons were set up as contrasts with Helmert coding (Venables and Ripley, 1999; Vasishth and Broe, 2011; Vasishth and Drenhaus, 2011), and they were included in the mixed effects models as fixed effect predictors (see below). Since the CS from the AQ questionnaire was the major subscale that has been shown to reflect speakers' pragmatic reasoning abilities in language processing (Nieuwland et al., 2010; Sikos et al., 2013), we will mainly focus on this subscale of the AQ[7]. Each participant's CS score was entered into the mixed effects models as an additional fixed effect predictor. Random effect structure included random intercepts for subjects and items, as well as random slopes of the fixed predictors. Before constructing the models, reading times longer than 2000 ms were removed, and all reading times were log-transformed.

NPI licensing

Acceptance rate. The averaged acceptance rate of each condition is presented in **Table 2**. As expected, the two licensed grammatical conditions (9a and 9c) have the highest acceptance rate (0.87 and 0.81), the unlicensed ungrammatical condition (9e) has the lowest acceptance rate (0.16). Critically, the interference conditions (9b and 9d) were accepted more often than the non-interference ungrammatical one (0.25 and 0.27), manifesting a standard interference effect.

We first analyzed all the data together, using a mixed effects logistic model. We defined three orthogonal contrasts: the first contrast examined the grammaticality effect (***Grammaticality***), in which the licensed grammatical conditions were contrasted with the ungrammatical conditions (i.e., *a, c* vs. *b, d, e*,); the second contrast examined the interference effect (***Interference***), in which the interference ungrammatical conditions *b* and *d* were contrasted with the unlicensed ungrammatical condition *e* (*b, d* vs. *e*); in the third contrast (***Licensor***) the two types of licensors were compared (*a, b* vs. *c, d*). These three contrasts were entered into the mixed effects model as fixed effect predictors. In addition, we included each participant's CS scores from the Autism Quotient as another fixed effect predictor in the model. Among the 88 participants included in this analysis, the minimum CS score was 0 and the maximum was 10, with a mean of 3.1 (median 3, and standard derivation 2.2). For the random effect structure, we included random intercepts for both subjects and items, as well as random slopes of the three user-defined contrasts above. The model output is presented in **Table 3** below:

Table 3 | NPI licensing acceptance rate: fixed effects from the mixed effect logistic model for the overall data.

	Estimate	**Std. Error**	***z* value**	***Pr*(>|*z*|)**
Grammaticality	5.08	0.66	7.74	9.63E-15***
CSscore	−0.11	0.08	−1.36	0.18
Interference	1.30	0.22	5.79	6.97E-09***
Licensor	0.65	0.30	2.20	0.03*
Gram:CSscore	0.08	0.17	0.47	0.64
Inter:CSscore	−0.13	0.06	−2.15	0.03*
Licensor:CSscore	−0.06	0.08	−0.81	0.42
Inter:Licensor	−1.03	0.77	−1.34	0.18
CSscore:Inter:Licensor	0.06	0.21	0.27	0.79

*lmer[acceptance ~ gram * CSscore + interfence * CSscore + licensor * CSscore + licensor:interference + licensor:interference:CSscore + (1 + gram + interference + licensor|subj) + (1 + gram + interference + licensor|item), data = dataframe, family = "binomial*[8]*"]. *p < 0.05; ***p < 0.001.*

[7]Among the other subscales, social skill showed a very similar effect as CS. Other subscales did not show any effect. We report the interactions between these sub-scales and other fixed effects in the appendix.

[8]The model in **Table 3** did not include all the interaction terms between the three user-defined contrasts. Note that the interaction between Grammaticality and Interference is irrelevant, since there is no interference to start with on the grammatical conditions (i.e., interference conditions are themselves all ungrammatical). For the interaction between Licensor and Grammaticality/Interference, since our experimental design is not a full 2 × 3 factorial design (i.e., there is only one plain unlicensed condition in **Table 1**), not all possible interaction combinations are possible. Therefore, we only included Licensor:Interference in the regression formula, which is essentially the same as the Licensor:Grammaticality interaction.

Not surprisingly, the model revealed a significant effect for both Grammaticality and Interference. What is crucial, is that there is a significant interaction between the effect of Interference and CS scores, indicating that the difference between the interference condition on the one hand, and the plain unlicensed condition on the other, is affected by participants' general pragmatic skills assessed via their CS scores. In addition, there is also an effect of Licensor, suggesting a difference between the *no* and *only* conditions.

The effect of licensor type

The data from licensor *no* and *only* are plotted separately in **Figure 1**. The unlicensed condition is shared by the two licensor groups (i.e., condition *e* in **Table 1**).

Paired comparisons between conditions showed that sentences licensed under *no* were accepted more often than those licensed under *only* (condition *a* vs. *c* in **Table 1**, $p < 0.001$); but the interference condition under *no* is not different from the interference condition under *only* (condition *b* vs. *d*, $p > 0.2$). Therefore, the effect of Licensor observed in **Table 3** was mainly driven by the grammatical conditions: subjects were slightly more resistant in accepting *only* as a grammatical licensor. We will come back to this observation in the general discussion.

The Interference-by-CS-scores interaction

Our model showed a robust interaction between the Interference effect and CS scores. We further discuss what this interaction entails in this section. Since our model revealed no interaction between CS scores and licensor type (i.e., neither three-way nor two-way interactions), we do not expect the effect of CS scores on interference to be conditioned by licensor type. For the completeness of our presentation, however, we present results from *no* and *only* separately.

In the analysis of licensor *no*, we only present the three relevant conditions in **Table 1**: conditions 9*a*, *b*, and *e*. The mixed effects model was constructed largely in the same way as before, except that only two contrasts were defined as fixed effect predictors: ***Grammaticality***, which contrasted the grammatical condition 9*a* with the other two ungrammatical conditions (i.e., *a* vs. *b* and *e*); and ***Interference***, which compared the interference ungrammatical condition 9*b* with the unlicensed ungrammatical condition *e* (*b* vs. *e*). For *only*, the three relevant conditions were 9*c*, *d*, and *e* in **Table 1**, and the two contrasts were defined as ***Grammaticality*** (*c* vs. *d*, *e*) and ***Interference*** (*d* vs. *e*). The model results for the fixed effects are presented in **Table 4**.

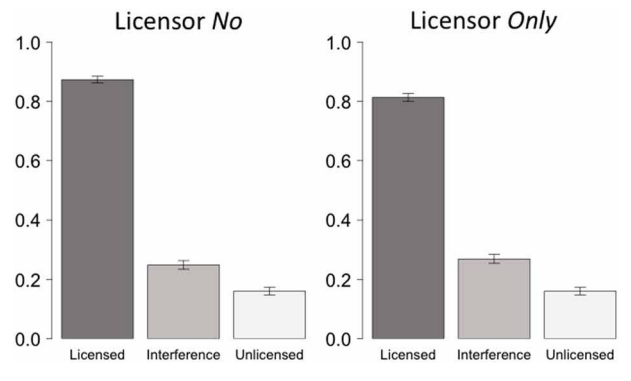


FIGURE 1 | Acceptance rate for the two different licensors.

As expected, both the Grammaticality effect and the Interference effect are highly significant, and the interaction between Interference and CS scores is also significant. To better understand the interaction between CS scores and the interference effect, we did the following two analyses for licensor *no* and *only* separately. For each subset of the data, we first carried out a correlation analysis between the size of the interference effect and individual participants' CS scores. For each participant, we calculated a difference score between their acceptance rates, averaged across items, in the interference condition and the plain unlicensed condition. This difference score represents the size of the interference effect for each subject. We then correlated these difference scores with their CS scores. There is a significant *negative* correlation between the difference scores and participants' CS scores for licensor *no* [Pearson's $r = -0.28$, $t_{(86)} = -2.7$, $p < 0.01$], as well as for licensor *only* (Pearson's $r = -0.21$, $p < 0.05$). The negative correlation suggests that the higher a participant's CS score, the smaller the difference between their

Table 4 | NPI licensing acceptance rate: fixed effects for two different NPI licensors.

	Licensor *No* **Pr(>|z|)**	**Licensor *Only*** ***Pr*(>|*z*|)**
Grammaticality	4.2e-11***	6.9e-09***
Interference	1.6e-08***	6.2 − 07***
CSscore	0.15	0.37
Gram:CSscore	0.32	0.27
Inter:CSscore	0.03*	0.07^

*model = lmer[acceptance ∼ gram * CSscore + inter * CSscore + (1 + gram + inter|subj) + (1 + gram + inter + CSscore|item), data = dataframe, family = "binomial"]. *p < 0.05; ***p < 0.001; ^p < 0.1.*

Table 5 | NPI licensing RTs: fixed effects from the maximal linear mixed effects model on the critical word and the spill-over word.

	CW ***p*-value**	**CW + 1** ***p*-value**
Grammaticality	**<0.0001*****	**0.03***
CSscore	0.76	0.36
Interference	**0.02***	0.33
Licensor	0.68	**0.006****
Gram:CSscore	0.88	0.28
Inter:CSscore	0.92	0.42
Licensor:CSscore	0.12	0.9
Inter:Licensor	0.22	0.33
CS:Inter:Licensor	0.4	0.53

*model = lmer[logRT ∼ gram * CSscore + inter * CSscore + licensor * CSscore + licensor:inter + licensr:inter:CSscore + (1 + gram + licensor + inter |subj) + (1 + gram + licensor + inter |item), data = dataframe]. *p < 0.05; **p < 0.01; ***p < 0.001.*

interference condition and plain unlicensed condition. In other words, participants with higher CS scores treated the interference conditions like the plain unlicensed condition, and rejected them both; on the other hand, participants with lower CS scores were more likely to erroneously accept the interference conditions. We plot the correlation results in **Figures 2A**, **3A**.

Second, we carried out a split-group analysis. We separated our participants into two groups along the median split of their CS scores: participants in one group had CS scores above 3 (high CS group, $n = 36$), and participants in the other group had scores below 3 (low CS group, $n = 43$). Participants who had a CS of exactly 3 were not included in either group. In **Figures 2B**, **3B**, we present the mean acceptance rate results for these two participant groups, separated by licensor type.

We carried out mixed effects models for each CS group under each licensor. Licensor *no* and *only* showed very similar patterns. For licensor *no* (**Figure 2B**), both high and low-CS groups showed the expected Grammaticality effect ($ps < 0.0001$); but only the low CS group showed an Interference effect (high CS: $p > 0.3$; low CS: $p < 0.0001$). For licensor *only* (**Figure 3B**), both high and low CS groups showed a clear Grammaticality effect ($ps < 0.0001$). The low CS group also showed a strong Interference effect ($p < 0.0001$), whereas this effect was much weaker for the high CS group ($p < 0.06$).

To summarize the acceptance rating data on NPIs, the group averaged data showed an interference effect, but this effect is crucially modulated by individual subjects' pragmatic-communicative skills, across different licensors.

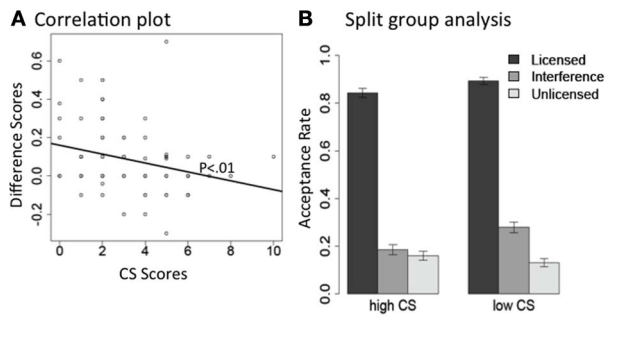


FIGURE 2 | The interaction between interference effect and CS scores in acceptance rate, for the NPI licensor *no*. (A) Correlation between each individual subject's difference scores between their interference condition and the plain unlicensed condition (*Y*-axis: acceptance rate of Interference—acceptance rate of Unlicensed) and their CS scores (*X*-axis: CS scores). **(B)** Acceptance rate for each condition plotted separately for the high and low CS groups.

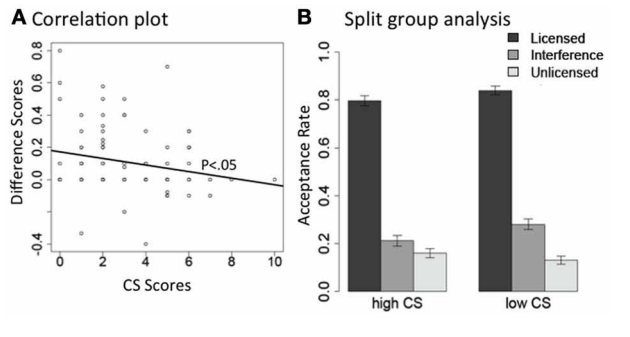


FIGURE 3 | The interaction between interference effect and CS scores in acceptance rate, for the NPI licensor *only*. (A) Correlation between each individual subject's difference scores between their interference condition and the plain unlicensed condition (*Y*-axis: acceptance rate of Interference—acceptance rate of Unlicensed) and their CS scores (*X*-axis: CS scores). **(B)** Acceptance rate for each condition plotted separately for the high and low CS groups.

Self-paced reading time

In **Figure 4**, we plot the reading time from four words prior to the NPI word *ever* and two words after it, with combined data from licensor *no* and licensor *only*. As shown in the plot, combined data from *no* and *only* showed differences among the licensed, interference, and unlicensed conditions only immediately at the critical NPI word (CW) *ever*. The grand average RTs on the CW are shown in **Table 2**.

We carried out mixed effects linear regression modeling on the RTs at the CW. The fixed and random effect structures are essentially the same as in our mixed effects logistic models discussed earlier. Prior to the analyses, we log-transformed all the RTs, and centered the CS scores. We first did analyses on the entire data set, and then did separate analyses for licensors *no* and *only*. The model output from the entire data set on the CW is presented below.

On the CW, the results revealed the expected effects for Grammaticality and Interference, but in contrast to the acceptance rate results, there was no interaction between CS scores and Interference. On the spill-over word CW + 1, there was a Grammaticality effect, but no Interference. There was also an unpredicted effect of Licensor. Further examination showed that this effect appeared because the grammatical and interference conditions under "only" were both read slower than the same two conditions under "no." Since this effect wasn't predicted under any of our hypotheses, we will not go into it further. We next analyzed data for *no* and *only* separately.

Licensor no

On the word CW. The word-by-word RTs (4 words prior and 2 words after the CW) are plotted in **Figure 5A**. On the CW *ever*,

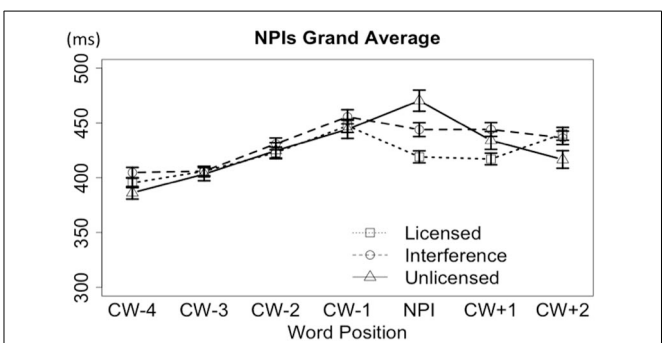


FIGURE 4 | Word-by-word reading times for the NPI stimuli set, collapsing the licensors "*no*" and "*only*."

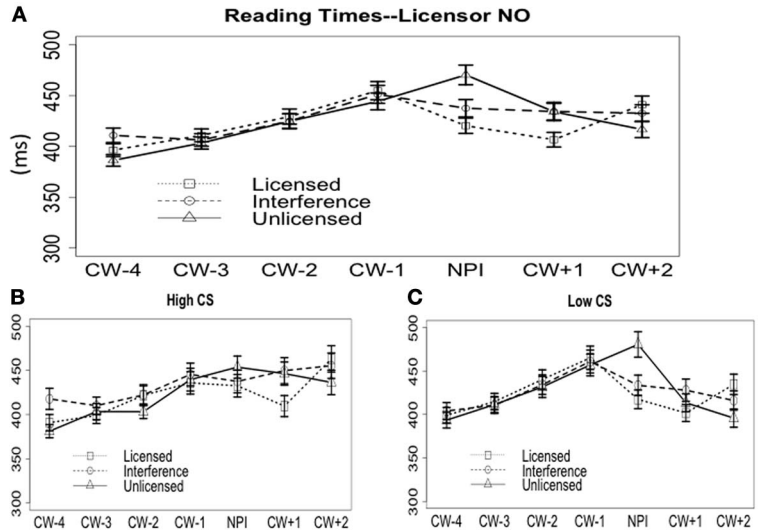


FIGURE 5 | Reading times when the NPI licensor was "*no.*" (A) Word-by-word reading time cross all subjects. **(B)** Word-by-word reading time for the high CS group. **(C)** Word-by-word reading time for the low CS group.

Table 6 | NPI licensing RTs: fixed effects from the linear mixed effect models, separated for two different licensors.

	Word CW		Word CW+1	
	Licensor *No* *p*-value	**Licensor *Only* *p*-value**	**Licensor *No* *p*-value**	**Licensor *Only* *p*-value**
Grammaticality	**0.0005*****	**<0.0001*****	**0.01***	0.2
Interference	**0.005****	0.15	0.5	0.1
CSscore	0.60	0.96	0.3	0.4
Gram:CSscore	0.35	0.53	0.4	0.1
Inter:CSscore	0.79	0.92	0.5	0.6

*model = lmer(logRT ~ gram * CSscore + inter * CSscore + (1+ gram + inter|subj) + (1+ gram + inter + CSscore|item),data = dataframe). *p < 0.05; **p < 0.01; ***p < 0.001.*

the grammatical condition was read faster (421 ms) than both the plain unlicensed condition (473 ms) and the interference condition (439 ms); but, the interference condition was also faster than the plain unlicensed condition, suggesting an interference effect.

The output of the mixed effects linear regression model for reading times on the CW is presented in **Table 6**. The model output shows a significant effect of Grammaticality and Interference, suggesting that the grammatical condition is read significantly faster than both ungrammatical conditions, while the interference condition is read faster than the plain unlicensed condition. But, the model did not show any effect of CS scores, nor any interaction between CS scores and any other effects.

Although the interaction between CS scores and Interference shown in **Table 6** isn't significant, to find out if there was any trend of an effect from CS scores, we carried out an exploratory correlation and split group analysis for the CW. The procedure was the same as with the analyses of acceptance rate data presented above. The first result is that there was no correlation between CS scores and interference [Pearson's $r = -0.05$, $t_{(86)} = -0.49$, $p > 0.6$]. For the split group analysis, we again separated participants into a high-CS ($n = 36$) and a low-CS group ($n = 43$) based on the median-split (CS = 3) of their CS scores. The mean RT for each group is plotted in **Figures 5B,C**. For the high-CS group, on the CW *ever*, neither the effect of Grammaticality nor Interference was significant (ps > 0.2)—there was no difference between any of the conditions (licensed, 433 ms; interference, 438 ms; unlicensed, 454 ms). For the low-CS group, however, both Grammaticality ($p < 0.01$) and Interference ($p < 0.05$) were significant. The licensed NPI (417 ms) was read faster than the unlicensed (480 ms) and the interference condition (433 ms); the interference condition was also faster than the unlicensed condition.

To summarize, on the CW, the averaged data showed the standard Grammaticality and Interference effect, but neither the mixed effects model nor the correlation analyses suggested any interaction between CS scores and the Interference effect. The split group analysis showed a small trend of modulation by CS scores: only the low-CS-scores group showed Interference, but it is difficult to draw any conclusions from this result since the high-CS-scores group did not show any difference between conditions, let alone an interference effect.

On the word CW + 1. On the spillover word (**Figure 5A**), the grand average over all the participants showed a faster reading time on the licensed condition (407 ms) than on the interference and unlicensed conditions (both were 434 ms). There is a significant Grammaticality effect, but no effect of Interference, CS scores, or any interactions (see **Table 6**). The exploratory correlation analysis found no correlation between the size of the interference effect and the CS scores [Pearson's $r = -0.07$, $t_{(86)} = -0.7$, $p > 0.4$]. The exploratory split-group analysis in

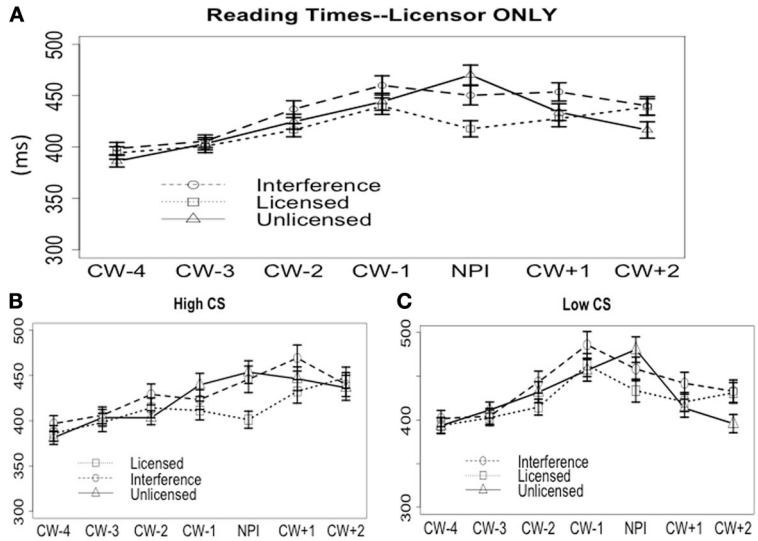


FIGURE 6 | Reading times when the NPI licensor was "*only.*" (A) Word-by-word reading time cross all subjects. **(B)** Word-by-word reading time for the high CS group. **(C)** Word-by-word reading time for the low CS group.

which we separated the high-CS and low-CS groups of participants, however, revealed different trends for the two groups (see **Figures 5B,C**). For the high-CS group, there was an effect of Grammaticality ($p < 0.01$), but no Interference ($p > 0.6$). The licensed condition (410 ms) was read faster than both the interference (450 ms) and the unlicensed conditions (446 ms), and there was no difference between the latter two. For the low-CS group, there was no effect of Grammaticality or Interference (Grammaticality, $p > 0.2$; Interference, $p > 0.5$; licensed, 402 ms; interference, 428 ms; unlicensed, 413 ms).

To summarize the results for the licensor *no*, grand average data showed a significant Grammaticality effect and an Interference effect on the CW. For the spillover word, there was only a Grammaticality effect. However, when we separated the high-CS group from the low-CS group, there was a trend of an effect of CS scores: for the high-CS group, there was no difference at the CW, but there was a grammaticality effect at the spill-over word, without an interference effect; for the low-CS group, there was both a grammaticality and an interference effect on the CW, yet no differences at the spill-over word. In other words, the low-CS group showed immediate sensitivity to ungrammaticality at the critical NPI word, but this sensitivity is also prone to an interference effect; the high-CS group, on the other hand, was slightly delayed in showing sensitivity to ungrammaticality, but, at the same time, was more resistant to the interference effect. Some caution is warranted, however, in interpreting the results from the split-group analysis, since based on the comprehensive model and the exploratory correlation analysis, the interaction between CS scores and Interference essentially presented a null result.

Licensor only

On the word CW. The grand average of the word-by-word RTs are shown in **Figure 6A**. On the CW *ever*, the licensed condition (420 ms) was read faster than the plain unlicensed condition (473 ms) and the interference condition (453 ms). The model output in **Table 6** shows only a significant Grammaticality effect, but no Interference effect. This is significantly different from licensor *no*, and we will discuss it further in the general discussion.

Table 7 | Number-agreement acceptance rate: fixed effects from the mixed effect logistic model.

	Estimate	Std. Error	*z* value	*Pr*(>|*z*|)
Grammaticality	1.99	0.2	9.4	<2e-16***
Interference	0.82	0.15	5.59	2.3e-08***
CSscore	0.09	0.07	1.25	0.2
Gram:CSscore	0.04	0.06	0.68	0.5
Inter:CSscore	−0.02	0.04	−0.61	0.5

*model = lmer(acceptance ~ gram * CSscore + inter * CSscore + (1+ gram + inter|subj) + (1+ gram + inter + CSscore|item),data = dataframe, family = "binomial"). ***p < 0.001.*

The results from the exploratory correlation analysis found no correlation [Pearson's $r = 0.03$, $t_{(86)} = 0.3$, $p > 0.7$]. Results from the split group analysis are shown in **Figures 6B,C**. For the high-CS group, there was an effect Grammaticality ($p < 0.001$), but no effect of Interference ($p > 0.4$): the licensed condition (401 ms) was read significantly faster than both the interference condition (446 ms) and the unlicensed condition (454 ms), and there was no difference between the latter two. For the low-CS group, there was also an effect of Grammaticality ($p < 0.01$). There seems to be a numerical trend of interference, but the effect of Interference wasn't significant ($p > 0.2$) (licensed 434 ms; unlicensed 480 ms; interference 458 ms).

On the word CW + 1. On the spillover word, the grand means (**Figure 6**) of the three conditions are: licensed 428 ms, interference 454 ms, and unlicensed 434 ms. The comprehensive mixed-effect model did not reveal any significant effects of Grammaticality or Interference (**Table 6**), and this was confirmed by the mixed effect models within each CS group (all *p*s > 0.1).

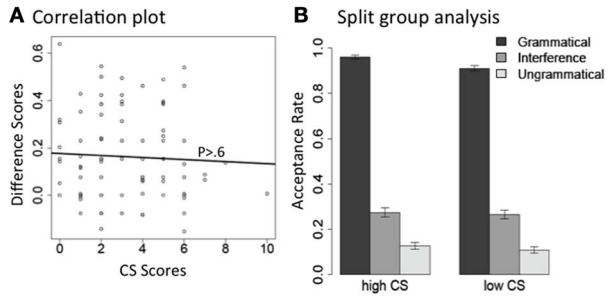


FIGURE 7 | The interaction between interference effect and CS scores in acceptance rate, for subject-verb agreement errors. (A) Correlation between each individual subject's difference scores between their interference condition and the plain unlicensed condition (*Y*-axis: acceptance rate of Interference—acceptance rate of Ungrammatical) and their CS scores (*X*-axis: CS scores). **(B)** Acceptance rate for each condition plotted separately for the high and low CS groups.

Table 8 | Number agreement RTs: fixed effects from the linear mixed effect model.

	Word CW ***p*-value**	**Word CW + 1** ***p*-value**
Grammaticality	**<0.0001*****	**<0.0001*****
Interference	**0.01***	0.5
CSscore	0.5	0.6
Gram:CSscore	0.6	0.5
Inter:CSscore	0.6	0.6

*model = lmer(logRT ~ gram * CSscore + inter * CSscore + (1+ gram + inter|subj) + (1+ gram + inter + CSscore|item), data = dataframe). *p < 0.05; ***p < 0.001.*

To summarize the self-paced reading time data from the licensor *only*, the grand average data showed a grammaticality effect without an interference effect. The same pattern largely holds for both high and low-CS groups, but the low-CS group showed a small trend of interference, as well.

Subject-verb number agreement

Within the subject-verb number agreement materials, two Helmert contrasts were defined in order to examine both the grammaticality effect: ***Grammaticality*** (10a vs. 10b and 10c in **Table 1**), and the interference effect: ***Interference*** (10b vs. 10c). Everything else about the mixed effects model structures was set up in the same way as for the analyses of NPI licensing.

Acceptability rating

The average acceptability rating (see **Table 2**) was 0.92 for the grammatical condition (10a), 0.12 on the plain ungrammatical condition (10c), and 0.28 on the interference ungrammatical condition (10b). The mixed effects logistic model showed a significant Grammaticality effect and a significant Interference effect. No other effects were significant. Crucially different from the NPI results (see **Tables 3**, **4**), CS scores did not affect participants' judgment of subject-verb agreement errors. The model output for the fixed effects is shown below in **Table 7**.

To make a parallel comparison with the NPI stimuli, **Figure 7A** presents the correlation between the interference effect and CS scores, and **Figure 7B** presents the median-split analysis. The lack of correlation in **Figure 7A** (Pearson's $r = 0.05, p > 0.6$) confirms that CS scores did not affect the interference effect in the agreement items. And the mixed effect models within each group also found the same Grammaticality and Interference effects for both high and low-CS groups (all *p*s < 0.0001).

Self-paced reading time

On the word CW. Word-by-word reading times are plotted in **Figure 8A**. The average RTs on the critical verb (e.g., *fail* in example 10b) are 463 ms for the grammatical condition (10a), 547 ms for the plain ungrammatical condition (10c), and 509 ms for the interference condition (10b). The mixed effects model showed a significant effect of Grammaticality and Interference (see **Table 8**) on the CW, such that the grammatical condition was read faster than the other two conditions (*a* vs. *b*, *c*, $p < 0.001$), and the interference condition was read faster than the plain ungrammatical condition (*b* vs. *c*, $p < 0.01$). There were no interactions between Interference and CS scores.

The split-group analysis, as shown in **Figures 8B,C**, revealed qualitatively similar patterns for high-CS and low-CS groups. For the high-CS group, there was an effect of Grammaticality ($p < 0.001$): the grammatical condition (439 ms) was read significantly faster than both the ungrammatical condition (511 ms) and the interference condition (481 ms); and an effect of Interference as well (although slightly weaker, $p < 0.07$). For the low CS group, the grammatical condition (480 ms) was read significantly faster than both the ungrammatical condition (576 ms) and the interference condition (521 ms); and the difference between the latter two was also significant (Grammaticality, $p < 0.01$; Interference, $p < 0.01$).

On the word CW + 1

At the spillover word, grand averages of the three conditions were: grammatical 428 ms, interference 480 ms, and ungrammatical 493 ms. The mixed effects model showed a significant effect of Grammaticality, but no effect of Interference (see **Table 8**). The split-group analysis (see **Figure 8**) revealed very similar results for both participant groups. For the high-CS group, the grammatical condition (429 ms) was read faster than the ungrammatical condition (502 ms) and the interference condition (500 ms), and there was no difference between the latter two (Grammaticality, $p < 0.0001$; Interference, $p > 0.8$). The low-CS group showed the same pattern: the grammatical condition (420 ms) was read faster than the ungrammatical (485 ms) and interference (477 ms) conditions (effect of Grammaticality, $p < 0.001$), with no significant difference between the latter two (effect of Interference, $p > 0.8$).

To summarize, for the agreement stimuli, we observed the grammaticality effect and the interference effect in both acceptability ratings and the self-paced reading time on the critical word. On the spillover word, self-paced RTs only showed a grammaticality effect, but no interference. In all these measures, the high and low-CS participants performed in very similar ways.

GENERAL DISCUSSION

The current study revealed three main findings. First, only NPI interference, but not agreement interference, is affected by

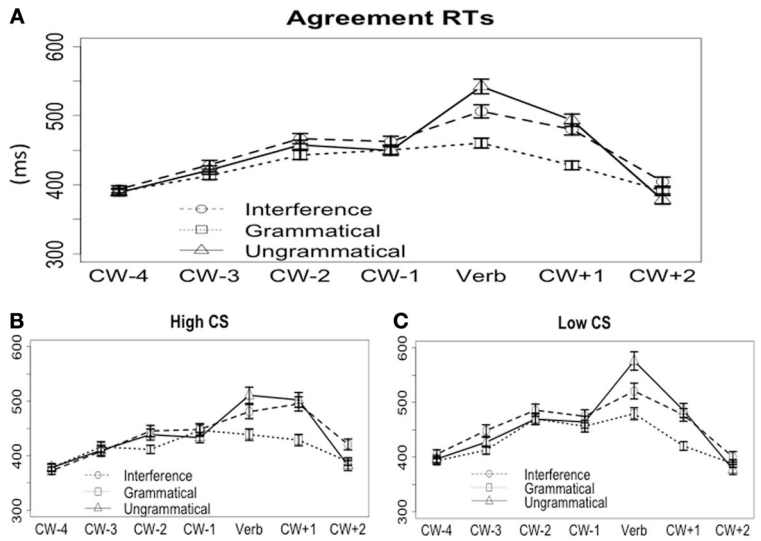


FIGURE 8 | Reading times for the agreement set of the stimuli. (A) Word-by-word reading times across all subjects. **(B)** Word-by-word reading times for the high CS group. **(C)** Word-by-word reading times for the low CS group.

individual subject's pragmatic-communicative skills. Second, the modulation of pragmatic-communicative skills mostly has its effect on offline acceptability rating, but not online reading time, although there seems to be a trend of effect in online RTs as well. And third, different NPI licensors, in particular, *no* and *only*, presented distinct interference profiles: while both showed offline interference in acceptability, NPIs under *only* did not show online interference. We turn below to the discussion of these observations.

INTERFERENCE IN ACCEPTANCE RATE AND THE EFFECT OF AUTISTIC TRAITS

A critical finding of the current study is that for the NPI materials, but not for the agreement interference stimuli, participants' acceptance rate was affected by their autistic-associated traits; in particular, their communication skills, as measured by the CS of the AQ questionnaire. Participants with higher CS scores, i.e., those that are relatively worse in their general pragmatic communicative skills, were less prone to NPI interference, as demonstrated by their more accurate acceptability judgments. On the other hand, participants with better communicative skills (lower CS scores) more often accepted the interference conditions. In contrast to the case with NPI licensing, participants' autistic traits did not seem to affect their acceptance of subject-verb agreement sentences, suggesting that subject-verb interference and NPI interference, although on the surface they look very similar, may arise from different sources.

We argue that the different interference profiles stem from the fact that NPI licensing and subject-verb agreement are different types of linguistic dependencies. It is uncontroversial that subject-verb number agreement involves a syntactic matching process that checks the number features on the subject and its corresponding verb. In incremental parsing, the subject of a sentence is likely to have been removed from focal attention when the verb is encountered (McElree, 2001); therefore, the real-time construction of a subject-verb agreement relationship depends on the successful retrieval of the subject's features. Memory-retrieval-based interference arises when the target of retrieval shares certain features with other items that have recently been processed (Lewis and Vasishth, 2005; Lewis et al., 2006). Under this account of subject-verb agreement and the corresponding interference effect, interference errors stem from misapplication of the mechanism by which number agreement is computed.

We likewise argue that NPI interference is closely tied to one of the mechanisms by which NPIs are regularly licensed. As discussed in the introduction, in addition to a logical-semantic mechanism, there is also a pragmatic component to NPI licensing in English. Particularly relevant to our purposes here, negative inferences are employed regularly as part of a pragmatic licensing mechanism for NPIs. During the comprehension process, the parser may over-apply the pragmatic licensing strategy, and use even unwarranted negative inferences to license NPIs, resulting in interference.

Under this account, interference in syntactic agreement and interference in NPI licensing are driven, at least partially, by different underlying sources. This shouldn't be totally unexpected, since these two linguistic phenomena involve different representations and computations in the first place: the agreement process is purely syntactic, whereas NPI licensing is at the interface of different systems, including syntax, semantics, and pragmatics. It is not surprising that the specific linguistic properties of each construction lead to substantial differences in how they are processed in comprehension.

The current results also add to the growing literature that autistic traits are present among the neurotypical population and they affect language processing in non-trivial ways. Our results, in line with previous findings, suggest that the two sub-scales from the AQ—Social Skill and Communication—may have particular

influence on pragmatic language processing. Since case studies in this regard are still relatively sparse, more future research is needed to further establish this association. There are many different kinds of pragmatic phenomena in language processing, and it is an open question whether they are in general affected by individual differences along the dimension of autistic traits. If it turns out that autistic traits only selectively target a subset of these phenomena, it would be very informative for the construction of a constrained pragmatic theory of language processing.

ONLINE INTERFERENCE AND THE (LACK OF) EFFECT OF AUTISTIC TRAITS

Although there is a strong effect of autistic traits on the offline acceptability rating, their effect on online reading time is much weaker. The split group analysis seems to show a trend where there is more interference for the low-CS group than the high-CS group, but the mixed effects models revealed no interaction between CS scores and interference effect. The lack of an interaction in the comprehensive model could be due to insufficient power in the data, in which case we may still consider the online interference effect as being qualitatively similar to the offline effect. This is a potential explanation, but also one that is difficult to validate given the null result. While keeping this possibility in mind, we will entertain the alternative possibility that there is genuinely no effect of CS scores on online interference and discuss the implications of that possibility. Another interesting observation about the online interference effect is that, for NPI licensing, we only observed interference for the licensor *no*, but not the licensor *only*. The difference between these two licensors is important for our explanation of the online interference effect, but we will focus only on *no* for the moment, and come back to *only* in the next section.

The lack of modulation by participants' communicative-pragmatic skills on the online interference effect suggests that NPI licensing may actually involve a syntactic matching process, like subject-verb agreement. This was the original hypothesis in Vasishth et al. (2008), which postulated a search process for a syntactic [+Neg] feature when an NPI word such as *ever* is encountered. We questioned this hypothesis earlier because it does not fully represent how NPIs are licensed—it overlooks the fact that NPI licensing is not *just* a syntactic process, but involves semantic and pragmatic mechanisms. However, the fact that NPI licensing is an interface phenomenon that involves multiple levels of representations and processes does not exclude the possibility that syntactic matching exists within one sub-component of the licensing process. The syntactic [+Neg] feature is a particularly suitable candidate to serve as the relevant matching feature, since, cross-linguistically, negation is the most robust NPI licensor. This line of reasoning would make NPI licensing similar to subject-verb agreement in some respects. If the regular memory retrieval mechanisms apply in both cases, one would expect similar online interference with no modulation from individual pragmatic skills.

We also want to point out that, by recognizing such a syntactic licensing process, at least for licensors such as *no* (see the contrast with *only* below), we acknowledge a syntactic process for NPI licensing that has not been fully recognized or emphasized in previous research for weak NPIs like *ever*, which can be licensed under a broad range of licensors. Polarity items that are only licensed by negation—called "strict" NPIs (Giannakidou, 1998, 2011; Zwarts, 1998)—are cross-linguistically common, for example, so-called "n-words" in Romance languages. Purely syntactic mechanisms like agreement have been proposed to account for the distribution of n-words (Haegeman and Zanuttini, 1991; Zanuttini, 1991; Zeijlstra, 2004, inter alia). But, traditionally, the general account of NPI licensing, especially for weak NPIs like *ever*, has been deliberately divorced from an agreement-based explanation. We agree with the traditional wisdom, but based on the current data, we also suggest that a syntactic feature matching process may exist in parallel with other licensing mechanisms, even for weak NPIs, at least for a subset of licensors—those that contain a syntactic [+Neg] feature.

Licensing as an integrated syntax-semantics process is to be expected if (a) we take seriously the idea that NPI licensing is a grammatical phenomenon driven by the logical properties of lexical expressions, and (b) there is a strict isomorphism between the syntax and the semantics. Under these two theses, the logical property of negation is mapped onto a morphosyntactic feature [+negative] (for an early discussion of such a model see Giannakidou, 1998). NPI licensing will then always involve at least this component of integrated syntax-semantics matching, and online processes access that. But importantly, even if we recognize a syntactic feature-matching component in NPI licensing, the overall process is still crucially different from subject-verb agreement in many ways. In particular, NPI licensing involves semantic *and* syntactic, as well as pragmatic mechanisms, as we discussed earlier. But for the agreement sentences, whether they are acceptable or not is determined only by whether or not the syntactic matching process on the relevant number feature is successful—there is no obvious connection between the processing of syntactic agreement and the final interpretation of a sentence. For instance, Lau et al. (2008) showed that when people were lured by interfering agreement number features, as in "*The phone by the toilets were . . .*," they nevertheless did not make mistakes in assigning the correct thematic role to the subject. NPI licensing, on the other hand, is a very different phenomenon. The presence of an NPI makes important contributions to the final propositional content. The acceptability of an NPI is not determined by the syntactic matching process alone, but is instead crucially regulated by semantic and pragmatic integration conditions. The effect of pragmatic inferences could be particularly strong in offline tasks, since participants are given enough time to reflect on what the target stimulus actually means, or could have meant.

The strong offline effect of individual subject's pragmatic skills leads to the question why such effects did not surface in online interference. One possibility is that the influence of pragmatic factors in online measures could have been masked by the strong presence of the memory-retrieval based effect, and hence was undetectable. This is not the most likely hypothesis, since for the licensor *only*, which we argue did not participate in a memory-retrieval based interference effect, we still did not observe a pragmatically driven interference in online measures. The other possibility is that since these interference sentences are ultimately

ungrammatical, the pragmatic inference may be a "last resort" strategy" in these situations, and hence have a delayed effect. We discuss these issues in more details in the next section, together with data from the licensor *only*.

Our current discussion about the source of online interference for *no* departs somewhat from our earlier work in Xiang et al. (2009), in which we conjectured that even the online interference effect for *no* (and *few*) was driven by pragmatic processes. In the current discussion we draw a distinction between online and offline interference effects, and argue that since a syntactic matching process is possible between *no* and an NPI, (some) online interference may thereby arise through a memory retrieval process, as argued in Vasishth et al. (2008) (modulo the possible additional contribution of a pragmatic process, as shown by the trend in the split group analysis). However, one important feature of the original analysis in Vasishth et al. (2008) is that memory search targets positions that are ruled out on syntactic grounds, and that NPI interference under *no* is a demonstration of syntactic interference, as accessing a licensor in a non-commanding position purportedly violates syntactic constraints. We do not think that the current results necessarily commit us to this position. We contend that questions about the search mechanism (i.e., whether or not the search process is blind/insensitive to syntactic constraints) and whether or not NPI licensing shows similarity-based feature interference may be two orthogonal issues. Although NPIs are generally c-commanded by their licensors, it is not obvious that a c-command requirement should be stated explicitly as part of the syntactic requirements on NPI licensing. It could simply be an epiphenomenon, within an isomorphic syntax-semantics level, of the semantic requirement that an NPI needs to stay in the semantic scope of its licensor. The computation of the semantic scope may track configurational relations like c-command, but this does not necessarily mean that the parser actually makes reference to the c-command condition in online processing. In other words, the memory retrieval process may target a [+negative] element, instead of a [+negative, +c-command] element, while there is simultaneously a semantic condition that checks whether or not the NPI falls within the semantic scope of the retrieved target. Of course this leaves open a number of non-trivial questions as to how semantic scope is tracked in online processing. One possibility is that we encode [+scope] in some way as a lexical feature on the retrieval target, and interference would arise largely in the same fashion as the proposal in Vasishth et al. (2008), but with the syntactic feature [+c-command] replaced by the semantic feature [+scope]. This approach calls for a detailed implementation as to how scope relations could be encoded as lexical features, when they obviously are not features stored in the lexicon. The other possibility is that scope relation can only fall out while propositional content is being incrementally composed, rather than being encoded on lexical items. If this is true, we need an explicit algorithm that can both derive correct scope relations at the proposition level, and also allow incorrect scope relations to be derived, in order to account for the interference effect. We do not have answers to these questions. But we think it would be too hasty to reach a conclusion about the exact search mechanism involved in NPI interference without fully exploring all of these logical possibilities.

DIFFERENT TYPES OF NPI LICENSORS

The perspective that multiple mechanisms are acting in parallel to license NPIs also helps explain the difference observed between the licensors *no* and *only*.

As discussed above, there is consensus that *only* does not license NPIs through a lexically encoded (syntactic) [+negative] feature, though the exact licensing mechanism for *only* as an NPI licensor is still under debate. The difference between *only* and other negative licensors such as *no* (or *few*) can be demonstrated by the syntactic diagnostics provided in Klima (1964), as was illustrated earlier.

In the current results, we saw that on grammatical conditions, NPIs licensed under *only* were accepted less often than those licensed under *no* (**Figure 1**). This could be due to a number of factors. For instance, *no* is a more frequent licensor than *only* in naturally occurring utterances (Xiang et al., 2009). This may have influenced the acceptability ratings of the two licensors. Alternatively, under the licensor *no*, an NPI can be licensed both syntactically and semantically. Syntactically, a feature-matching process may search and identify a target with a [+Neg] feature; semantically, a negative meaning may also be calculated. Syntactic and semantic processes converge on the final representation in which an NPI is licensed. With the licensor *only*, however, the syntactic feature-matching process fails, since *only* does not contain a morphosyntactic feature [+Neg]. Then, only the semantic route (via the exceptive entailment "nobody other than") would be available. The failure of isomorphism between syntax and semantics, in contrast with *no,* may have reduced the acceptability of NPIs under *only*. In a recent study (Xiang et al., 2013), acceptability ratings were collected for a larger set of licensors, including *no*, *few*, *only*, and emotive factives such as *amazed*, *surprised*, etc. It was found that the two syntactically negative licensors *no* and *few* are judged more acceptable than *only* and emotive factives, which are both non-negative. This is completely in line with the results reported here. Furthermore, since *few* is also much less frequent than *no* as an NPI licensor (Xiang et al., 2009), this result also suggests that lexical frequency *per se* does not completely determine the degree to which a licensor is accepted.

It is worth noting that the current study also revealed some difference between *no* and *only*: there was no obvious online interference effect for *only* (modulo the possibility that the low-CS group may have shown a trend of online interference). In a previous self-paced reading study, Xiang et al. (2006) also showed a lack of interference effect in online RTs for *only*. As mentioned earlier, the interference effect of *only*, or the lack of one, has not been widely tested. Although the current results showed a difference between *no* and *only* when we analyzed these two licensors separately (**Table 6**), there wasn't a Licensor by Interference interaction in the overall model in **Table 5**. This could be due to insufficient power in the data. We are fully aware that the lack of an interaction may undermine our proposed account of *only* here. More studies are needed to verify whether or not *only* is indeed resistant to online interference. But if it is, such a result is completely in line with the distinction we draw here between *no* and *only*: the former, but not the latter, is targeted by a syntactic feature-matching process. Therefore, similarity-based interference, which crucially relies on specific lexical features, will

not arise for *only* online. The immediate question for this explanation is why pragmatically-driven interference does not appear online.

One possibility is that the pragmatic inferences that drive the interference effect cannot be generated in time to trigger online interference. Instead, they are delayed until a later stage, and therefore only offline tasks can detect them. This is not the most likely hypothesis, however, given the large literature that has suggested that pragmatic inferences can be incrementally generated online (Altmann and Steedman, 1988; Sedivy et al., 1999; van Berkum, 2009; Nieuwland and Kuperberg, 2008). We propose that the reason pragmatically-driven interferences were predominately observed offline in the present study is not that such inferences failed to become available in time, but rather that the available inferences were not immediately adopted by the comprehension system to license NPIs.

First of all, the pragmatic licensing mechanism could in general be a more costly strategy than regular syntactic and semantic mechanisms. In a recent ERP study, Xiang et al. (2012, 2013) showed that, even for grammatically acceptable sentences, there is difference between NPIs that are licensed under pragmatically-derived negation (e.g., the negative implicature from emotive factive predicates) and those that are licensed under regular semantic negation (such as *no*)—only the semantic negation, but not the pragmatic negation, had a small P600 compared to the ungrammatical control condition. Second, as we mentioned above, while we recognize that pragmatically-derived negative implicatures can license NPIs, we also recognize that not all implicatures can do so. The specific conditions characterizing the "usable" implicatures are yet to be isolated, but we have conjectured that the kind of pragmatic implicatures that trigger interference effects are normally insufficient to actually license NPIs. It is likely that the comprehension system does not resort to such implicatures unless it is pushed into a corner, as in the presence of an ungrammatical sentence. If pragmatically-driven interference were the result of a last-resort strategy, it wouldn't necessarily surface in online processing.

A MULTI-DIMENSIONAL SYSTEM OF NPI LICENSING

NPI licensing reveals a case in which syntactic, semantic, and pragmatic processes act in parallel during parsing, which makes NPI licensing qualitatively different from purely syntactic dependencies, such as subject-verb agreement. The processing profile of NPI licensing is therefore much more complicated. Some licensors, such as *no*, which can participate in a syntactic licensing relation with an NPI, may be targeted by the same memory retrieval mechanisms that target other syntactic dependencies; but, for other licensors that do not bear the relevant syntactic features, memory retrieval of a lexical feature does not apply. In addition, since pragmatic licensing is a regular mechanism for NPI licensing, at least for English weak NPIs, the comprehension system may stretch it to cases in which pragmatic licensing normally does not apply, leading to pragmatically-driven interference.

To account for the full complexity of NPI licensing and the interference effect associated with it, a number of open issues need to be addressed in future work. First of all, if, as we argued above, interference associated with feature similarity only arises for licensors that contain a lexical [+negative] feature, we predict that online interference should be observed for some NPI licensors, but not others. Expressions that can license NPIs and, at the same time, are categorized as real negative expressions (i.e., under Klima, 1964) include *no*, *none*, *not*, *never*, *few*, *hardly*, *scarcely*, *seldom*, etc.; on the other hand, licensors that are not negative in the regular sense include the examples we mentioned earlier, such as *only*, *every*, comparatives, conditionals, emotive factives, questions, etc. Neither of these two groups has been tested exhaustively.

Second, we have argued that NPI interference is partially driven by pragmatic inferences, especially in the case of offline interference. We have made suggestions both about how such inferences arise and why they seem to be more prominent in offline measures. Our account of pragmatic interferences is closely associated with a particular construction that has been heavily tested by other researchers, as well as in our current work—that is, relative clauses. We made use of the well-known fact that modifiers, with relative clauses as a prime example, invite contrastive inferences. This gives rise to the following prediction: complement clauses (such as "*The fact that no student passed the exam …*"), which are minimally different from relative clauses but do not serve a modifier function, should not show interference effects, or at least not the kind of interference effect we have shown that can be modulated by individual subjects' pragmatic skills. Some results from Parker and Phillips (2011) provide preliminary support for this prediction. These authors showed a reduced interference effect for complement clauses, compared to relative clause structures, under the licensor *no*. Furthermore, with a licensor like *only* (as in "*The fact the only the best students passed the exam …*"), we predict that the interference effect on such clauses should be reduced to minimum, since the pragmatic source of interference has been entirely eliminated by the complement clause, and, in the meantime, "*only*" does not trigger syntactically associated interference either.

Finally, most of the current work on NPI interference has focused on languages that allow weak NPIs. These have a broad distribution and can be licensed under a variety of licensors. We conjectured that, for such NPIs, an independently available pragmatic licensing mechanism is over-applied in some situations, resulting in interference. Cross-linguistically, however, many languages have NPIs that are much more restricted in their distribution. It is possible that for some of these stricter NPIs, pragmatic-licensing mechanisms are never available in the grammar. This territory—interference with such NPIs—is still largely uncharted (for a recent examination of this sort, see Yanilmaz and Drury, 2013)[9]; and if they do show interference, we predict

[9]The recent ERP findings in Yanilmaz and Drury (2013) tested Turkish NPIs that have a very limited distribution. Interference was found on these NPIs. Since the constructions tested there were very different from the ones tested here (sentential complement of a matrix verb was tested), we won't go into further details. But the additional novel factor in Yanilmaz and Drury (2013) is that NPIs in Turkish come before their licensors in linear order, which may result in a forward expectation of a licensor (e.g., similar to a regular filler-gap dependency), rather than just backward search, as in the case of English NPIs.

it to be syntactically-driven interference, and not to be subject to individual differences in pragmatic skills.

CONCLUSION

In this study, we compared the interference effects in syntactic agreement and NPI licensing, especially with respect to their modulation by individual subjects' pragmatic skills. We showed that the interference profile for NPI licensing is more complicated than that for syntactic agreement, due to their representational differences. In particular, NPI interference is affected by (a) the type of NPI licensors involved, (b) the particular experimental tasks, and (c) individual subjects' pragmatic-communicative skills. All together, our results show that NPI licensing, different from the pure syntactic processes involved in agreement, evokes multiple different processes corresponding to different levels, or dimensions, of linguistic representations.

ACKNOWLEDGMENTS

The authors thank Lelia Glass, Steven SanPietro and Genna Vegh for their assistance in data collection, Matt Wagers for sharing his sentence material, and Brian Dillon and Ellen Lau for their constructive comments and suggestions.

REFERENCES

Atlas, J. D. (1993). The importance of being only: testing the neo-gricean versus neo-entailment paradigms. *J. Sem.* 10, 301–318. doi: 10.1093/jos/10.4.301

Altmann, G. T., and Steedman, M. (1988). Interaction with context during human sentence processing. *Cognition* 30, 191–238. doi: 10.1016/0010-0277(88)90020-0

Baayen, R. H., Davidson, D. J., and Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *J. Mem. Lang.* 59, 390–412. doi: 10.1016/j.jml.2007.12.005

Baker, C. L. (1970a). Double negatives. *Linguist. Inq.* 1, 169–186.

Baker, C. L. (1970b). "Problems of polarity in counterfactuals," in *Studies Presented to Robert B. Lees by his Students*, PIL Monograph Series 1, eds J. Sadock and A. Vanek (Edmonton, AB: Linguistic Research, Inc.), 1–15.

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., and Clubley, E. (2001). Autism-Spectrum Quotient (AQ): evidence from Asperger syndrome/high functioning autism, males and females, scientists and mathematicians. *J. Autism Dev. Disord.* 31, 1, 5–17. doi: 10.1023/A:1005653411471

Barr, D., Levy, R., Scheepers, C., and Tily, H. (2013). Random effects structure for confirmatory hypothesis testing: keep it maximal. *J. Mem. Lang.* 68, 255–278. doi: 10.1016/j.jml.2012.11.001

Bates, D. M., Maechler, M., and Bolker, B. (2012). *lme4: Linear Mixed-Effects Models using S4 Classes*. R package version 0.999999-0.

Bernardi, R. (2002). *Reasoning with Polarity in Categorial Type Logic*. Ph D dissertation. University of Utrecht.

Bishop, D. V. M. (1989). Autism, Asperger's syndrome and semantic-pragmatic disorder: where are the boundaries? *Br. J. Disord. Commun.* 24, 107–121. doi: 10.3109/13682828909011951

Bock, J. K., and Miller, C. A. (1991). Broken agreement. *Cogn. Psychol.* 23, 45–93. doi: 10.1016/0010-0285(91)90003-7

Bock, J. K., and Eberhard, K. M. (1993). Meaning, sound, and syntax in English number agreement. *Lang. Cogn. Process.* 8, 57–99. doi: 10.1080/01690969308406949

Chierchia, C. (2006). Broaden your views. Implicatures of domain widening and the 'logicality' of language. *Linguist. Inq.* 37, 535–590. doi: 10.1162/ling.2006.37.4.535

Condoravdi, C. (2010). NPI licensing in temporal clauses. *Nat. Lang. Linguist. Theory* 28, 877–910. doi: 10.1007/s11049-010-9115-z

Dillon, B., Mishler, A., Sloggett, S., and Phillips, C. (2013). Contrasting interference profiles for agreement and anaphora: experimental and modeling evidence. *J. Mem. Lang.* 69, 85–103.

Drenhaus, H., Saddy, D., and Frisch, S. (2005). "Processing negative polarity items: when negation comes through the backdoor," in *Linguistic Evidence—Empirical, Theoretical, and Computational Perspectives*, eds S. Kepser and M. Reis (Berlin; New York: Mouton de Gruyter), 145–165. doi: 10.1515/9783110197549.145

Duffley, P., and Larivée, P. (2010). Anyone for non-scalarity? *Engl. Lang. Linguist.* 14, 1–17. doi: 10.1017/S1360674309990402

Eberhard, K. M., Cutting, J. C., and Bock, J. K. (2005). Making syntax of sense: number agreement in sentence production. *Psychol. Rev.* 112, 531–559. doi: 10.1037/0033-295X.112.3.531

Gajewski, J. R. (2005). *Neg-raising: Polarity and Presupposition*. Doctoral Dissertation, MIT.

Giannakidou, A. (1998). *Polarity Sensitivity as (non)Veridical Dependency*. Amsterdam; Philadelphia: John Benjamins.

Giannakidou, A. (2006). Only, emotive factives, and the dual nature of polarity dependency. *Language* 82, 575–603. doi: 10.1353/lan.2006.0136

Giannakidou, A. (2011). "Negative and positive polarity items," in *Semantics: An International Handbook of Natural Language Meaning*, eds K. von Heusinger, C, Maienborn, and P. Portner (de Gruyter), 1660–1712.

Giannakidou, A., and Yoon, S. (2012). Weakening and reanalysis of EVEN in non-emphatic negative polarity items: evidence from Greek and Korean. *Chic. Linguist. Soc.* (to appear).

Grice, H. P. (1975). "Logic and conversation," in *Syntax and Semantics*, Vol. 3, *Speech Acts*, eds P. Cole and J. L. Morgan (New York, NY: Academic Press), 41–58.

Grodner, D., Dalini, M., Pearlstein-Levy, S., and Ward, A. (2012). "Factors that contribute to the use of perspective in referent identification," in *Talk Presented at the 25th Annual CUNY Conference on Human Sentence Processing* (New York, NY).

Haegeman, L., and Zanuttini, R. (1991). Negative heads and the Neg criterion. *Linguist. Rev.* 8, 233–251. doi: 10.1515/tlir.1991.8.2-4.233

Happé, F. (1993). Communicative competence and theory of mind in autism: a test of relevance theory. *Cognition* 48, 101–119. doi: 10.1016/0010-0277(93)90026-R

Hoeksema, J. (1994). "On the grammaticalization of negative polarity items," in *Proceedings of the Annual Meeting of the Berkeley Linguistics Society (BLS) 20*, eds S. Gahl, A. Dolbey, and C. Johnson (Berkeley, CA: Berkeley Linguistics Society), 273–282.

Horn, L. R. (1989). *A Natural History of Negation*. Chicago: The University of Chicago Press.

Horn, L. R. (1996). Exclusive company: only and the dynamics of vertical inference. *J. Sem.* 13, 1–40. doi: 10.1093/jos/13.1.1

Horn, L. R. (2002). "Assertoric inertia and NPI-licensing," in *Proceedings of the Annual Meeting of the Chicago Linguistic Society (CLS)*, Vol. 38. eds M. Andronis, E. Debenport, A. Pycha, and K. Yoshimura (Chicago), 55–82.

Kadmon, N., and Landman, F. (1993). Any. *Linguist. Philos.* 16, 353–422. doi: 10.1007/BF00985272

Klima, E. S. (1964). "Negation in english," in *The Structure of Language*, eds J. Fodor and J. Katz (Englewood Cliffs, NJ: Prentice-Hall), 246–323.

Krifka, M. (1995). The semantics and pragmatics of polarity items in assertion. *Linguist. Anal.* 15, 209–257.

Krifka, M. (2010). Before and after without coercion. Comment on the paper by Cleo Condoravdi. *Nat. Lang. Linguist. Theory* 28, 911–929. doi: 10.1007/s11049-010-9116-y

Ladusaw, W. A. (1979). *Negative Polarity Items as Inherent Scope Relations*. Ph D Dissertation, University of Texas at Austin.

Ladusaw, W. (1980). On the notion "affective" in the analysis of negative polarity items. *J. Linguist. Res.* 1, 1–16.

Lahiri, U. (1998). Focus and negative polarity in Hindi. *Nat. Lang. Sem.* 6, 57–123. doi: 10.1023/A:1008211808250

Lau, E., Wagers, M., Stroud, C., and Phillips, C. (2008). "Agreement and the subject of confusion," in *Talk Given at the CUNY Sentence Processing Conference*, (Chapel Hill, NC).

Lewis, R., and Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cogn. Sci.* 29, 375–419. doi: 10.1207/s15516709cog0000_25

Lewis, R., Vasishth, S., and Van Dyke, J. (2006). Computational principles of working memory in sentence comprehension. *Trends Cogn.*

Sci. 10, 447–454. doi: 10.1016/j.tics.2006.08.007

Lin, J.-W. (1996). *Polarity Licensing and WH-Phrase Quantification in Chinese*, Ph.D. dissertation. Amherst, MA: University of Massachusetts.

Linebarger, M. (1980). *The Grammar of Negative Polarity*. PhD dissertation. Cambridge, MA: MIT.

Linebarger, M. (1987). Negative polarity and grammatical representation. *Linguist. Philos.* 10, 325–387. doi: 10.1007/BF00584131

McElree, B. (2001). Working memory and focal attention. *J. Exp. Psychol. Learn. Mem. Cogn.* 27, 817–835. doi: 10.1037/0278-7393.27.3.817

McElree, B., Foraker, S., and Dyer, L. (2003). Memory structures that subserve sentence comprehension. *J. Mem. Lang.* 48, 67–91. doi: 10.1016/S0749-596X(02)00515-6

Nieuwland, M. S., and Kuperberg, G. R. (2008). When the truth is not too hard to handle. *Psychol. Sci.* 19, 1213–1218. doi: 10.1111/j.1467-9280.2008.02226.x

Nieuwland, M. S., Ditman, T., and Kuperberg, G. R. (2010). On the incrementality of pragmatic processing: an ERP investigation of informativeness and pragmatic abilities. *J. Mem. Lang.* 63, 324–346. doi: 10.1016/j.jml.2010.06.005

Ozonoff, S., and Miller, J. N. (1996). An exploration of right-hemisphere contributions to the pragmatic impairments of autism. *Brain Lang.* 52, 411–434. doi: 10.1006/brln.1996.0022

Parker, D., MacMillan, G., and Phillips, C. (2013). "Illusory NPI licensing: now you see it, now you don't," in *Poster Presented at the 26th Annual CUNY Human Sentence Processing Conference* (Columbia, SC: University of South Carolina).

Parker, D., and Phillips, C. (2011). "Illusory negative polarity item licensing is selective," in *Poster Presented at the 24th Annual CUNY Human Sentence Processing Conference* (Stanford, CA).

Pearlmutter, N. J., Garnsey, S. M., and Bock, K. (1999). Agreement processes in sentence comprehension. *J. Mem. Lang.* 41, 427–456. doi: 10.1006/jmla.1999.2653

Sedivy, J. C., Tanenhaus, M. K., Chambers, C. G., and Carlson, G. N. (1999). Achieving incremental semantic interpretation through contextual interpretation. *Cognition* 71, 109–147. doi: 10.1016/S0010-0277(99)00025-6

Sikos, L., Tomlinson, S., Traut, H., and Grodner, D. (2013). "Incremental computation of scalar implicatures: an ERP study," in *Poster Presented at the 26th Annual CUNY Sentence Processing Conference* (Columbia, SC).

Stewart, M., and Ota, M. (2008). Lexical effects on speech perception in individuals with "autistic" traits. *Cognition* 109, 157–162. doi: 10.1016/j.cognition.2008.07.010

Tager-Flusberg, H., Paul, R., and Lord, C. E. (2005). "Language and communication in autism," in *Handbook of Autism and Pervasive Developmental Disorder, 3rd Edn,* eds F. Volkmar, R. Paul, A. Klin, and D. J. Cohen (New York, NY: Wiley), 335–364.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., and Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science* 268, 1632–1634. doi: 10.1126/science.7777863

van Berkum, J. J. A. (2009). "The neuropragmatics of 'simple' utterance comprehension: an ERP review," in *Semantics and Pragmatics: From Experiment to Theory*, eds U. Sauerland and K. Yatsushiro (Basingstoke: Palgrave Macmillan), 276–316.

Van Dyke, J. A., and McElree, B. (2006). Retrieval interference in sentence comprehension. *J. Mem. Lang.* 55, 157–166. doi: 10.1016/j.jml.2006.03.007

Van Dyke, J. A., and Lewis, R. L. (2003). Distinguishing effects of structure and decay on attachment and repair: a retrieval interference theory of recovery from misanalyzed ambiguities. *J. Mem. Lang.* 49, 3, 285–316. doi: 10.1016/S0749-596X(03)00081-0

Vasishth, S., Drenhaus, H., Saddy, D., and Lewis, R. (2005). "Processing negative polarity," in *Talk Presented at the 18th CUNY Sentence Processing Conference* (University of Arizona).

Vasishth, S., and Drenhaus, H. (2011). Locality effects in German. *Dialogue Discourse* 2, 59–82.

Vasishth, S., Lewis, R. L., Brüssow, S., and Drenhaus, H. (2008). Processing polarity: how the ungrammatical intrudes on the grammatical. *Cogn. Sci.* 32, 685–712. doi: 10.1080/03640210802066865

Vasishth, S., and Broe, M. (2011). *The Foundations of Statistics: A Simulation-based Approach*. Berlin: Springer. doi: 10.1007/978-3-642-16313-5

Venables, W. N., and Ripley, B. D. (1999). *Modern Applied Statistics with S-PLUS*. NewYork, NY: Springer. doi: 10.1007/978-1-4757-3121-7

von Fintel, K. (1999). NPI-licensing, Strawson-entailment, and context-dependency. *J. Sem.* 16, 97–148. doi: 10.1093/jos/16.2.97

Wagers, M. W., Lau, E. F., and Phillips, C. (2009). Agreement attraction in comprehension: representations and processes. *J. Mem. Lang.* 61, 206–237. doi: 10.1016/j.jml.2009.04.002

Wang, A. T., Lee, S. S., Sigman, M., and Dapretto, M. (2006). Neural basis of irony comprehension in children with autism: the role of prosody and context. *Brain* 129, 932–943. doi: 10.1093/brain/awl032

Xiang, M., Dillon, B., and Phillips, C. (2006). "Testing the strength of the spurious licensing effect for negative polarity items," in *Talk Presented at the 19th Annual Meeting of the CUNY Conference on Human Sentence Processing* (New York, NY).

Xiang, M., Dillon, B., and Phillips, C. (2009). Illusory licensing effects across dependency types: ERP evidence. *Brain Lang.* 108, 40–55. doi: 10.1016/j.bandl.2008.10.002

Xiang, M., Giannakidou, A., and Grove, J. (2012). "Two stages of NPI licensing: an ERP study," in *Poster Presented at the 25th Annual Meeting of the CUNY Conference of Human Sentence Processing* (New York, NY).

Xiang, M., Grove, J., and Giannakidou, A. (2013). "Multiple dimensions of NPI licensing: evidence from ERPs," in *Talk Presented at the 5th Experimental Pragmatics conference (XPrag)* (Utrecht).

Yanilmaz, A., and Drury, J. (2013). "Licensing in Turkish: an ERP study," in *Poster Presented at the 20th Cognitive Neuroscience Society Meeting* (San Francisco, CA).

Yu, A. C. L. (2010). Perceptual compensation is correlated with individuals' "Autistic" traits: implications for models of sound change. *PLoS ONE* 5:e11950. doi: 10.1371/journal.pone.0011950

Zanuttini, R. (1991). *Syntactic Properties of Sentential Negation. A Comparative Study of Romance Languages*. University of Pennsylvania Dissertation.

Zeijlstra, H. H. (2004). *Sentential Negation and Negative Concord*. Ph D Dissertaion, University of Amsterdam.

Zwarts, F. (1995). Nonveridical contexts. *Linguist. Anal.* 25, 286–312.

Zwarts, F. (1986). *Categoriale Grammatica en Algebraische Semantiek. Een Studie Naar Negatie en Polariteit in het Nederlands*. Ph D thesis, University of Groningen.

Zwarts, F. (1996). "A hierarchy of negative expressions," in *Negation. A Notion in Focus*, ed H. Wansing (Berlin; New York: De Gruyter), 169–194. doi: 10.1515/9783110876802.169

Zwarts, F. (1998). "Three types of polarity," in *Plural Quantification*, eds F. Hamm and E. Hinrichs (Dordrecht: Kluwer), 177–238. doi: 10.1007/978-94-017-2706-8_5

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 14 June 2013; accepted: 16 September 2013; published online: 07 October 2013.

*Citation: Xiang M, Grove J and Giannakidou A (2013) Dependency-dependent interference: NPI interference, agreement attraction, and global pragmatic inferences. Front. Psychol. **4**:708. doi: 10.3389/fpsyg.2013.00708*

This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology.

Copyright © 2013 Xiang, Grove and Giannakidou. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

APPENDIX

The interactions between the other four sub-scales from AQ and each of the fixed effect predictors, for NPI licensing and number agreement separately. The mixed effect models are constructed in similar ways as the models in **Tables 3**, **5**, **7**, **8**. As shown below, the only significant effect observed is the interaction between the Social Skill sub-scale and the offline NPI interference effect (i.e., acceptance rate). This is similar to the effect of the Communication sub-scale. No other interaction was observed.

Table 9 |

	NPI licensing			Number agreement	
Acceptance rate (*p*-value)	**Interaction with Interference**	**Interaction with Licensor**	**Interaction with Grammaticality**	**Interaction with Interference**	**Interaction with Grammaticality**
Social skill	0.006**	0.14	0.89	0.16	0.96
Attention to details	0.44	0.17	0.92	0.16	0.73
Attention switching	0.98	0.9	0.63	0.13	0.56
Imagination	0.15	0.44	0.94	0.61	0.096
RT at the critical word (*t*-value)	**Interaction with Intrusion**	**Interaction with Licensor**	**Interaction with Grammaticality**	**Interaction with Intrusion**	**Interaction with Grammaticality**
Social skill	−1.28	0.65	0.97	0.43	0.56
Attention to details	0.58	0.89	−0.74	0.65	−1.02
Attention switching	−1.46	0.98	0.54	0.86	0.54
Imagination	0.13	0.57	0.07	0.54	0.17

$^{**}p < 0.01$.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read, for greatest visibility

COLLABORATIVE PEER-REVIEW
Designed to be rigorous – yet also collaborative, fair and constructive

FAST PUBLICATION
Average 85 days from submission to publication (across all journals)

COPYRIGHT TO AUTHORS
No limit to article distribution and re-use

TRANSPARENT
Editors and reviewers acknowledged by name on published articles

SUPPORT
By our Swiss-based editorial team

IMPACT METRICS
Advanced metrics track your article's impact

GLOBAL SPREAD
5'100'000+ monthly article views and downloads

LOOP RESEARCH NETWORK
Our network increases readership for your article

Frontiers
EPFL Innovation Park, Building I • 1015 Lausanne • Switzerland
Tel +41 21 510 17 00 • Fax +41 21 510 17 01 • info@frontiersin.org
www.frontiersin.org

Find us on