EXPERIMENTAL APPROACHES TO PRAGMATICS

EDITED BY: Valentina Cuccio, Pietro Robert Perconti, Gerard Steen, Yury Y. Shtyrov and Yan Huang

PUBLISHED IN: Frontiers in Psychology

frontiers Research Topics

frontiers

Frontiers eBook Copyright Statement

The copyright in the text of individual articles in this eBook is the property of their respective authors or their respective institutions or funders. The copyright in graphics and images within each article may be subject to copyright of other parties. In both cases this is subject to a license granted to Frontiers.

The compilation of articles constituting this eBook is the property of Frontiers.

Each article within this eBook, and the eBook itself, are published under the most recent version of the Creative Commons CC-BY licence. The version current at the date of publication of this eBook is CC-BY 4.0. If the CC-BY licence is updated, the licence granted by Frontiers is automatically updated to the new version.

When exercising any right under the CC-BY licence, Frontiers must be attributed as the original publisher of the article or eBook, as applicable.

Authors have the responsibility of ensuring that any graphics or other materials which are the property of others may be included in the CC-BY licence, but this should be checked before relying on the CC-BY licence to reproduce those materials. Any copyright notices relating to those materials must be complied with.

Copyright and source acknowledgement notices may not be removed and must be displayed in any copy, derivative work or partial copy which includes the elements in question.

All copyright, and all rights therein, are protected by national and international copyright laws. The above represents a summary only. For further information please read Frontiers' Conditions for Website Use and Copyright Statement, and the applicable CC-BY licence.

ISSN 1664-8714
ISBN 978-2-88976-132-6
DOI 10.3389/978-2-88976-132-6

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: frontiersin.org/about/contact

EXPERIMENTAL APPROACHES TO PRAGMATICS

Topic Editors:
Valentina Cuccio, University of Messina, Italy
Pietro Robert Perconti, University of Messina, Italy
Gerard Steen, University of Amsterdam, Netherlands
Yury Y. Shtyrov, Aarhus University, Denmark
Yan Huang, The University of Auckland, New Zealand

Citation: Cuccio, V., Perconti, P. R., Steen, G., Shtyrov, Y. Y., Huang, Y., eds. (2022). Experimental Approaches to Pragmatics. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-88976-132-6

Table of Contents

05 ***Editorial: Experimental Approaches to Pragmatics***
Valentina Cuccio, Pietro Perconti, Gerard Steen, Yury Shtyrov and Yan Huang

08 ***Pragmatics Always Matters: An Expanded Vision of Experimental Pragmatics***
Raymond W. Gibbs Jr. and Herbert L. Colston

16 ***Complex Inferential Processes Are Needed for Implicature Comprehension, but Not for Implicature Production***
Irene Mognon, Simone A. Sprenger, Sanne J. M. Kuijper and Petra Hendriks

34 ***The Role of Literal Features During Processing of Novel Verbal Metaphors***
Camilo R. Ronderos, Ernesto Guerra and Pia Knoeferle

59 ***Children's and Adults' Sensitivity to Gricean Maxims and to the Maximize Presupposition Principle***
Francesca Panzeri and Francesca Foppolo

69 ***Context Breeds False Memories for Indeterminate Sentences***
Levi Riven and Roberto G. de Almeida

86 ***Information Structure and Word Order Canonicity in the Comprehension of Spanish Texts: An Eye-Tracking Study***
Carolina A. Gattei, Luis A. París and Diego E. Shalom

102 ***Conceptual Conflicts in Metaphors and Pragmatic Strategies for Their Translation***
Ilaria Rizzato

110 ***The Role of Working Memory in the Processing of Scalar Implicatures of Patients With Schizophrenia Spectrum and Other Psychotic Disorders***
Walter Schaeken, Linde Van de Weyer, Marc De Hert and Martien Wampers

122 ***Exposure to Literary Fiction Is Associated With Lower Psychological Essentialism***
Emanuele Castano, Maria Paola Paladino, Olivia G. Cadwell, Valentina Cuccio and Pietro Perconti

132 ***Child-Like Adults: Dual-Task Effects on Collective vs. Distributive Sentence Interpretations***
Anna M. B. de Koster, Petra Hendriks and Jennifer K. Spenader

148 ***The Double Framing Effect of Emotive Metaphors in Argumentation***
Francesca Ervas, Maria Grazia Rossi, Amitash Ojha and Bipin Indurkhya

172 ***The Cost of the Epistemic Step: Investigating Scalar Implicatures in Full and Partial Information Contexts***
Maria Spychalska, Ludmila Reimer, Petra B. Schumacher and Markus Werning

198 ***Processing Non-at-Issue Meanings of Conditional Connectives: The* wenn/falls *Contrast in German***
Mingya Liu

216 ***Referential Chains Reveal Predictive Processes and Form-to-Function Mapping: An Electroencephalographic Study Using Naturalistic Story Stimuli***
Ingmar Brilmayer and Petra B. Schumacher

232 ***Face Management and Negative Strengthening: The Role of Power Relations, Social Distance, and Gender***
Nicole Gotzner and Diana Mazzarella

245 ***A Study of Zero Anaphora Resolution in Chinese Discourse: From the Perspective of Psycholinguistics***
Ning Yang, Jingyu Zhang, Lijun Ma and Zhi Lu

260 ***Correction by Focus: Cleft Constructions and the Cross-Linguistic Variation in Phonological Form***
Markus Greif and Stavros Skopeteas

280 ***Persistence Conditions of Institutional Entities: Investigating Copredication Through a Forced-Choice Experiment***
Elliot Murphy

frontiers | Frontiers in Psychology

EDITORIAL
published: 19 April 2022
doi: 10.3389/fpsyg.2022.865737

Editorial: Experimental Approaches to Pragmatics

Valentina Cuccio[1]*, *Pietro Perconti*[2], *Gerard Steen*[3], *Yury Shtyrov*[4] and *Yan Huang*[5]

[1] Department of Ancient and Modern Civilizations, University of Messina, Messina, Italy, [2] Department of Cognitive Sciences, Psychology, Education and Cultural Studies, University of Messina, Messina, Italy, [3] Language and Communication, Faculty of Humanities, University of Amsterdam, Amsterdam, Netherlands, [4] Center for Functionally Integrative Neuroscience, Institute for Clinical Medicine, Aarhus University, Aarhus, Denmark, [5] Faculty of Arts, The University of Auckland, Auckland, New Zealand

Keywords: experimental pragmatics, neuropragmatics, context, clinical pragmatics, figurative language

Editorial on the Research Topic

Experimental Approaches to Pragmatics

INTRODUCTION

Often the starting point of the study of the biological bases of language is the question: How is language represented in the brain? This question might suggest that linguistic meaning is a form of knowledge stored in the human brain that we retrieve when we listen to or read words and sentences. Undoubtedly, we do have a knowledge of language which is represented in neural networks in the perisylvian cortex. However, the way this question is formulated can be misleading. A better way to rephrase it would be: how meaning is constructed in the brain? The difference between the first and the second question is the model of language they imply. The first question presupposes that language is a code whereas the second suggests that linguistic meaning is always the result of a contextually based process of interpretation. Many linguistic phenomena, such as metaphors, irony and other forms of figurative meaning, could hardly be explained if we defined language as a code. A pragmatic approach is, thus, fundamental if we aim at providing a full account of language processing. We need to explain how we use symbols and how we make meanings out of them. And we need to do so in a psychologically and neurologically plausible framework.

Since the pioneering work of Wittgenstein and Grice, Pragmatics, the study of how language is used in context, has been traditionally addressed by philosophers and linguists from a theoretical perspective. However, classic pragmatic notions such as communicative intentions, implicatures or usage-based meaning must now be understood in light of a psychological and neural account of language. Thus, today, Pragmatics is a highly interdisciplinary enterprise that is investigated by psychologists, neuropsychologists and neuroscientists as well as philosophers and linguists. Recently, this experimental approach to the pragmatics of language has gathered momentum and has given rise to the birth of a new field of study: Experimental Pragmatics (Cuccio, 2022; Gibbs and Colston for an overview). This refers to a set of different but strictly interrelated disciplines: *Neuropragmatics*, which aims at identifying the neural infrastructures underlying pragmatic processes in language production/comprehension; *Clinical Pragmatics*, which aims at studying pragmatic disorders in clinical populations and *Experimental Pragmatics stricto sensu*, which aims at empirically validate theoretical accounts of the pragmatic of language by means of behavioral experiments.

With no ambition to provide the precise geography of this research field, we can say that *Experimental Pragmatics stricto sensu* is, no doubt, the discipline most represented in this collection. Thirteen papers out of eighteen accepted for publication in the Research Topic *Experimental*

OPEN ACCESS

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

Reviewed by:
Igor Douven,
Université Paris-Sorbonne, France

***Correspondence:**
Valentina Cuccio
vcuccio@unime.it

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

Received: *30 January 2022*
Accepted: *08 March 2022*
Published: *19 April 2022*

Citation:
Cuccio V, Perconti P, Steen G, Shtyrov Y and Huang Y (2022) Editorial: Experimental Approaches to Pragmatics. Front. Psychol. 13:865737. doi: 10.3389/fpsyg.2022.865737

approaches to Pragmatics investigated crucial theoretical issues in the pragmatics of Language by means of behavioral studies and surveys. Two theoretical contributions (Gibbs and Colston; Rizzato) provided a critical overview of current perspectives in the experimental approaches to pragmatics with a specific focus on metaphors. One paper (Schaeken et al.), investigating the role of working memory in the processing of scalar implicature in Schizophrenic patients, was framed within the discipline of *Clinical Pragmatics*. Two papers (Brilmayer and Schumaker; Spychalska et al.) used electrophysiological measures in two Event Related Potential (ERP) experiments to investigate, respectively, scalar implicature in full and partial information context and the relationship between referential chains and predictive processes. These ERP studies were carried out in the framework of *Neuropragmatics*. The latter seems to be still the most underrepresented branch of investigation in the study of the neural correlates of language.

PRAGMATICS AND THE NEURAL CORRELATES OF LANGUAGE

A Place for the Pragmatics of Language

Experimental pragmatics is today a burgeoning field of study with a very lively debate (Gibbs and Colston). However, the research about the brain areas underpinning the pragmatic processing involved in non-literal usages of language (i.e., Neuropragmatics; see Bambini, 2010; Bara and Bara, 2010; Haggort and Levinson, 2014) is relatively recent compared to the study of the neural correlates of syntax and semantics.

To understand the reasons of this gap we need to acknowledge that in the second half of the 19° century and for a long time in the Neuroscience of language, Pragmatics was not even considered as one of the levels, along with phonology/orthography, syntax and semantics, to be taken into consideration when exploring the neural correlates of language. Linguists did not recognize the pragmatic dimension of language until the work of philosophers of language such as Wittgenstein and Anscombe (1953), Austin and Urmson (1962), and Grice (1989). And, when Pragmatics was finally introduced in the theoretical study of language, in the second half of the 20° century, it was first considered as a far less important feature compared to syntax and semantics. In fact, as Mey (2001) clearly explains in his introduction to Pragmatics, it was first considered as the "waste-basket of semantics," a place where linguists were used to relegate problematic aspects of language, such as its figurative usages, which they could hardly explain in semantic theories. Thus, Pragmatics struggled to find its identity and its own place in Linguistics (Mey, 2001). And if linguists for a long time did not sufficiently consider Pragmatics, so did, later, the neuroscientists working on the identification of the anatomical bases of language. For this reason, Neuropragmatics, compared to the study of the neurobiology of syntax and semantics, is the most recent branch of Neurolinguistics.

Furthermore, the possibilities of experimental pragmatics have long been undermined by the difficulties of modeling context dependence in an adequate way. Formal semantics has made many important contributions by attempting to bring to light the logic underlying the fact that the meaning of words and sentences often seems to depend on the context of production and evaluation of the linguistic act we are considering. On the whole, however, the adventure of formal semantics and, in particular, the attempt to provide a satisfying logical model of the phenomenon of context dependence has not been a success. The problem is that, if we are without a logical model of context dependence, quantitative research has a poor basis, and with it the very possibilities of doing experimental research in the field of pragmatics.

From Aphasiology to the Contemporary Cognitive Neuroscience of Language

The beginning of the identification of the neural structures subserving language dates back to the second half of the 19° century, when the development of aphasiology made possible the first description of the brain areas underlying the processing of language. Broca (1861) discovered that language is lateralized to the left hemisphere and identified a region in the frontal lobe, the pars triangularis of the Inferior Frontal Gyrus (IFG—Brodmann area 45) which seemed to be responsible for language production. A few years later, Wernicke (1874) identified in the temporal lobe another area linked to language, the Superior Temporal Gyrus (STG—Brodmann area 22), which, in turn, seemed to be related to the comprehension of language. On these bases, Wernicke (1874) proposed a first model of the brain mechanisms underlying both language production and comprehension, which was then further developed by Lichtheim (1885) and, in the second half of the 20° century, renewed by Norman Geschwind. The Wernicke-Lichtheim model, also known as the Wernicke-Geschwind model, for the processing of language has been influential for a long time. Generally speaking, aphasiology certainly gave a fundamental impulse to the study of the brain bases of language. However, today models of language processing based on aphasiology have been largely revised (for a discussion, Kandel et al., 2013). Recent years have witnessed an enormous technological growth. Functional brain imaging research allowed us to study *in vivo* the brain of both healthy subjects and patients with language impairments while these perform linguistic tasks. Techniques such as the functional Magnetic Resonance Imaging (fMRI) or the magnetoencephalography (MEG) provided us with the possibility to observe the neural mechanisms underlying the processing of language. On these bases, more complex models of the functional neuroanatomy of language have been proposed (e.g., Hickok and Poeppel, 2004, 2007; Friederici and Gierhan, 2013). Today we know that several systems underpin the processing of language and that the language network is far more sophisticated and extended than it was first believed. Broca's and Wernicke' area are still considered as the cornerstone of this network but they are functionally characterized in a partially different way. In fact, these brain areas not only subserve production and comprehension of language, as it was first believed on the basis of neurological data. In the field of language processing, they are today mainly characterized in

terms of their involvement, respectively, in the processing of syntax and semantics. Furthermore, we know that the arcuate fasciculus, that links Broca and Wernicke areas, previously considered to be unidirectional, conveying information from Wernicke to Broca, is bidirectionally linked to these two brain regions (Hickok and Poeppel, 2004, 2007). Most importantly, Broca and Wernicke areas are also connected through two other streams of information, beyond the arcuate fasciculus: (i) a ventral stream, bilaterally distributed in the brain, has been identified in the superior and middle temporal lobes, although with some differences in the recruitment of the left and right hemispheres. This ventral stream processes speech signals for language comprehension (i.e., it maps sounds to meanings, according to Hickok and Poeppel, 2004, 2007 model); (ii) a dorsal stream, lateralized to the left hemisphere, includes structures in the posterior frontal lobe and in the posterior-dorsal area of the temporal lobe. This dorsal stream maps acoustic representations of language to articulatory networks (i.e., it maps sounds to articulatory gestures; for a discussion of the model, Hickok and Poeppel, 2004, 2007). In addition to this, other regions linked to language processing have also been identified in the perisylvian cortex (see Kandel et al., 2013 for anatomical and functional description of these areas). Furthermore, techniques with high temporal resolution, such as, for example, the MEG or the electroencephalography (EEG), also gave us the possibility to investigate the neural time course of language processing, which is of paramount importance to develop a model of how we produce and comprehend language since language production/comprehension is a multilayered process where different kinds of information need to be handled.

Thanks to this enormous technological growth and, most of all, to the introduction of more fine-grained models of language use, today we know that the neural network recruited by the processing of context-based meaning is bilaterally represented in the brain and it includes regions such the Inferior Frontal Gyrus (IFG; left and right BA 45, left BA 47), the Temporal-Parietal Junction (TPJ, right and left BAs 22 and 39), the right Anterior Cingulate Cortex (ACC, right BAs 24 and 32) and the right dorsolateral Prefrontal Cortex (DLPFC), specifically the middle frontal gyrus (MFG, right BA 9).

Many questions are open in the experimental study of the pragmatics of Language. Experimental research in this research field is intense today. To understand the psychological and neural processes subserving our ability to use language in context is of paramount importance for a better comprehension of many clinical conditions and, most of all, for a deeper understanding of what makes us human.

AUTHOR CONTRIBUTIONS

Although all the authors discussed and designed the article together, Sections Introduction and From Aphasiology to the Contemporary Cognitive Neuroscience of Language were written by VC. Section A Place for the Pragmatics of Language was written by VC and PP. All authors contributed to the article and approved the submitted version.

REFERENCES

Austin, J. L., and Urmson, J. O. (1962). *How to Do Things With Words*. Oxford: Clarendon Press.

Bambini, V. (2010). Neuropragmatics: a forward. *Riv. Ling.* 22, 1–20.

Bara, B. G., and Bara, B. G. P. (2010). *Cognitive Pragmatics: The Mental Processes of Communication*. Cambridge, MA: MIT Press.

Broca, P. P. (1861). *Sur le siege de la faculte du langage, avec deux observations d'aphemie, etc.* Paris: Extrait des Bulletins de la Societe anatomique de Paris.

Cuccio, V. (2022). "The figurative brain," in *The Routledge Handbook of Neurosemiotics*, eds. A. M. García and A. Ibáñez. London: Routledge.

Friederici, A. D., and Gierhan, S. M. (2013). The language network. *Curr. Opin. Neurobiol.* 23, 250–54. doi: 10.1016/j.conb.2012.10.002

Grice, H. P. (1989). *Studies in the Way of Words*. Cambridge, MA: Harvard University Press.

Haggort, P., and Levinson, S. (2014). "Neuropragmatics," in *The Cognitive Neurosciences*, 5th edition, eds. M. S. Gazzaniga and G. R. Mangun (Cambridge, MA: MIT Press), 667–674.

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

Kandel, E. R., Schwartz, J. H., Jessell, T. M., Siegelbaum, S., and Hudspeth, A. J. (2013). *Principles of Neural Science (5th ed.)*. New York, NY: McGraw-Hill.

Lichtheim, L. (1885). On aphasia. *Brain*, 7, 433–484.

Mey, J. (2001). *Pragmatics: An Introduction (2nd ed.)*. Oxford: Blackwell.

Wernicke, C. (1874). *Der aphasische Symptomencomplex*. Wernicke: Eine psychologische Studie auf anatomischer Basis.

Wittgenstein, L., and Anscombe, G. E. M. (1953). *Philosophical Investigations* (Eng. Trans. 2nd ed. 1968). Oxford: Blackwell.

Conflict of Interest: 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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Cuccio, Perconti, Steen, Shtyrov and Huang. 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) and the copyright owner(s) 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

CONCEPTUAL ANALYSIS
published: 24 July 2020
doi: 10.3389/fpsyg.2020.01619

Pragmatics Always Matters: An Expanded Vision of Experimental Pragmatics

Raymond W. Gibbs Jr.[1]* *and Herbert L. Colston*[2]

[1] Independent Researcher, Soquel, CA, United States, [2] Department of Linguistics, University of Alberta, Edmonton, AB, Canada

Much of the work in experimental pragmatics is devoted to testing empirical hypotheses that arise within the study of linguistic and philosophical pragmatics. The focus in much of this work is focused on those aspects of communicated meaning that are "inferred" rather than understood through linguistic "coding" processes. Under this view, pragmatic meanings emerge secondarily after purely linguistic meanings are accessed or computed. Our aim in this article is to greatly broaden the scope of experimental pragmatic studies by calling for much greater emphasis on the complete pragmatics of language use. Pragmatics is continuously present and constrains people's real-time production and processing of language in context. Experimental pragmatics should attend more to the particularities of pragmatic experience through closer examination of the people we study, the specific tasks used to assess understanding, as well as the actual complex meanings people interpret in diverse contexts. The many specifics of human pragmatics demand the study and theoretical inclusion of many bodily, linguistic, and situational factors that make up each instance of meaning making.

Keywords: pragmatics, experimental pragmatics, individual differences, task demands, psycholinguistics

OPEN ACCESS

Edited by:
Valentina Cuccio,
University of Messina, Italy

Reviewed by:
Francesca Marina Bosco,
University of Turin, Italy
Beate Hampe,
University of Erfurt, Germany

***Correspondence:**
Raymond W. Gibbs Jr.
raymondwgibbs@gmail.com

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

Received: *18 March 2020*
Accepted: *16 June 2020*
Published: *24 July 2020*

Citation:
Gibbs RW Jr and Colston HL (2020) Pragmatics Always Matters: An Expanded Vision of Experimental Pragmatics. Front. Psychol. 11:1619. doi: 10.3389/fpsyg.2020.01619

INTRODUCTION

Experimental pragmatics has had a complex history in its 40 or so years of existence. The field emerged back in the 1970s as various psychologists, both those studying developmental psychology and psycholinguistics, as well as linguistics, began to explore people's understandings of pragmatic meaning, which was quite a departure from the traditional emphasis in psycholinguistics on lexical, syntactic, and semantic processing of individual sentence meaning. Certain critics within linguistics and psychology were skeptical about the possibility of scientifically examining pragmatic language production and interpretation. One often repeated refrain from the 1970s and 1980s was that "pragmatics is the wastebasket of linguistics," a claim that suggests the impossibility of making proper scientific order out of a human endeavor which is so messy and intractable. Still, psycholinguists found much inspiration, and even testable hypotheses, in the writings of linguists and philosophers interested in pragmatics (Clark, 1996; Noveck and Sperber, 2004; Bara, 2010; Noveck, 2018; Gibbs, 2019). The field of experimental pragmatics has continued to survive, and make its mark, within the larger interdisciplinary world of cognitive science.

Many practitioners of experimental pragmatics see their work as explicitly devoted to testing the claims of those studying linguistic and philosophical pragmatics. A tremendous body of experimental work has spoken positively and negatively about different facets of various linguistic pragmatic theories (Noveck, 2018; Huang, 2019). One lingering assumption in much

experimental pragmatics research is the idea that "pragmatics" refers, somewhat narrowly, to those aspects of linguistic processing that are inferential, and not due to temporarily earlier linguistic coding/decoding processes. Under this view, people begin understanding what speakers mean by first engaging in many fast-acting linguistic processes in which sounds are recognized and then syntactic and semantic analyses are completed. Pragmatic meaning is created later on via special pragmatic inferential processes that may be generally applied to all utterances or are optionally applied given specific forms of linguistic input (e.g., different processes are needed to determine metaphor as opposed to ironic speaker meaning) (e.g., the standard pragmatic model, see Gibbs, 1994). Noveck (2018) argues that part of this view is motivated by ideas about modularity within cognitive science, more generally.

A related emphasis in experimental pragmatics is on the role that "theory of mind" or "mind-reading" plays in pragmatic language interpretation (Noveck, 2018). The focus here has been to explore the ways that understanding what people say or write depends on creating a theory of that person's mind, or specific thoughts in some communicative situation (Nichols and Stich, 2003). Experimental studies on theory of mind in pragmatic interpretation have examined a number of ways that people's cognitive abilities, and sometimes inabilities, to infer speakers' possible mental states are a critical facet of interpersonal communication (Kissine, 2016; Bosco et al., 2018). Some pragmatic theories go so far as to suggest that there is a "relevance theoretic comprehension procedure" module that is embedded within a larger "theory of mind" module (Sperber and Wilson, 2002).

Our argument in this article is that these traditional views on pragmatic meaning, despite their contributions to experimental pragmatics, under-estimate the true, and complex reality of pragmatic meaning making. We maintain that experimental pragmatics should be more than the testing of ideas from linguistic pragmatic theory. Experimental investigations must pay much greater attention to the larger ways that pragmatics always shapes our use and understanding of both linguistic and non-linguistic meanings, as seen in research on multimodal communication (Shockley et al., 2009; Hollers and Levinson, 2019). Pragmatics is much greater than the study of particular inferential processing stages, because people are always doing pragmatics within each moment of their lives. This includes people's pragmatic participation in experimental studies. We suggest the need for an expanded vision of experimental pragmatics, one that extends more deeply into the different ways that our doing pragmatics shape experimental participants' performances. Pragmatics is not merely a specific type of inferential processing, and it is not just a type of knowledge that differs from that accessed during various parts of language production and processing (e.g., lexicon, grammar, and semantics). Pragmatics is more fundamentally the entirety of people's adaptive performances in varying circumstances and contexts.

This article discusses several research practices within the field of experimental pragmatics over the last few decades. Our aim is not to criticize particular people. Both of us have engaged in some of the practices we take issue with in what follows. Some readers may also suggest that the situation we outline is not as bad as we make it out to be. Our aim, though, is to encourage discussion and debate in order to move experimental pragmatics studies forward to more adequately addressing "pragmatics" in a broader, psychologically real, fashion than it has been in the past.

THE PROBLEM

Experimental pragmatics studies typically explore what kinds of pragmatic processing emerges at what points during people's use and interpretation of language. Early theories in the field often assumed that pragmatic knowledge and inferential processes were recruited relatively late in the understanding process, especially when compared to the access of other sources of linguistic information (e.g., lexical, syntactic, and semantic) (see Gibbs, 1994; Gibbs and Colston, 2012). But the strong trend in experimental findings over the last several decades shows that pragmatic knowledge and pragmatic inferences comes into play very early during the online interpretation of language in context (Gibbs, 1994, 2019; Noveck and Sperber, 2004). People do not perform purely linguistic analyses first on a word string and only later recruit pragmatics to infer what speakers/writers aim to communicate. Instead, pragmatics has its influence through the immediate, automatic construction of what people imply by the words they speak and write (Gibbs, 1994; Gibbs and Colston, 2012). Pragmatics does not come into play only at certain temporal points in language use, and is not turned on and off in people's linguistic and non-linguistic experiences. Theoretical models in psycholinguistics now mostly embrace the idea that pragmatics, often through access to prior pragmatic background knowledge and more proximate contextual information, constrains all facets of the understanding process (Campbell and Katz, 2012; McRae and Matsuki, 2013; McClelland et al., 2014).

Our concern, however, is with two unacknowledged assumptions in the traditional study of experimental pragmatics. First, there is surprisingly little discussion of what it really means to say that some pragmatic message (e.g., "This soup needs salt" implies "Pass me the salt") has been "understood." Pragmatic understanding is assumed to be a general goal that all people in all contexts aim to achieve. But people differ in their cognitive and personal make-up, as well as their understanding motivations, in various circumstances. These individual variations, both between and within people, are critical to take into account in any theoretical characterization of how people interpret pragmatic messages.

Second, experimental pragmatics examines people's language understanding abilities by asking participants to perform a wide range of experimental tasks. These task demands constitute a big part of the inherent pragmatics within any experimental study (e.g., developmental studies have long struggled with how implicit and explicit task demands affect behavioral outcomes in cognitive and linguistic studies). Yet this aspect of pragmatic experience is not sufficiently acknowledged in scholars' theoretical interpretations of experimental results

within psycholinguistics and cognitive neuroscience. As is often the case in experimental studies of human perception and cognition, we too often strip away the task demands in creating theories of pragmatics as if this critical feature of experimental studies is irrelevant to characterizing the role that pragmatics has in people's use and understanding of language in context.

In addition to these difficulties, there is also the problem that experimental pragmatics focuses mostly on the "processes" by which language is acquired, produced, and understood, but is far less dedicated to explaining meaning "products" that people really convey or interpret in real-world language situations. The relative neglect of pragmatic "products" in experimental pragmatics comes with a great cost. We too often assume that people experience a definitive "click of comprehension" when pragmatic messages are singularly encountered and understood. Yet this mistakenly assumes that experimental pragmatics should focus on the use and understanding of different types of pragmatic meanings (e.g., scalar implicatures, presuppositions, politeness, negation, and metaphor), but not the very specific tokens of meaning that people may often infer in discourse. This difficulty also alerts us to the need to significantly broaden our vision of pragmatics by looking more closely at what participants are fully engaged in during different experimental situations.

INDIVIDUAL DIFFERENCES

Most theories within linguistic pragmatics offer detailed proposals on the ways ordinary people use and understand pragmatic meanings (Huang, 2019). These theoretical proposals typically assume some idealized speaker/hearer who is an adult possessing relatively intact neural, cognitive and linguistic abilities. Of course, there is an extensive body of research looking at variations in pragmatic language talents, such as children who are still acquiring pragmatic language skills, and atypical children and adults who may be limited because of brain injury, disease (e.g., Alzheimer's) or developmental disorders (e.g., autism) (Cummings, 2019). The classic assumption, nonetheless, is that differences in pragmatic language performances are mostly evidence of pragmatic deficits in which the typical, normative module of pragmatic competence is not functioning as expected.

But there exists a range of evidence showing important individual differences that shape pragmatic performances in experimental pragmatic studies. For example, there is an emerging body of research showing many variations within, and between, experimental participants. Consider some of the individual differences that have been empirically shown to influence figurative language use and understanding, including language experience, gender, occupation, social status and culture, political background/beliefs, cognitive differences (e.g., IQ, working memory capacity), bodily action, geographic origin, personality, social relationship, and common ground (Gibbs and Colston, 2012). These factors have their assorted influences on both the processing of figurative language, such as metaphor and irony, and the exact meaning products people infer when they encounter different tropes in various experimental situations.

Many scholars in experimental pragmatics may argue that it should be possible to control for, or factor away, individual variations in order to create normative theories of pragmatic language abilities without regard to complex arrays of individual differences. Our reply is that trying to control for, and then eliminate the need to account for, individual differences turns a blind eye to the real complexities of pragmatic experiences. Individual differences are not mere representations of "noise" around some normative mechanism of pragmatic meaning understanding. The fact of the matter is that individual differences always have a critical role in the psychology of pragmatic behaviors.

There are also within-individual variations that affect pragmatic performances in experimental situations. For example, a typical study in experimental pragmatics will present individual participants a set of stimuli, representing different independent variables, which they will respond to in some instructed manner. We often compute averages of people's behavioral performances across the many stimuli in each experimental condition. The aim here is to capture something about the central tendencies in people's reactions to different experimental conditions and looking at means is widely viewed as the most appropriate descriptive statistic by which to achieve this goal.

But means or averages hide the fuller complexity of people's pragmatic behaviors in experimental studies. There is a good deal of work within experimental psychology that demonstrates how individual people's in-experimental performances vary in systematic ways (Raczaszek-Leonardi and Kelso, 2007; Gibbs and Van Orden, 2010). Looking at the distributions of responses, such as reading times, can offer more insightful explanations for people's experimental performances, including the idea that people are behaving as self-organizing dynamical systems within the experiment (Gibbs and Van Orden, 2010; Gibbs, 2017). For this reason, we must be careful not to assume, as is too often done, that the independent variable must only be caused by a specific, isolated mechanism in mind (e.g., pragmatic competence). Many independent variables may only have partial, probabilistic influence on people's behaviors in experimental pragmatic tasks (Gibbs and Santa Cruz, 2012).

Our point is that the data obtained in experimental pragmatic studies do not simply reflect people's responses to different experimental conditions and the independent variables these are meant to tap into. Instead, people's individual pragmatic behaviors in any experimental situation are subtly shaped by their specific bodies, cultural expectations, personalities, and histories (Paxton and Dale, 2017; Abney et al., 2018). Pragmatics is, in this way, always a part of experiments we conduct and the data obtained from these investigations.

EXPERIMENTAL TASKS

It is challenging to characterize the diversity of tasks employed in experimental pragmatics (Jucker et al., 2018). Nonetheless, a typical study in experimental pragmatics will present participants with a set of stimuli to which they are to respond in one of many possible ways. Among the most widely used

experimental techniques are full-sentence reading times, word-by-word reading times (including both moving-window and eye-movement measures), self-paced listening, paraphrase judgment response times, priming methods, mouse-tracking, eye-tracking in visual world environments, free recall, cued recall, mental imagery studies, summarization and paraphrase of meaning tasks, question answering, cooperative conversation tasks, bodily enactment tasks, and various brain scanning measures such as evoked-related potentials (ERPs) and functional magnetic resonance imaging (fMRI) methods.

Each of these experimental techniques presumably taps into how people "understand" pragmatic meaning. But these measures reflect different facets of pragmatic understanding. For example, full phrase or sentence reading time studies offer evidence on the total cognitive effort required to interpret a particular kind of pragmatic meaning at the phrasal or sentence level, such as a figurative utterance (e.g., metaphor, idiom, and irony) or other kinds of conversational implicature (e.g., scalar implicature). Methods examining the time it takes people to read individual words in linguistic expressions conveying different kinds of pragmatic meaning, via moving-window or eye-movement techniques, are useful for exploring local processing of specific word meanings in context. These online techniques, along with brain scanning measures such as ERPs, provide insights into the interaction of linguistic, social/pragmatic and cognitive knowledge during real-time pragmatic language understanding. Asking people to paraphrase the meanings of different pragmatic messages, rapidly judge suggested paraphrases of utterance meaning, or engage in specific task-related conversations provide evidence that enables scholars to characterize the meaning products understood when people process pragmatic meanings. Similarly, imagery tasks provide another method for exploring the contents of what people have understood having just quickly read or heard a specific kind of pragmatic message. Bodily engagement tasks, where people are asked to perform specific gestures or adopt different postures, are critical for investigating the role of embodied experience and action in creating pragmatic understandings of words, phrases, and longer stretches of discourse.

In general, no single method is capable of examining all facets of pragmatic understanding. Each technique may reveal different aspects of what happens during people's inferring of pragmatic meanings. In some cases, these insights into pragmatic language processing are specific to particular temporal dimensions of the online construction of pragmatic meaning. For instance, word-based processing measures aim to assess more local pragmatic processing as experimental participants read or listen to linguistic messages word-by-word. Full-time reading and priming tasks are better able to assess more global aspects of pragmatic meaning understanding, such as when an overall message is understood (e.g., does this phrase, sentence, in context convey metaphorical meaning or a specific scalar implicature?).

Our concern here is that there is still an overwhelming tendency in the literature for scholars to make generalizations from their task-specific studies to larger, comprehensive theories of pragmatics. A vast number of studies on figurative language use employs an extensive range of experimental methods in which participants are instructed to engage in different tasks, such as fast, word-by-word reading, full phrase or sentence reading, making quick judgments on whether a particular figurative utterance makes sense, or fits into the previously read story context, or determine if an utterance conveys literal or some kind of figurative meaning (e.g., metaphorical and ironic), and whether a figurative utterance is apt or creative (Gibbs and Colston, 2012; Colston, 2015). Each of these dependent measures may affect participants' "understanding" performances in experimental situations given the different forms of attention they must pay to the stimulus materials. The results of these varying studies, and the theoretical interpretations scholars offer for explaining these findings, will differ depending on the explicit task required of the participants in a study. Yet these task influences are rarely acknowledged in linguistic pragmatic theories.

One possible response to this concern is to place most credibility in those experimental findings that converge across different experimental tasks (i.e., converging operations) (Gibbs, 2019). But it may still be difficult, if not impossible, to find experimental results that are truly universal across various people, languages, cultures, and task demands (Kecskes, 2014). A related response would be to argue that those that have the *greatest* convergence across people and tasks should be given the most weight in theoretical debates. However, arguments based on the "weight" of empirical evidence may be far less satisfactory to scientists who demand reliability and consistency in experimental findings.

Another response to the task demand problem in experimental pragmatics is when individual scholars argue for the superiority of some task environments (e.g., measures of eye-movements) over others (e.g., full phrasal or sentence reading times). The arguments along this line typically suggest that some specific task measures are better indicators of "real-world" pragmatic language use than others. Experiments that employ those privileged methods should, under this view, be afforded the most weight in debates over the content of pragmatic theories. It is fair to observe, however, that this type of response to the task demand issue typically ends in complete empirical stalemates as different scholars merely embrace results from preferred methods while ignoring or dismissing findings obtained from less preferred experimental paradigms.

The alternative position that is part of our broader vision of experimental pragmatics suggests that pragmatic language use is always task-specific both in and outside of experimental studies. Pragmatic language processing is not a uniform activity that operates in a task-free manner. Speakers and listeners always approach any language interaction or situation with explicit or implicit goals in mind. For instance, a listener can hear a political speech and wonder, even if implicitly, as to whether or not the message conveyed was persuasive, or whether or not he/she appreciated what a speaker has stated or an author wrote. People listen to language hoping to remember what was stated, in some circumstances, and may, therefore, pay close attention to the individual words and their meanings differently than when engaged in a very casual conversation. People's criteria for

understanding speakers' messages will greatly vary depending on the circumstances.

More generally, the time is ripe for scholars to incorporate task demands as an enduring part of any experimental pragmatic situation. Theories of pragmatics may need to be specifically tailored to the various tasks people perform in different experiments. It may be impossible to create comprehensive theories that supervene over experimental task demands. In this manner, the pragmatic constraints inherent in any experimental task offer another reason for claiming that pragmatics always matters in people's experiences of language use.

THE SUPERFICIALITY AND RICHNESS OF PRAGMATIC EXPERIENCE

Another challenge in conducting experimental pragmatic studies is that there are more complicated relations between task-dependent performances and pragmatic theories than are typically acknowledged. Consider a typical reading-time study that explores the cognitive effort required to understand pragmatic meaning, such as drawing a scalar implicature, inferring an ironic message, or quickly comprehending a novel metaphor. The reading time data are typically analyzed to test different hypotheses on the process by which people understand these different forms of pragmatic meaning.

However, we question whether people only infer a specific kind of meaning (e.g., literal vs. figurative, non-metaphorical vs. metaphorical, familiar metaphorical meaning vs. novel metaphorical meaning) when they read or hear language in discourse. Our motivations as readers, for example, are not simply centered on the recovery of a specific "meaning," but involve a vast assortment of human phenomenological experiences, such as drawing more context-specific pragmatic inferences, experiencing different emotional reactions or esthetic pleasures, or imagining what you, even as an isolated participant in an experiment, may say in response to what some other person has stated. Each of these impressions, reactions, and esthetic responses may be part of the total time it takes someone to read and understand, for example, a simple metaphorical phrase as having "metaphorical" and not "literal" meaning in context.

We often fail to appreciate people's pragmatic experiences of language in our quest to test specific hypotheses from linguistic pragmatics. To take one example, studies show that people take different times to interpret a metaphorical statement, such as "Lawyers are also sharks," depending on whether that expression is intended to simply affirm a pre-existing belief in some discourse, add new information, or contradict a previously asserted belief (Gibbs et al., 2011). People do not simply understand a metaphor as only expressing a metaphorical meaning, but interpret it more precisely in terms of its specific pragmatic messages in context (e.g., that a speaker wishes to strengthen an existing assumption, add new information, or contradict a previously stated belief about some topic).

A different example illustrates how the amount of effort devoted to processing a speaker's message depends on what meanings become most optimally relevant (Sperber and Wilson, 1995). For instance, reading the metaphorical phrase "My marriage is an icebox" takes longer to do in a context in which a speaker describes the state of his marriage than in a situation in which a speaker makes this reply to the question "Are you happy in your marriage? (Gibbs, 2010). The expectation set up by the prior question makes it unnecessary for readers to infer the many possible metaphorical meanings of "My marriage is an icebox" (e.g., my marriage is confining, emotionally cold, and not moving forward), precisely because the utterance quickly communicates a "no" answer to the prior "Are you happy in your marriage?" question.

Pragmatic "understanding" is not simply a matter recovering a particular type of meaning, as it also involves understanding what a speaker pragmatically, socially and esthetically intends to achieve by the use of some discourse. More attention to the exact pragmatic meanings people really infer, including their esthetic and emotional responses, in context will be an important part of broadening the vision of experimental pragmatics. We need to create experimental situations that systematically investigate when and how specific pragmatic messages are conveyed and inferred, as well as when vague, or less specific, meanings and attitudes are interpreted.

A CASE STUDY EXAMPLE

The experimental literature on pragmatic language use is enormously complex. As noted earlier, many studies offer conflicting findings in regard to how people pragmatically produce and interpret various aspects of communicative meaning. These profound variations in experimental outcomes relate to a broader concern within psychology and elsewhere, dubbed as the "replication crisis." Failures to replicate are now being published more than ever with some scholars claiming that any variation from some empirical standard should be interpreted as casting doubt on the validity of some earlier obtained experimental result (both for exact and conceptual replications) (Shrout and Rodgers, 2019).

We view the replication "crisis" in the behavioral sciences in a more positive light because it affords a perfect opportunity to explore all of the pragmatic nuances that shape human performances in different experimental studies. These replication problems are not problems at all, but concrete indications of how individual differences and task demands, for instance, are critical to explaining the experimental findings obtained, and why these factors are important to acknowledge in larger theories of human performance.

Consider the case of experimental research on irony understanding (Gibbs and Colston, 2007, 2012). There are many studies showing relatively fast understanding of ironic utterances in discourse, which suggests how pragmatic knowledge, of various sorts, quickly plays a role in people's online understanding of ironic meaning (e.g., Gibbs, 1986a,b; Ivanko and Pexman, 2003). At the same time, there is data suggesting that pragmatics comes in only later on during linguistic processing when irony is encountered (e.g., Giora, 2003; Filik and Moxey, 2010). There is also considerable research on the importance of cognitive abilities

related to mind-reading and executive functioning during both the learning and understanding of ironic speech and writing (e.g., Filippova and Astington, 2008).

How can we discriminate between those findings that are valid and worthy of theoretical consideration and those that are irrelevant? Replications efforts are important. Our point, though, is that replication attempts are not the solution to the diversity of experimental findings on irony comprehension, or any other pragmatic phenomena. It is far better to see the numerous experimental findings as pointing to many of the pragmatic nuances that really shape people's complex ironic language use. For instance, many studies show that the speed with which ironic utterances are understood may vary depending on whether the experiments assessed self-paced, full statement reading time, eye-movements in which regressions back to earlier text is allowed, word-by-word moving-window measures in which regressions are not possible, paraphrase judgment times, lexical decisions to words reflecting literal or ironic meanings, judgments over whether some phrase expressed irony or not, and so on. The stimuli used in these studies included variations in the length and syntactic complexity of ironic phrases, familiar vs. novel ironic statements, different forms of irony (e.g., blame by praise vs. praise by blame), different contextual circumstances (e.g., did the context set up an ironic situation, did the context provide an explicit echo to the irony mentioned), whether the ironic statements were addressed to participants or were participants overhearers of ironic exchanges, the accent in which an ironic utterance was spoken, cases where people had to make verbal responses to ironic phrases after quickly reading them, and so on. There are also individual differences between the experimental participants in these studies which include people with different ages, language backgrounds, cultural backgrounds, occupations, personality types, organic brain disorders and injuries, different cognitive abilities (e.g., working memory capacity, mind-reading abilities), and so on. These variations in the tasks and people studied in experiments on timed irony understanding have their individual effects, but also interact in many complex ways to reveal different emergent combinations of factors that may contribute to whether verbal irony is seen as easy or more difficult to interpret.

All of these varying empirical results are subject to exact and conceptual replication attempts (and some have been replicated in one form or another). But it seems unlikely that replication efforts will somehow clean up this catalog of experimental findings to reveal a simple, comprehensive set of data which clearly points to one theoretical model of irony understanding that can be applied to all people in all situations of verbal irony use. Nonetheless, the various, sometimes complex patterns of experimental results may highlight different systems of constraint that flexibly operate to produce relevant irony interpretations in different task-specific and people-specific contexts (e.g., constraint-satisfaction models, see Campbell and Katz, 2012; Caffarra et al., 2019).

Any instance of linguistic communication fundamentally constitutes a different task for the participants given their idiosyncratic histories, dispositions, and situations. No single task captures the complex underlying psychological reality when people encounter particular combinations of word strings or utterances. Each different configuration of task demands as task constraints requires a differently self-organized mind and body. The flexible capacity to self-organize to suit task constraints exists because mind and body compose a complex system. Specifically, the embodiment of task demands constrains the mind and body to anticipate task appropriate utterances in critical states and respond as needed within an experimental setting (e.g., timed comprehension responses).

Finally, virtually all experimental studies on irony comprehension, similar to many other areas of pragmatic meaning, assume that the final product of understanding is an "ironic" message. Yet these messages vary considerably in discourse, depending on a wide range of contextual and interpersonal factors. A person may hear "A fine friend you are!" in some situation and properly infer that the speaker was not making a compliment. But the exact interpretation created is usually much more than "You are not a good friend," and likely involves more specific meaning products, including that "the speaker had expected me to help him in my capacity as a good friend and was now scolding me with the hope that my future behaviors will be more cooperative." All of these more nuanced pragmatic effects may be understood as part of any simple behavioral response in an experimental situation (e.g., measuring eye-movements during reading of irony in written discourse). The future challenge is to assess the relations between task-specific experimental situations and the particular, in this case, ironic messages interpreted, along with the possible emotional and affective responses of people when reading, or listening to, ironic statements. Again, the inherent complexities among people and their explicit task requirements, as well as their implicit personal motivations, may all be constitutive of pragmatics when conducting experimental pragmatic studies.

CONCLUSION: EMBRACING A DIFFERENT THEORETICAL GOAL

These numerous challenges for experimental pragmatics may be overcome by adopting a broader vision for experimental pragmatics. There are several immediate steps toward a better understanding of the complexities of pragmatic language use.

First, researchers need to fully acknowledge the particular people they study and the implicit or explicit tasks presented to participants in experimental studies. There is no neutral point of view, no context-free, task-free environment from which utterance interpretation begins and eventually unfolds to produce pragmatic meanings. All language use is pragmatically situated from the early stages of linguistic processing, and theories of linguistic pragmatics must embrace this omnipresent reality. An experimental effect (i.e., the influence of an independent variable on a dependent variable) may be caused by a confluence of factors, most of which are not necessarily being manipulated within the context of a single study (e.g., individual differences, task demands, and the overall dynamical system that is created as a person performs in a specific task

environment) (Raczaszek-Leonardi and Kelso, 2007; Gibbs and Van Orden, 2010). Experimental psycholinguistics has obtained many important empirical findings demonstrating how various pragmatic knowledge (e.g., background knowledge, contextual information, and various cognitive abilities) shape ordinary language use (Clark, 1996; Gibbs, 2019). There is still a greater need to show how the pragmatic conditions within which experimental participants operate have their influence in different facets of linguistic communication.

Second, scholars need to more fully explore the meaning products that people create when they interpret pragmatic messages in different contexts given their different understanding of goals or tasks. People do not always understand utterances in the same way, as expressing the same meanings, a fact that is true both between and within people (e.g., a single person may infer different messages from the same utterance in the same context at different times) (e.g., "good enough language comprehension," see Ferreira and Patson, 2007). Our ultimate goal is to create a theory of pragmatics that is capable of generating the diverse meanings that people actually understand, not merely the idealized, and too often more socially and esthetically decontextualized, meanings that pragmatic theories typically discuss.

Pragmatic performances are not an isolated part of human behavior, divorced from other psychological processes and systems. People use utterances for various communicative purposes that are deeply connected with other bodily behaviors such as those responsible for tone of voice, eye-movement or gaze, laughter, bodily postures, hand and arm gestures, and so on. These bodily actions are all "coupled" in both time and space, as much cognitive science research indicates (Clark, 1996; Gibbs, 2006), to enable people to better coordinate and collaborate in order to achieve various personal and social goals (Gibbs, 2006; Shockley et al., 2009; Cols ton, 2019). Too much research in experimental pragmatics ignores these complex pragmatic realities when they analyze their data and go on to draw larger theoretical conclusions on the basis of the specific results they have obtained.

A general theory of pragmatics may also be characterized as part of a human dynamical system, not as its own isolated system (Gibbs, 2017). How people interpret utterances may, therefore, share many properties and processes that are related to many kinds of intentional human actions. The task that people explicitly or implicitly adopt when they produce and understand pragmatic messages, or the particular complex make-up of the participants in our studies, and the ways we analyze the full range of information that is obtained from participants are all part of the inherent pragmatic nature of human communication processes. We cannot, and should not, assume that there are ways of scrapping away the complexities in our experimental studies so that we can create a normative theory of pragmatics apart from the messy descriptive realities of real human performance.

Pragmatics is not just a temporally isolated inferential process that arises only at later points during real-life language use. Instead, pragmatics reflects the entire bodily system in action as people engage in different task-specific performances under the multiple influences of broader interpersonal, social, and cultural landscapes. Pragmatics is best understood as systems of varying constraints that have interactive influences on people's adaptive behaviors. This broader vision embraces the view that pragmatics *always* matters, to varying degrees, and must be acknowledged, and systematically investigated, within experimental pragmatic studies.

Our call for an expanded vision of experimental pragmatics is ultimately aimed at broadening what is considered to be "pragmatics" in contemporary theories of linguistic pragmatics. Linguists and philosophers, for example, may not see questions of individual differences and task demands as being relevant to their own respective writings on pragmatic theory. However, pragmatic theories should not be divorced from the pragmatic realities of human performances. Shouldn't these considerations of real people doing pragmatic actions be at the forefront of research and theory in linguistic pragmatics?

AUTHOR CONTRIBUTIONS

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

REFERENCES

Abney, D., Dale, R., Kello, C., and Louwerse, M. (2018). The burst and lulls of multimodal interaction: temporal distributions of behavior reveal differences between verbal and non-verbal communication. *Cogn. Sci.* 42, 1297–1316. doi: 10.1111/cogs.12612

Bara, B. (2010). *Cognitive Pragmatics*. Cambridge, MA: MIT Press.

Bosco, F., Tirassa, M., and Gabbatore, I. (2018). Why pragmatics and theory of mind do not (completely) overlap. *Front. Psychol.* 9:1453.

Caffarra, S., Motamed Haeri, A., Michell, E., and Martin, C. (2019). When is irony influenced by communicative constraints? ERP evidence supporting interactive models. *Eur. J. Neurosci.* 50, 3566–3577. doi: 10.1111/ejn.14503

Campbell, J., and Katz, A. (2012). Are there necessary conditions for inducing a sense of sarcastic irony? *Discourse Process.* 49, 459–480. doi: 10.1080/0163853x.2012.687863

Clark, H. (1996). *Using Language*. New York, NY: Cambridge University Press.

Colston, H. (2015). *Using Figurative Language*. New York, NY: Cambridge University Press.

Colston, H. L. (2019). *How Language Makes Meaning: Embodiment and Conjoined Antonymy*. New York, NY: Cambridge University Press.

Cummings, L. (2019). "Clinical pragmatics," in *Oxford Handbook of Pragmatics*, ed. Y. Huang (New York, NY: Oxford University Press), 346–361.

Ferreira, F., and Patson, N. (2007). The 'good enough' approach to language comprehension. *Lang. Linguist. Compass* 1, 71–83. doi: 10.1111/j.1749-818x.2007.00007.x

Filik, R., and Moxey, L. (2010). The on-line processing of written irony. *Cognition* 116, 421–436. doi: 10.1016/j.cognition.2010.06.005

Filippova, E., and Astington, J. (2008). Further development in social reasoning revealed in discourse irony understanding. *Child Dev.* 79, 126–138. doi: 10.1111/j.1467-8624.2007.01115.x

Gibbs, R. Jr., and Colston, H. (2012). *Interpreting Figurative Meaning*. New York, NY: Cambridge University Press.

Gibbs, R. (1986a). Comprehension and memory for nonliteral utterances: the problem of sarcastic indirect requests. *Acta Psychol.* 53, 41–57. doi: 10.1016/0001-6918(86)90004-1

Gibbs, R. (1986b). On the psycholinguistics of sarcasm. *J. Exp. Psychol. Gen.* 115, 3–15. doi: 10.1037/0096-3445.115.1.3

Gibbs, R. (1994). *The Poetics of Mind: Figurative Thought, Language, and Understanding*. New York, NY: Cambridge University Press.

Gibbs, R. (2006). *Embodiment and Cognitive Science*. New York, NY: Cambridge University Press.

Gibbs, R. (2010). "The wonderful, chaotic, creative, heroic, challenging world of researching and applying metaphor: a celebration of the past and some peeks into the future," in *Metaphor in the Real World*, ed. L. Cameron (Amsterdam: Benjamins), 3–17.

Gibbs, R. (2017). "Metaphor and dynamical systems," in *Routledge Handbook of Metaphor and Language*, eds V. Koller, E. Semino, and Z. Demjén (London: Routledge), 56–70.

Gibbs, R. (2019). "Experimental pragmatics," in *Oxford Handbook of Pragmatics*, ed. Y. Huang (New York, NY: Oxford University Press), 310–325.

Gibbs, R., and Colston, H. (eds) (2007). *Irony in Thought and Language: A Cognitive Science Reader*. Hillsdale, NJ: Erlbaum.

Gibbs, R., Okonski, L., and Tendahl, M. (2011). Inferring pragmatic messages from metaphor. *Lodz Pap. Pragm.* 7, 3–28.

Gibbs, R., and Santa Cruz, M. (2012). Temporal unfolding of conceptual metaphoric experience. *Metaphor Symb.* 27, 299–311. doi: 10.1080/10926488.2012.716299

Gibbs, R., and Van Orden, G. (2010). Adaptive cognition without massive modularity. *Lang. Cogn.* 2, 147–169.

Giora, R. (2003). *On Our Mind: Salience, Context, and Figurative Language*. New York, NY: Oxford.

Hollers, J., and Levinson, S. (2019). Multimodal language processing in human communication. *Trends Cogn. Sci.* 23, 639–652. doi: 10.1016/j.tics.2019.05.006

Huang, Y. (ed.) (2019). *Oxford Handbook of Pragmatics*. Oxford: Oxford University Press.

Ivanko, S., and Pexman, P. (2003). Context incongruity and irony processing. *Discourse Process.* 35, 241–279. doi: 10.1207/s15326950dp3503_2

Jucker, A., Schneider, K., and Bublitz, W. (eds) (2018). *Methods in Pragmatics*. Berlin: DeGruyter.

Kecskes, I. (2014). *Intercultural Pragmatics*. New York, NY: Oxford University Press.

Kissine, M. (2016). Pragmatics as meta-cognitive control. *Front. Psychol.* 6:2057.

McClelland, J., Mirman, D., Bolger, D., and Khaitan, P. (2014). Interactive activation and mutual constraint satisfaction in perception and cognition. *Cogn. Sci.* 38, 1139–1189. doi: 10.1111/cogs.12146

McRae, K., and Matsuki, K. (2013). "Constraint-based models of sentence processing," in *Current Issues in the Psychology of Language: Sentence Processing*, ed. R. van Gompel (New York, NY: Psychology Press), 51–77.

Nichols, S., and Stich, S. (2003). *Mindreading: An Integrated Account of Pretence, Self-Awareness, and Understanding Other Minds*. Oxford: Oxford University Press.

Noveck, I. (2018). *Experimental Pragmatics: The Making of a Cognitive Science*. New York, NY: Cambridge University Press.

Noveck, I., and Sperber, D. (2004). *Experimental Pragmatics.* London: Palgrave MacMillan.

Paxton, A., and Dale, R. (2017). Interpersonal movement synchrony responds to high-and-low conversational constraints. *Front. Psychol.* 8:1135. doi: 10.3389/fpsyg.2017.01135

Raczaszek-Leonardi, J., and Kelso, S. (2007). Reconciling symbolic and dynamic aspects of language: toward a dynamic psycholinguistics. *New Ideas Psychol.* 26, 193–207. doi: 10.1016/j.newideapsych.2007.07.003

Shockley, K., Richardson, D., and Dale, R. (2009). Conversation and coordinative structures. *Top. Cogn. Sci.* 1, 305–319. doi: 10.1111/j.1756-8765.2009.01021.x

Shrout, P., and Rodgers, J. (2019). Psychology, science, and knowledge construction: broadening perspectives from the replication crisis. *Annu. Rev. Psychol.* 69, 487–510. doi: 10.1146/annurev-psych-122216-011845

Sperber, D., and Wilson, D. (1995). *Relevance: Communication and Cognition*. London: Blackwell.

Sperber, D., and Wilson, D. (2002). Pragmatics, modularity and mind-reading. *Mind Lang.* 17, 3–23. doi: 10.1111/1468-0017.00186

Conflict of Interest: 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 © 2020 Gibbs and Colston. 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) and the copyright owner(s) 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: 05 January 2021
doi: 10.3389/fpsyg.2020.556667

Complex Inferential Processes Are Needed for Implicature Comprehension, but Not for Implicature Production

Irene Mognon[1], Simone A. Sprenger[1], Sanne J. M. Kuijper[2] and Petra Hendriks[1]*

[1] Center for Language and Cognition Groningen, University of Groningen, Groningen, Netherlands, [2] Department of Inclusive and Special Needs Education, Faculty of Behavioral and Social Sciences, University of Groningen, Groningen, Netherlands

Upon hearing "Some of Michelangelo's sculptures are in Rome," adults can easily generate a scalar implicature and infer that the intended meaning of the utterance corresponds to "Some but not all Michelangelo's sculptures are in Rome." Comprehension experiments show that preschoolers struggle with this kind of inference until at least 5 years of age. Surprisingly, the few studies having investigated children's production of scalar expressions like *some* and *all* suggest that production is adult-like already in their third year of life. Thus, children's production of implicatures seems to develop at least 2 years before their comprehension of implicatures. In this paper, we present a novel account of scalar implicature generation in the framework of Bidirectional Optimality Theory: the Asymmetry Account. We show that the production–comprehension asymmetry is predicted to emerge because the comprehension of *some* requires the hearer to consider the speaker's perspective, but the production of *some* does not require the speaker to consider the hearer's perspective. Hence, children's comprehension of scalar expressions, but not their production of scalar expressions, is predicted to be related to their theory of mind development. Not possessing fully developed theory of mind abilities yet, children thus have difficulty in comprehending scalar expressions such as *some* in an adult-like way. Our account also explains why variable performance is found in experimental studies testing children's ability to generate scalar implicatures; moreover, it describes the differences between children's and adults' implicature generation in terms of their ability to recursively apply theory of mind; finally, it sheds new light on the question why the interpretation of numerals does not require implicature generation.

Keywords: scalar implicatures, language acquisition, horn scales, asymmetries, semantics–pragmatics interface, optimality theory

OPEN ACCESS

Edited by:
Valentina Cuccio,
University of Messina, Italy

Reviewed by:
Lyn Tieu,
Western Sydney University, Australia
Jacee Cho,
University of Wisconsin-Madison, United States

****Correspondence:***
Irene Mognon
i.mognon@rug.nl

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

Received: *28 April 2020*
Accepted: *13 November 2020*
Published: *05 January 2021*

Citation:
Mognon I, Sprenger SA, Kuijper SJM and Hendriks P (2021) Complex Inferential Processes Are Needed for Implicature Comprehension, but Not for Implicature Production. Front. Psychol. 11:556667. doi: 10.3389/fpsyg.2020.556667

INTRODUCTION

From the earliest age, humans exhibit extraordinary communicative abilities and a pro-social, cooperative attitude. By their first year of life, for instance, infants are able to use nonverbal pointing gestures to direct other individuals' attention (Carpenter et al., 1998) and, just a few months later, they appear to grasp the cooperative and mental essence of communication: from 18 months of age,

infants can interpret pointing gestures on the basis of the experience they have shared with others (Liebal et al., 2009), and tend to repair episodes of miscommunication irrespective of whether the result of the communicative act is in their favor (Grosse et al., 2010). Moreover, some studies demonstrate the existence of a relationship between early pragmatic abilities such as gaze following and pointing and later language development (Brooks and Meltzoff, 2005; Colonnesi et al., 2010), suggesting that the pragmatic component plays a critical role in language acquisition in general. In light of this, children's difficulties with particular forms of pragmatic inferencing appear rather puzzling. In the last two decades, a steadily growing body of literature has focused in particular on Scalar Implicatures (SIs) (Noveck, 2001; Papafragou and Musolino, 2003; Guasti et al., 2005; Barner et al., 2011; Foppolo et al., 2012; Stiller et al., 2015; Skordos and Papafragou, 2016, among others). Consider the sentence in (1), which adults normally interpret as (2):

(1) Some roses in William's garden are red.
(2) Some but not all roses in William's garden are red.

According to the classical Gricean account of SI generation (Horn, 1972; Grice, 1975; Gazdar, 1979), listeners infer (2) from (1) because of the presence of a non-pronounced alternative utterance, namely (3):

(3) All roses in William's garden are red.

Even though the semantic (literal) meaning of *some* is AT LEAST ONE, POSSIBLY ALL (notice that forms are presented in italics and meanings are presented in small caps), the quantifiers *some* and *all* are considered as being part of a Horn scale, so named after Horn (1972). Horn scales are lexical scales organized by informativeness: *some*, the first element of the scale <*some*, *all*>, is less informative than the second element, *all*. Informativeness is generally considered to be based on the semantic relation of entailment: *all* entails *some*, but not vice versa. When speakers use the less informative term of a Horn scale, uttering sentence (1) instead of sentence (3), they manifestly violate Grice's Quantity Maxim, according to which cooperative speakers should always provide as much information as possible. To reconcile the apparent violation of the maxim with the expectation that the speaker is cooperative, listeners can infer that the speaker believes that the non-pronounced sentence (3) is not true. Hence, upon hearing the form *some*, and by negating its stronger alternative *all*, the meaning SOME BUT NOT ALL can be pragmatically derived.

It is worth mentioning that different and sometimes conflicting hypotheses concerning SI generation have been proposed. According to the defaultist view developed by Levinson (2000), implicature generation is automatically triggered by the scalar term *some*; so, by default, irrespective of the context, whenever *some* is used, SOME BUT NOT ALL is derived. On the other hand, according to the grammatical view, SIs emerge at the level of semantic computation (Chierchia et al., 2012; see also Magri, 2009, and subsequent works). According to the defaultist approach and the grammatical approach, the SOME BUT NOT ALL meaning of *some* is not considered as emerging from an online pragmatic process, and it should not be referred to as "pragmatic meaning." In light of this, here we will use the more theory-neutral expression "upper-bounded meaning of *some*" (i.e., the interpretation that excludes the upper bound of the scale, *all*).

Despite knowing the semantic meaning of the quantifier *some* from an early age (Pouscoulous et al., 2007), children struggle to infer its upper-bounded meaning. Until at least 4 or 5 years of age, they tend to accept sentences that for adults would be underinformative, such as sentence (1) in a context in which the full set of roses is in fact red (e.g., Noveck, 2001; Papafragou and Musolino, 2003; Guasti et al., 2005; Pouscoulous et al., 2007; Huang and Snedeker, 2009; Katsos and Bishop, 2011; Foppolo et al., 2012; Skordos and Papafragou, 2016; Horowitz et al., 2018). This issue has been investigated in a large body of literature. However, to date, there is still considerable disagreement about the reason behind children's non-adult-like behavioral pattern, with some researchers focusing on the detrimental effect of task demand (e.g., Papafragou and Musolino, 2003) and others holding that children's problems are intrinsically linked to the pragmatic inferencing process (e.g., Huang and Snedeker, 2009).

In this paper, we will propose the Asymmetry Account, a new account of SI generation, couched in the framework of Bidirectional Optimality Theory (Bi-OT). Importantly, Bi-OT allows us to analyze production and comprehension as separate processes (Blutner, 1998, 2000). Moreover, following Hendriks and Spenader (2006) (but *contra* Blutner, 2006, 2010), we will argue that Bi-OT has psychological validity, and we will show that it correctly predicts children's performance. In particular, we will start presenting some acquisition findings (see section "Different Tasks, Conflicting Results"). We will then introduce two influential accounts of children's difficulties and illustrate some recent corpus data (Eiteljoerge et al., 2018) that point to the fact that children are able to produce SIs already in their third year of life (see section "Previous Accounts of Children's Difficulties"). We will see that this finding casts doubts on the idea that children's difficulties lie in the process of SI generation itself (see section "The Pragmatic Tolerance Account"). We will then show that, contrary to this view, the production–comprehension asymmetry is real (see section "Challenges for the Pragmatic Tolerance Account"). In Section "Carving Quantity-Based Implicature at Its Joints: *Ad Hoc* and Horn Scales," we will rigorously define and discuss some features of SIs. Then, we will present our Asymmetry Account (see section "The Asymmetry Account: A Cognitively Plausible Model of Children's Difficulties") and discuss its predictions (see section "Discussion"). Specifically, we will show that children's comprehension difficulties emerge because implicature generation imposes a cognitive burden on hearers, but not on speakers. Accordingly, children's pattern of performance is explained by the fact that complex inferential processes are not needed in production, but only in comprehension (see section "When Speakers Are More Logical Than Hearers"); in Section "Scalar Implicature Generation and Theory of Mind," the relationship between SIs and theory of mind (ToM) is described; in Section

"When Speakers Become Less Logical and More Pragmatic," the differences between children's and adults' ability to generate implicatures are illustrated. This paper ends with a discussion of the reasons behind children's variable performance in comprehension studies (see section "Explaining Children's Variable Performance in Comprehension Studies") and on the reason why children's interpretation of numerals does not require implicature generation (see section "Why Children Interpret *n* as EXACTLY *n*").

SCALAR IMPLICATURES IN ACQUISITION

Different Tasks, Conflicting Results

One of the striking characteristics of studies on children's implicature generation is that the particular task used and the contextual support provided to participants substantially influences the outcome of the experiments, to such an extent that the age at which children have been reported to acquire the adult-like interpretation varies between age 5 and preadolescence.

Noveck (2001) is one of the first studies to systematically investigate SIs in language acquisition (but see also Paris, 1973; Smith, 1980; Chierchia et al., 1998, 2001). In this study, children were asked to evaluate sentences such as "Some giraffes have long necks" uttered in isolation (Statement Evaluation Task). Noveck's (2001) results indicated that even at the age of 11, children do not reliably reject underinformative sentences containing the quantifier *some*. However, tested with this paradigm, even the adult participants in this study did not draw inferences at a high rate (59% for adults vs. 15% for 11-year-olds). Hence, albeit being useful in revealing a difference between children and adults, the Statement Evaluation Task does not seem the most reliable tool to measure SI generation, given that, as demonstrated in later studies (e.g., Guasti et al., 2005), this paradigm favors the emergence of logical interpretations also in adult participants. Probably, the reason for this lies in the abstract nature of the task, which consists in judging world-knowledge statements in isolation.

Subsequent studies (Lidz and Musolino, 2002; Papafragou and Musolino, 2003; Guasti et al., 2005; Foppolo et al., 2012) adopted another kind of comprehension task, namely, the binary Truth Value Judgment Task (TVJT), and showed that the age at which children are able to generate SIs can be lowered considerably. For instance, in Experiment 1 of Foppolo et al. (2012), adults and children aged 4 to 7 were asked to evaluate a sentence in combination with a particular picture (e.g., "Some Smurfs are going on a boat" presented in combination with a picture in which five out of five Smurfs are on a boat). Six-year-old children demonstrated to be able to generate SIs almost at an adult-like rate (83% for 6-year-olds vs. 87% for adults).

Interestingly, the same study also illustrates the largely overlooked difference between the ability to generate SIs and the ability to identify the most informative between two given alternatives. In Experiment 5 of Foppolo et al. (2012), the group of 5-year-olds who failed to compute the SI in the previously administered TVJT was administered a Felicity Judgment Task (FJT). In FJTs (a paradigm first introduced by Chierchia et al., 2001), participants are provided with two statements and are asked which one best describes a given picture. In the critical items of Foppolo et al.'s (2012) Experiment 5, children heard a sentence containing *all* and a sentence containing *some* (e.g., "All the chipmunks are taking a shower" vs. "Some chipmunks are taking a shower" in combination with a picture showing five out of five chipmunks taking a shower). Quite surprisingly, children's performance in this task was 95% correct overall (see also Chierchia et al., 2001, for similar results). Thus, children's difficulties with SI generation do not appear to emerge in connection with an inability to grasp the difference (in terms of informativeness) between *some*- and *all*-sentences (see section "Previous Accounts of Children's Difficulties" for discussion).

That the experimental manipulation can drastically influence children's performance in SI experiments was further demonstrated also by a study conducted by Pouscoulous et al. (2007). These authors adopted an Act-Out Task (AOT), a methodology that allows children to indirectly exhibit their ability to generate SIs by performing an action, instead of giving a verbal judgment. According to the authors, task demand is to be held responsible for hampering children's SI generation in TVJT and similar paradigms. In line with this hypothesis, their results showed that, if the task is simple enough, from the age of 5, children rather robustly generate SIs: in their task, 73% of 5-year-olds (and 88% of adults) demonstrated to have interpreted *some* as SOME BUT NOT ALL.

Previous Accounts of Children's Difficulties

Various explanations have been proposed for why children experience difficulties in generating SIs. In what follows, we discuss two of the most influential accounts: the Lexicalist account by Barner et al. (2011) and the Pragmatic Tolerance account by Katsos and Bishop (2011).

The Lexicalist Account

According to Barner et al. (2011), children's problems do not stem from pragmatic immaturity or processing difficulties, but rather lie in a particular step of SI generation, namely, the retrieval of the scale of alternative lexical terms. In fact, accessing the scale and recognizing the existence of an alternative is clearly a prerequisite for generating implicatures. Barner et al. (2011) argued that preschoolers fail in generating the relevant scalar alternative (e.g., *all*) when hearing a scalar item (e.g., *some*) (see also Foppolo et al., 2012, for a similar claim). Notably, this hypothesis can explain why children struggle with Truth Value Judgment Tasks but show adult-like performance in Felicity Judgment Tasks (Foppolo et al., 2012). In the latter case, the strong alternative (the sentence with *all*) is already given in the task, and the task can be carried out simply by recognizing that in critical trials, the *all*-sentence is more appropriate.

However, as we will now see, further experimental evidence (Eiteljoerge et al., 2018) casts doubt on the plausibility of the lexicalist account.

The Pragmatic Tolerance Account

Despite the ever-growing number of studies devoted to the topic, researchers have mainly focused on children's comprehension of SIs and hardly any experiment examined children's *production*. There are, however, a few notable exceptions.

Production data are presented by Foppolo and Guasti (2005). In this study, an Elicitation Task was used in order to assess whether children can use *some* and *all* in an adult-like manner. With the aim of eliciting sentences containing quantified NPs, Italian children aged 3;7 to 5;8 were presented with stories and asked to describe what had happened to a set or subset of characters. Children's mistakes were rather infrequent: *all* was used correctly 95% of the time [only in 4 out of 71 utterances children used *tanti* (many) instead of *tutti* (all) to refer to a full set of characters]. On the other hand, to refer to a subset of characters, children produced 53 utterances containing different lexical items that appear equivalent to the English *some* (the exact number of instances of the different items used is not reported in the paper). Foppolo and Guasti's (2005) conclusion was that children can appropriately use *all* and *some* in production: the former when describing a full set of relevant characters, the latter when describing a subset of the relevant characters. Importantly, children never used *some* underinformatively to refer to a full set of characters.

A discrepancy between the correct use of quantifiers in production and the difficulties in comprehension emerges also in the study of Katsos and Smith (2010). In Experiment 1, children were tested in both comprehension and production. The comprehension part consisted in a classical TVJT, in which children listened to stories and were asked to indicate whether the fictional character Mr. Caveman replied correctly to some questions. In critical trials, Mr. Caveman would say, for instance, that some of the carrots had been picked up when in fact all of them had been picked up. In the production task, on the other hand, children would see a scenario in which a subset of objects was acted upon; this time, however, not Mr. Caveman, but the children themselves were asked to provide an appropriate description of the situation. Performance in the comprehension task confirmed previous findings: children overwhelmingly failed to reject underinformative sentences, thus showing not to have generated the implicature. The same group of children, however, was able to produce informatively appropriate utterances, using the quantifier *all* (or a numeral, or a plural noun phrase such as *the carrots*) instead of the underinformative *some* when describing the so-called ALL-scenario. So, despite accepting underinformative sentences in comprehension, children demonstrated to be fully informative speakers.

In this study, the intriguing asymmetry between children's adult-like production and children's non-adult-like comprehension was interpreted as evidence in favor of the Pragmatic Tolerance Hypothesis (Davies and Katsos, 2010; Katsos and Smith, 2010; Katsos and Bishop, 2011). According to this hypothesis, children are pragmatically competent and are aware of the underinformativeness of *some*-sentences in ALL-scenarios. Nevertheless, they do not penalize pragmatic violations as adults do. As a result, in the binary judgment tasks that are typically employed to test SIs, children tend to accept underinformative sentences—which, in fact, are not semantically false. Nevertheless, in particular paradigms (such as Katsos and Smith's production experiment), they can exhibit their pragmatic abilities. Their non-adult-like behavior is simply due to an overly tolerant pragmatic attitude.

Further evidence that children in production can show adult-like competence is provided by a recent study carried out by Eiteljoerge et al. (2018). These authors conducted a corpus study analyzing the production of sentences containing the quantifier *some*. Spontaneous utterances (N = 2883) of five English children aged 2;00 to 5;01 were inspected and categorized according to the likeliness to contain a SI. The classification was based on the linguistic context (i.e., three lines of context before and after each occurrence of *some* were examined) and structural features (e.g., partitive constructions, plural noun phrases, etc.). An implicature was categorized as *Possible* or *Probable* if a quantifiable set could be recognized and the speaker was probably referring to a subset of the quantified set using *some* with the NOT ALL meaning. Among the included utterances, *Implicature Implausible*-sentences (i.e., sentences in which most likely the speaker was not implying NOT ALL) were the majority (70.76%). Nevertheless, in 19.46% of utterances, an implicature was *Possible* or *Plausible* (e.g., "The puzzle is missing some pieces," while describing a puzzle). Strikingly, even 2-year-old children were able to use *some* in a way that clearly triggers implicature generation: one child, Fraser, did so at 2;03 years of age; all the others did so before or around 3;00 years of age. In light of their data, Eiteljoerge et al. (2018) criticized Barner et al.'s (2011) lexicalist account, claiming that: "If toddlers have not associated *some* with its lexical scale (*many*, *most*, *all*), this should affect their ability to produce, as well as comprehend, implicatures" (Eiteljoerge et al., 2018, p. 14).

Moreover, as Eiteljoerge et al. (2018) observed, the low rate of produced implicatures in a children's corpus should not come as a surprise. In fact, children's production was in line with mothers' usage, as the analysis of mothers' child-directed speech revealed. Among adults' sentences, only 16% of the instances of *some* could be analyzed as *Implicature Possible* or *Plausible*. Interestingly, although in the literature it is almost always implicitly assumed that "scalar implicatures arise more often than not when the lexical item *some* is used" (Degen, 2013, p. 164), this assumption, as shown by Degen (2013), is not borne out by corpus studies.

This being said, the finding that children use *some* with its upper-bounded meaning at least 2 or 3 years before they show an adult-like comprehension of the same term suggests that a purely lexicalist account along the lines of Barner et al. (2011), albeit intriguing, cannot be wholly satisfactory. Moreover, as mentioned by Eiteljoerge et al. (2018), their data are in line with an explanation of children's non-adult-like comprehension pattern in terms of non-linguistic factors, as proposed by Katsos and Bishop (2011).

Challenges for the Pragmatic Tolerance Account

The idea that the difficulties in generating SIs in binary comprehension tasks lie in children's excessive pragmatic

tolerance—and not in the generation itself or in particular steps required for the generation—is extremely appealing, in that it dismisses the issue of the production–comprehension asymmetry. However, a careful examination of further data casts some doubts on the explanatory power of the Pragmatic Tolerance hypothesis.

Firstly, on a general level, it can be argued that if children were more pragmatically tolerant than adults, excesses in pragmatic tolerance would emerge in other contexts too. Contrary to this, however, we know that children are endowed with an astonishing pragmatic sensitivity, which appears incompatible with a hypothetical overly tolerant pragmatic attitude. As an example of children's extraordinary sensitivity to communicative intentions, consider the aforementioned study of Grosse et al. (2010). In this work, the authors showed that infants as young as 18 months of age recognize and tend to repair episodes of miscommunication even if those same episodes accidentally lead to children's desired outcomes. In this experiment, children were prompted to ask for an object. In critical trials (*Happy Accident Conditions*), an experimenter would pretend to have misunderstood the request but at the same time accidentally provide the child with the desired object, placing it in a target position. Despite having received the requested object, 18-, 24-, and 30-month-olds tried to repair the communication, through gestures, vocalizations, or verbal sentences.

One could argue that if children were excessively tolerant toward pragmatic violations in general, they would ignore communicative failures and welcome the desired outcome without trying to repair. However, this is not the case. Children as young as 18 months of age do not regard communication as a simple tool to manipulate others' behavior. On the contrary, they are alert and aware of communicative pragmatic deviances. In light of this observation and of evidence coming from numerous other studies that point to children's extraordinary pragmatic skills (Matthews, 2014, for an overview), it is safe to claim that children are not, generally speaking, more pragmatically tolerant than adults. Consequently, if pragmatic tolerance is the factor responsible for children's non-adult-like behavior in SI generation, we have to assume that pragmatic tolerance is restricted to violations of underinformativeness only. This, however, seems an unwelcome result given that we would have to postulate a phenomenon-specific pragmatic tolerance.

Secondly, apart from children's early pragmatic abilities, it seems quite hard to understand why preschoolers' pragmatic tolerance would emerge just in comprehension, and not in production too. If children simply required sentences to be true and not also pragmatically appropriate, they should also produce, at least some of the times, pragmatically infelicitous sentences using *some* instead of *all*. However, this does not seem to be the case (Foppolo and Guasti, 2005; Katsos and Smith, 2010).

Thirdly, eye-tracking research (although data are still scarce) seems to suggest that 5-year-old children struggle—or at the very least, require significantly more time than adults—at a processing level, to generate *some*-implicatures (Huang and Snedeker, 2009). If problems emerge in SI processing, the locus of children's difficulties with SI in general should lie in the inferencing process, or in particular steps of this process. This would be at odds with the Pragmatic Tolerance Hypothesis, according to which there are no inherent difficulties in children's SI generation.

In sum, although Katsos and Bishop's (2011) account elegantly explains the asymmetry between the production and comprehension of *some*, it faces substantial challenges and the search for alternative explanations seems to be warranted.

In what follows, adopting the framework of Bidirectional Optimality Theory (Bi-OT), we develop a novel account of children's SI generation and of the production–comprehension asymmetry that emerges in connection with *some*. As shown by Blutner (1998, 2000), Bi-OT is particularly suited to model Gricean pragmatics (see also Schulz and Van Rooij, 2006; Aloni, 2007; Krifka, 2007, 2010, 2011). We start by rigorously defining SIs, in the belief that any account of children's difficulties makes terminological clarity particularly important (see section "Carving Quantity-Based Implicature at Its Joints: *Ad Hoc* and Horn Scales"). Then, we describe two constraints that determine the semantics of the scale <*some*, *all*> (see section "Translating Horn Scales in Constraints"). We show how these constraints interact (see section "Constraint Interaction: Modeling Speakers' and Hearers' Perspectives Separately") and how implicatures can be modeled (see section "Bidirectional Optimization: Generating the Implicature"). Lastly, we illustrate the predictions of our Asymmetry Account (see section "Discussion"), which, we argue, explains why children experience difficulties comprehending SIs, although they are able to produce *some* with its upper-bounded meaning from a very young age.

CARVING QUANTITY-BASED IMPLICATURE AT ITS JOINTS: *AD HOC* AND HORN SCALES

As mentioned in the Section "Introduction," according to the traditional Gricean approach, conversational implicatures can be seen as non-truth-functional meanings emerging in connection with the Principle of Cooperativity. Quantity-based implicatures (QBIs), in particular, are those implicatures that are based on the two submaxims of Quantity (here in the formulation of Matsumoto, 1995, p. 23).

(4) First submaxim of Quantity: Make your contribution as informative (strong) as possible.
Second submaxim of Quantity: Do not make your contribution more than is required in the context of the exchange.

Under the label QBI, we can include SIs as well as at least some instances of *ad hoc* implicatures. The distinction between scalar (or generalized) implicatures and *ad hoc* (or particularized) implicatures, introduced by Grice (1975, 1989), is based on inferences' inherent degree of (in)dependence from the context. To illustrate, consider the following sentences and the relevant inferences.

(5) a: I ate some of the apples.
b: I ate some but not all of the apples.

(6) a: My friend wears glasses.
b: My friend wears glasses and not a hat.

The inference in (5b), a SI, appears to naturally follow from the sentence in (5a). In contrast, the inference in (6b), an *ad hoc* implicature, seems not to follow automatically from (6a). In (6a), the implicature emerges only if the context is such that *glasses* and *glasses and hat* constitute relevant alternatives. This happens, for instance, when the sentence in (6a) is uttered in a situation in which there is a person who is wearing a hat and glasses, and another person who is wearing just glasses.

The distinction between *ad hoc* implicatures and SIs has been challenged, among others, by advocates of Relevance Theory (Sperber and Wilson, 1986)[1]. Irrespective of whether the distinction is cognitively legitimate, according to the Gricean tradition, both classes of implicatures are generated when a speaker intends to communicate a particular meaning that goes beyond the literal meaning of the uttered words and does so by uttering a sentence in which the quantity of information is reduced with respect to what the listener could have expected.

Defining formally what is meant with *quantity of information* is rather problematic. A viable solution, proposed by Horn (1972), is to describe *informativeness* in terms of asymmetric semantic entailment. Roughly, an item P asymmetrically entails Q if P is true in all set of circumstances in which Q is true, but not vice versa (see also Gazdar, 1980). To exemplify, if the sentence in (7) is true, the sentence in (8) is also true, but not the other way around.

(7) All of my friends are linguists.
(8) Some of my friends are linguists.

Accordingly, the so-called Horn scales are those ordered sets of lexical items whose members have a similar structural complexity (cf. Katzir, 2007, for an in-depth discussion) and stand in an asymmetrical relationship of entailment, and because of this, are particularly prone to give rise to SIs. So, if in the case of *ad hoc* implicatures, what counts as a relevant alternative is determined by the context (as shown in 6), relevant alternatives are lexically defined in the case of implicatures that emerge from Horn scales (as shown in 5).

It should be observed that, as argued by Hirschberg (1985), the Horn/*ad hoc* scales dichotomy is perhaps a false one. As pointed out by Horn himself in later works (Horn, 2006), what we can call Horn-Scalar Implicatures are, to a certain degree, context-sensitive too. Nevertheless, being based on terms that are strongly associated at the lexical level, they are inevitably *less* context-sensitive than other QBIs (see Barbet and Thierry, 2018, for experimental evidence).

In this regard, it is relevant to mention that the association between the scale mates that constitute Horn scales appears to be demonstrated experimentally. Adopting a masked priming paradigm, de Carvalho et al. (2016) showed that less informative items of scales can prime stronger items of the same scale. Conversely, priming from stronger items to the less informative one is weak. This points to the fact that stronger words are evoked when weaker ones need to be interpreted, but not the other way around. The association operates in one direction only, so scalar weaker terms are *asymmetrically* associated with certain alternatives at the level of the mental lexicon. Scales, in essence, appear to have a psychological reality. Most importantly for our purposes here, the existence of such links between scalar items has been, by and large, taken for granted in the acquisition literature, and the cognitive reality of Horn scales is at the core of Barner et al.'s (2011) lexicalist account.

The controversy in language acquisition is predominantly centered around SIs, strictly defined as being based on the Quantity Maxim and Horn scales. Thus, with the aim of providing an adequate and cognitively plausible explanation of children's SI generation, as tested in an ever-growing number of studies, we focus on a particular Horn scale, namely, <*some*, *all*>. With slight modifications, the analysis presented in the remainder of this paper, however, can be applied to SIs that emerge from the whole class of Horn scales.

THE ASYMMETRY ACCOUNT: A COGNITIVELY PLAUSIBLE MODEL OF CHILDREN'S DIFFICULTIES

Describing Scalar Implicature generation presupposes an understanding of Horn scales functioning. Here, we argue that the comprehension and production of *some* and *all*, and consequently, SI generation, are regulated by two semantic principles, or constraints.

Our account is couched in the constraint-based framework of Bidirectional Optimality Theory (Bi-OT). In Bi-OT, production and comprehension of lexical elements are seen as optimization processes in which, given an input, an optimal output needs to be identified. Specifically, in production the input is a meaning and the output is a form (i.e., what will be finally uttered). In comprehension, the input is a form and the output is a meaning (i.e., the interpretation that will be chosen). Clearly, both in production and in comprehension, given an input, there are several possible outputs. When we want to communicate a meaning, we need to choose among different forms, and when we hear a form, we need to choose among different meanings. The evaluation process is guided by constraints. In OT (Hendriks and Spenader, 2006 for OT semantics; cf. Prince and Smolensky, 2004 for OT phonology), these constraints are violable and hierarchically organized. Stronger constraints are more important than weaker ones

[1]The claim that *ad hoc* implicatures and scalar implicatures belong to the same class of phenomena (e.g., Hirschberg, 1985; Geurts, 2010) has been challenged on experimental grounds. Stiller et al. (2015), for instance, showed that children, as young as 3;5 years of age, can generate *ad hoc* implicatures (see also Horowitz et al., 2018, for similar results, but see Katsos and Bishop, 2011; Schaeken et al., 2018, for conflicting evidence). Given the experimental methods adopted in *ad hoc* implicatures studies, however, we believe that children's success with *ad hoc* implicatures may be achieved through a strategy that does not require implicature generation. In particular, we speculate that children' good performance could be linked to (and possibly explained by) contrast inferencing (e.g., a contrast between the most relevant features of two visually presented pictures) rather than implicature generation (see Sullivan et al., 2019, for experimental evidence that contrast/exclusion inferences may play a role children's success in various implicatures studies). See also footnote 2 for further evidence that *ad hoc* implicatures and scalar implicatures may be different phenomena.

and, whenever two constraints are in conflict, the weaker one can be violated.

We now show how the interaction between SI constraints can explain children's comprehension failures as well as their production successes (for a more formal treatment of these constraints and of their interaction in the Bi-OT framework, see Mognon et al., in press).

Translating Horn Scales in Constraints

The first constraint we introduce arises directly in connection with Grice's Cooperative Principle. Consider the first submaxim of Quantity, mentioned above in (4). This general and fundamental principle of communication is mirrored in the constraint that we call Strength (cf. Zeevat, 2000, and Hogeweg, 2009, for the production counterpart of this constraint):

(9) Strength: Use the strongest element on the Horn scale.

According to this constraint, if two terms (in this case, *some* and *all*) stand in a relation of entailment and can both be used, *salva veritate*, in a given context, then speakers should lean toward choosing the most informative term (here, *all*).

The constraint Strength interacts with a family of constraints, which, like Strength, is relevant to the comprehension and production of scalar elements. In particular, this family of constraints is essential to introduce a link between forms and the dimension conveyed by Horn scales. First, let us consider a virtually ignored but fundamental feature of scales. Lexical scales are always polarized toward a culmination point, which can be a lower or an upper bound. We call this culmination point the *apex* of the scale. The apex is the maximization of the dimension conveyed by the scale. Equivalently, it represents the strongest lexical meaning of the scale. To give an example, the apex of the scale <*possible*, *certain*> is NECESSITY. It is possible to identify apices also in the case of *ad hoc* scales, even if, needless to say, these are *ad hoc* apices.

To better grasp the nature of apices, consider the following context. A traveler is going from Europe to Vladivostok via the Trans-Siberian route and utters the following:

(10) I've reached Novosibirsk.

The utterance in (10) is likely to give rise to a "not Vladivostok"-inference. The *ad hoc* scale here consists of the various stops along the Trans-Siberian route, and the *ad hoc* apex is something like LAST STOP OF TRANS-SIBERIAN ROUTE, which corresponds to the city name *Vladivostok*. It is worth noting that experimentally demonstrating the cognitive reality of scales amounts to demonstrating the cognitive existence of apices. At least for what concerns Horn scales, as mentioned, evidence has already been found (de Carvalho et al., 2016).

Turning back to Horn scales, and in light of the existence of apices, we can now introduce the aforementioned family of constraints: FaithHorn. The family of FaithHorn constraints promotes the mapping between the strongest lexical element on a Horn scale (i.e., the element of the scale that entails the other weaker elements) and a particular meaning, namely, the apex of the relevant Horn scale. When applied to the scale <*warm*, *hot*>, for instance, FaithHorn promotes the mapping between *hot* and the apex of the scale, namely, HEAT. In the case of the <*some*, *all*>-scale, FaithHorn links the term *all* with complete sets. We label this specific constraint FaithAll.

(11) FaithAll: *All* corresponds to complete sets.

Trivial as it seems, FaithAll is a fundamental, primitive constraint of the semantics of the <*some*, *all*>-scale. It is violated by an association between *all* and a non-complete set.

Constraint Interaction: Modeling Speakers' and Hearers' Perspectives Separately

Having introduced the two constraints that are relevant for our account of SIs, we now illustrate their interaction. As mentioned above, in Bi-OT, constraints are seen as violable and hierarchically organized. Complying with a stronger constraint is more important than complying with a weaker constraint, and if two constraints are in conflict, then the weaker constraint can be violated in order to satisfy the stronger constraint.

Production and comprehension of linguistic expressions can be seen as independent but related processes. They are guided by the same constraints, but in production, speakers need to map meanings onto forms, whereas in comprehension, hearers need to map forms onto meanings (Hendriks, 2016). Thus, the effects of the application of the same constraints may yield different results in production and comprehension (Smolensky, 1996).

Let us describe the interaction of constraints, first, taking the perspective of speakers and hence considering the production processes (**Figure 1**).

Consider a speaker who wants to refer to a complete set of items, in which five out of the five roses are red (**Figure 1A**). Given the choice between the form *some* and the form *all*, the speaker can easily exclude *some* because choosing it would violate the constraint Strength ("Use the strongest element on the Horn scale") and generate an underinformative message. Choosing *all* to refer to a complete set, on the other hand, does not violate any constraint: as Strength requires, *all* is the strongest term of the scale at hand and, as stated by FaithAll, can be associated with complete sets. In other words, *all* is the optimal candidate to refer to complete sets. The speaker, thus, can readily utter the following sentence:

(12) All the roses are red.

A different situation arises when the speaker wants to refer to a set that is not complete, where, for instance, three out of the five roses are red (**Figure 1B**). Choosing *all* violates FaithAll, given that, according to this constraint, *all* should always be associated with a complete set. Choosing *some*, on the other hand, violates Strength, given that there is a stronger term on the scale. However, FaithAll is higher-ranked than Strength. Therefore, the violation of Strength is less grave than the violation of FaithAll. Hence, *some* is a better option than *all* to describe a set that is not complete. So, when a speaker wants to describe a scenario in which three out of five roses are red, using the quantity scale at hand, the speaker's only option is to utter (13):

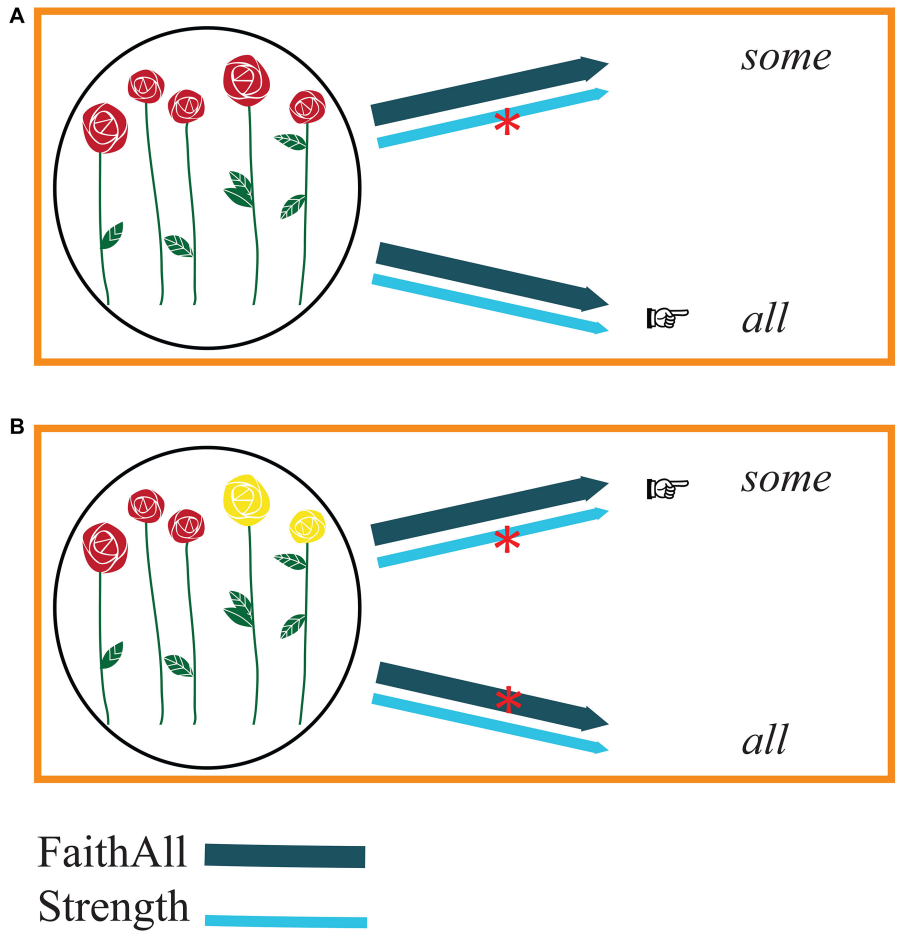

FIGURE 1 | Production (speakers' perspective). The arrows representing FaithAll (dark blue) and Strength (light blue) link meanings with possible forms. Constraint violations are represented by red asterisks on the arrows. The relative strength of the constraints is indicated by the weight of the line of the arrows: FaithAll is stronger than Strength. Pointing fingers indicate which form proves to be optimal on the basis of the constraints. Panel **(A)** illustrates reference to a complete set of flowers: the meaning to be expressed corresponds to a complete set of elements. In this case, choosing *some* would violate Strength. The optimal form, hence, is *all*. Panel **(B)** illustrates reference to a subset of red flowers in a larger set of flowers of different colors: the situation in which the meaning to be expressed corresponds to a non-complete set of elements. In this case, choosing *some* would violate Strength, whereas choosing *all* would violate FaithAll. However, given that FaithAll is stronger than Strength, the optimal form is *some*.

(13) Some of the roses are red.

The two production processes just described are carried out by speakers whenever they need to refer to sets using the quantity scale <*some*, *all*>.

The hearers' perspective (**Figure 2**) differs from the speakers' perspective. In the hearers' perspective, the constraint Strength has no effect because this constraint expresses a preference for the choice of forms. Hence, it influences production but cannot influence comprehension. In other words, in the comprehension process, the form is already given—it has been uttered by the speaker. Thus, "Prefer the strongest element on the Horn scale" has no effect and it is simply not relevant when deciding how to interpret a form such as *all* or *some*. The comprehension of the elements of the <*some*, *all*>-scale depends uniquely on the constraint FaithAll. How, then, does this constraint affect the interpretation of the two quantifiers?

When the form *all* is heard and needs to be interpreted (**Figure 2A**), FaithAll ("*All* corresponds to complete sets") rules out every interpretation but the complete set. Thus, following FaithAll, the form *all* is straightforwardly associated with a complete set meaning.

What about the interpretation of *some*? When the form *some* is heard and needs to be interpreted (**Figure 2B**), FaithAll does not rule out non-complete sets, nor complete sets: in fact, FaithAll only requires an association between *all* and a complete set. Hence, when *some* has to be interpreted, FaithAll is simply not relevant. So, as outputs of the comprehension process of *some*, complete sets and non-complete sets are both optimal candidates. The result of this is that, from hearers' perspective, *some* is ambiguous because it is compatible with complete sets and non-complete sets.

The analysis of production and comprehension processes of the <*some*, *all*>-scale just proposed, then, gives rise to the

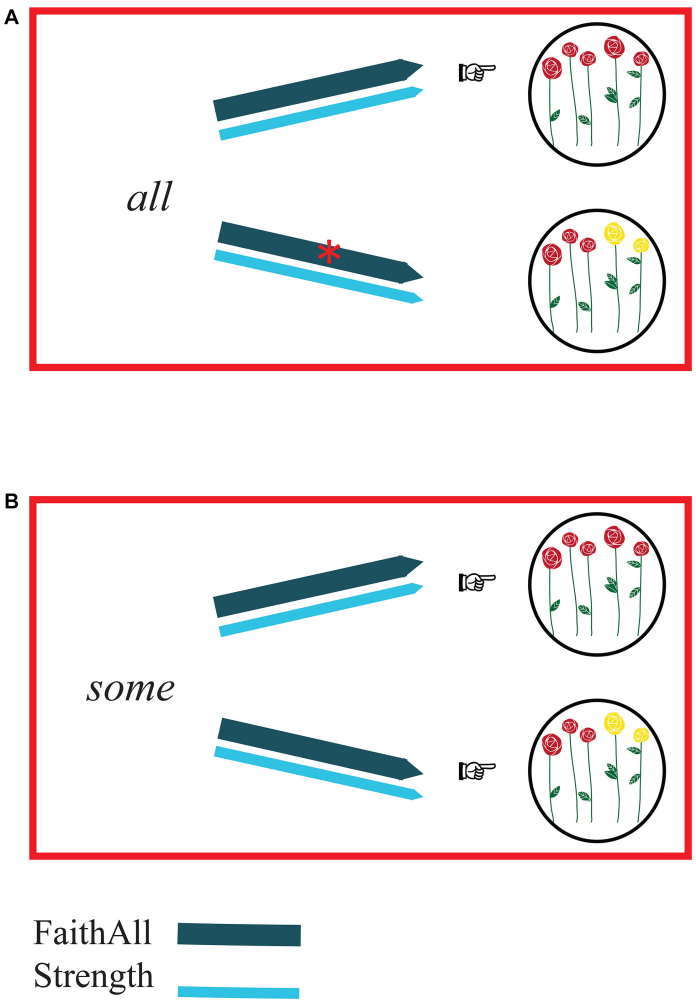

FIGURE 2 | Comprehension (hearers' perspective). The arrows representing FaithAll (dark blue) and Strength (light blue) link forms with possible meanings. As in Figure 1, constraint violations are represented by red asterisks on the arrows, the relative strength of the constraints is indicated by the weight of the line of the arrows, and pointing fingers indicate which meaning proves to be optimal on the basis of the constraints. Panel **(A)** illustrates the comprehension of the form *all*: FaithAll rules out the interpretation consisting of a non-complete set of elements, whereas Strength does not have any effect. The optimal interpretation appears to be the one consisting of a complete set of elements (here: the complete set of flowers). Panel **(B)** illustrates the comprehension of the form *some*. In this case, both interpretations are possible, because, irrespective of the chosen interpretation, neither FaithAll nor Strength are violated. So, the form *some* turns out to be ambiguous between two interpretations: a complete set of elements (here: the complete set of flowers) and a non-complete set of elements (here: a subset of flowers) are both optimal meanings.

following result: in production, reference to complete sets is made using the form *all* (**Figure 1A**) and reference to non-complete sets is made using the form *some* (**Figure 1B**). In comprehension, *all* is straightforwardly interpreted as referring to complete sets (**Figure 2A**). The comprehension of *some*, on the other hand, is problematic because *some*, in hearers' perspective, proves to be ambiguous (**Figure 2B**).

This indeed is what we find when we test children on the comprehension and the production of the most popular Horn scale, <*some*, *all*>: the comprehension and production of *all* are adult-like, and so is the production of the upper-bounded *some* (i.e., SOME BUT NOT ALL). The comprehension of the form *some*, however, is problematic for children: they tend to overaccept *some*-sentences, showing that they do not spontaneously generate the *some*-implicature. This, again, is in line with our model, which predicts that *some* is ambiguous between two interpretations. How do adults resolve this ambiguity that stems from the semantics of *some*? The process bringing to light the SOME BUT NOT ALL interpretation of *some* is bidirectional optimization.

Bidirectional Optimization: Generating the Implicature

So far, we have seen how two semantic constraints account for children's production and comprehension of the lexical element of the <*some*, *all*>-scale. Remarkably, according to our analysis, the comprehension of *some* results in ambiguity (see **Figure 2B**). This ambiguity, however, can disappear in

the adult comprehension thanks to a process of bidirectional optimization. In a bidirectional optimization process, the effect of the constraints in production and the effects of the constraints in comprehension are both taken into account (Blutner, 1998, 2000 and subsequent works). Informally, bidirectional optimization can be thought as a perspective-taking mechanism (Van Rij et al., 2010). Let us describe its functioning.

Suppose an opinionated speaker wants to refer to a situation in which three out of five roses are red using an expression on the <*some*, *all*>-scale (and not, say, a cardinal number). In light of the two constraints Strength and FaithAll, the speaker has no choice but to use the form *some*, with its upper-bounded reading (see **Figure 1B**). Notice that no implicature has been generated here. As we will see in Section "Discussion," producing *some* to refer to a subset is not equivalent to generating an implicature.

What about the hearer? When hearing *some*, the hearer's language system is faced with an ambiguity (see **Figure 2B**): from the hearer's perspective, given Strength and FaithAll, it is impossible to choose between a complete or a non-complete set. However, *some* can be disambiguated if the hearer considers also the speaker's perspective.

To do so, the hearer needs to consider the effects of the constraints Strength and FaithAll not just from the hearer's own perspective (comprehension perspective, in which the output is a meaning) but, simultaneously, also from the speaker's perspective (production perspective, in which the output is a form). So, rather than simply finding the optimal meaning of the form the hearer has heard, the hearer needs to assess whether, in production, that optimal form would have been chosen for that meaning. In other words, the hearer needs to evaluate all *form-meaning associations* on the basis of the constraints (see **Figure 3**). Concretely, the bidirectional process proceeds as follows. Taking into consideration both hearer's and speaker's perspective, a first optimal association, which does not violate any constraint, can be identified: the association between the complete sets and the form *all* (association in **Figure 3A**). Globally, both from the production perspective and from the comprehension perspective, this association does not violate Strength ("Use the strongest element on the Horn scale") or FaithAll ("*All* corresponds to complete sets"). For example, in production, uttering "All roses are red" to refer to a complete set of roses does not violate any constraint. Likewise, in comprehension, interpreting "All

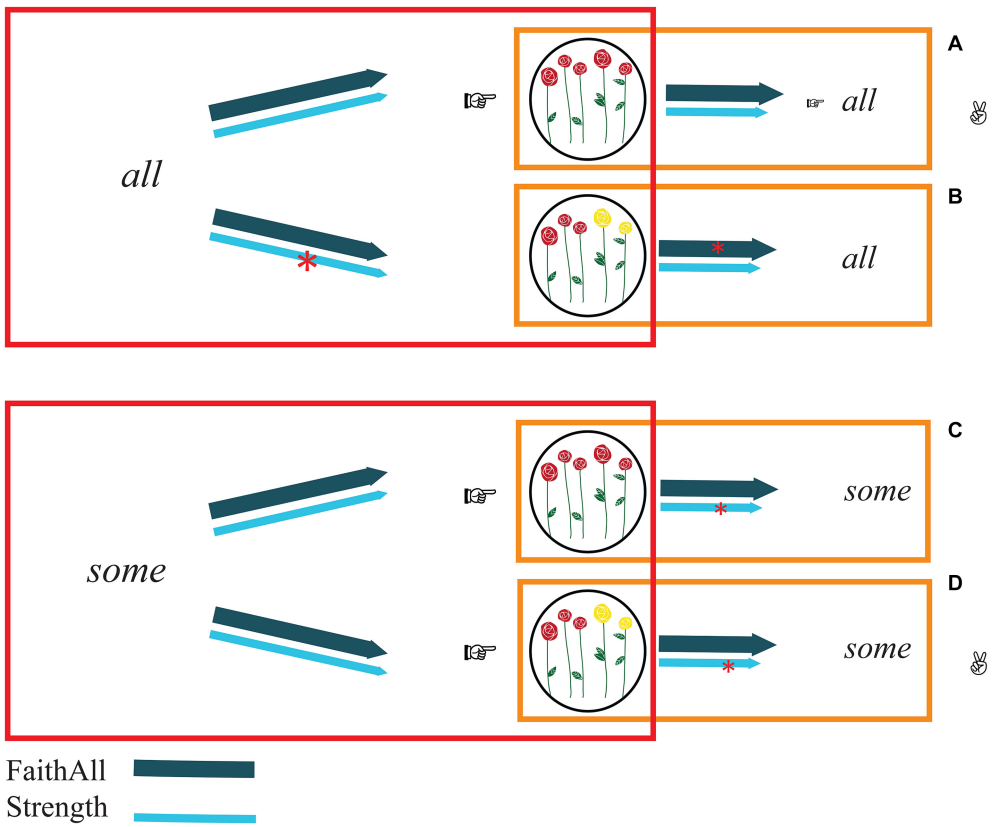

FIGURE 3 | Bidirectional optimization (in comprehension). Panels **(A–D)** represent the four possible form-meaning associations. The process starts with an evaluation of the optimal meanings given the forms *all* and *some* (red boxes) and continues with a second step in which the associations are evaluated according to speakers' perspective (orange boxes). The first bidirectionally optimal association that can be determined is panel **(A)**: it does not violate any constraint in comprehension (see red boxes) nor in production (see orange boxes). All the other possible associations violate at least one constraint (see red asterisks). However, if hearers also consider the speaker's perspective, panels **(B,C)** must be excluded. Hence, panel **(D)** emerges as the second optimal association: there is no better meaning than reference to the non-complete set to interpret *some*, and there is not better form than *some* to refer to a non-complete set of elements.

roses are red" as referring to a complete set of roses does not violate any constraint.

What about the other associations? The other possible associations are the following: *all*-non-complete set (**Figure 3B**), *some*-complete set (**Figure 3C**), and *some*-non-complete set (**Figure 3D**). Crucially, the first two (**Figures 3B,C**) must immediately be excluded. Specifically, the association *all*-non-complete set (**Figure 3B**) cannot be considered an optimal association because the form *all* can be better interpreted as referring to the complete set (i.e., association in **Figure 3A**). Likewise, the s*ome*-complete set (**Figure 3C**) cannot be considered an optimal association because the complete set can be better referred to using *all* (so, **Figure 3A**). By exclusion, then, the association between *some* and the non-complete set (**Figure 3D**) can be established. As a matter of fact, there is no better interpretation for *some* than the non-complete set, and there is no better form than *some* to express the non-compete set.

Due to this process of bidirectional optimization, the hearer has considered all the possible ways in which *some* and *all* can be interpreted (the associations in **Figures 3A–D**) and can conclude that in uttering *some*, the speaker could have in mind only one of *some*'s meanings. Thus, the hearer is able to associate the word *some* with the upper-bounded reading (**Figure 3D**). SI generation consists precisely in this disambiguation of *some* on the part of the hearer.

One observation is now in order. That hearers and speakers have different roles in SI generation is undisputed. As Horn (2006) rightly claimed: "Speakers implicate, hearers infer." Nonetheless, the idea that production and comprehension are distinct processes has not been incorporated in theories of implicatures. Most importantly, it hardly plays a role in any explanations of children's difficulties. We will now discuss the advantages of our Bi-OT approach, which allows one to consider the production and comprehension processes separately.

DISCUSSION

The constraints we introduced in Section "Translating Horn Scales in Constraints," FaithAll and Strength, and their interaction, are at the core of our Bi-OT analysis. In this section, we examine in detail the predictions that arise from our Asymmetry Account, concerning in particular the *some*-implicature asymmetry (see section "When Speakers Are More Logical Than Hearers"), the relationship between ToM and implicature generation (see sections "Scalar Implicature Generation and Theory of Mind" and "When Speakers Become Less Logical and More Pragmatic"), children's variable performance in comprehension studies (see section "Explaining Children's Variable Performance in Comprehension Studies"), and children's interpretation of numerals (see section "Why Children Interpret *n* as EXACTLY *n*").

When Speakers Are More Logical Than Hearers

In Bidirectional Optimality Theory, the same set of constraints can affect production and comprehension differently (Smolensky, 1996; Hendriks, 2016). In presenting our Asymmetry Account, in line with Hendriks and Spenader (2006), we maintain that Bi-OT has psychological validity and should not be considered merely in a diachronic perspective (cf. Blutner, 2010, and subsequent works).

Specifically, we argue that as soon as children master the two semantic constraints Strength and FaithAll, they start to produce *all* and *some* in an adult-like manner. Importantly, this means that they are able, in production, to use *some* with its upper-bounded meaning (SOME BUT NOT ALL) from a very early age. However, as we have seen, on the basis of Strength and FaithAll, the form *some* happens to be ambiguous in comprehension. Consequently, in the early stages of language acquisition, the child language system cannot distinguish between the two possible interpretations of the quantifier *some*. In order to acquire the ability to comprehend *some* as adults do, children need to acquire the ability to carry out bidirectional optimization, which can be seen as the formalization in OT of perspective-taking (Hendriks et al., 2010). Only when children optimize bidirectionally they can generate an implicature, interpreting *some* with the upper-bounded meaning.

The first prediction of our analysis, then, is the following: no complex inferential process is needed in order to produce *some* with its upper-bounded meaning. In this, a clear asymmetry emerges. The comprehension of *some* requires perspective-taking (in the form of bidirectional optimization) and is thus more complex than the production of *some*. This can explain the findings of the elicitation task of Foppolo and Guasti (2005), Katsos and Smith's (2010) production results, as well as corpus data presented by Eiteljoerge et al. (2018). In all these studies, children demonstrated an ability to produce *some* with its upper-bounded reading at an age at which they cannot yet interpret *some* associating it with its upper-bounded reading.

A related important observation is that the adult-like production of the form *some* with its upper-bounded meaning is not equivalent to the production of an implicature. When speakers produce *some* with its upper-bounded meaning, their production of *some* makes *hearers* generate an implicature in order to arrive at the SOME BUT NOT ALL meaning. Nonetheless, speakers do not generate implicatures themselves. As the language acquisition findings of Eiteljoerge et al. (2018) suggest, for an implicature to emerge, speakers do not need to have an *intention* to produce implicatures (to *implicate*). Consequently, *contra* Hirschberg (1985) and Horn (2006), we believe that shifting the focus from speakers' intentions to hearer's perspective-taking process can be greatly beneficial in defining and studying implicatures. The different roles of hearers and speakers in implicature generation will be discussed further in Section "When Speakers Become Less Logical and More Pragmatic." Before doing that, it is worth considering more in detail the relation between the ability to generate implicatures and a particular cognitive ability: ToM.

Scalar Implicature Generation and ToM

So far, we claimed that the ability to generate implicatures develops with age in parallel with the ability to optimize bidirectionally and that bidirectional optimization can be seen

as a kind of perspective-taking mechanism. In light of this, it is worth considering the role played by ToM, broadly defined as the understanding of others' feelings, desires, intentions, and beliefs (Wellman, 2018).

Evidence suggesting a connection between bidirectional optimization and ToM comes from another production–comprehension asymmetry in child language, namely, the asymmetry observed with object pronouns. Despite the fact that children experience problems with the interpretation of object pronouns until at least the age of 6 (see Hamann, 2011, for an overview), children's pronoun production is almost adult-like from the age of 4;6 (Spenader et al., 2009; cf. De Villiers et al., 2006). This asymmetry has been accounted for in the framework of Bi-OT by Hendriks and Spenader (2006). These authors claimed that pronoun interpretation, but not pronoun production, requires bidirectional optimization, and it is inextricably linked to ToM. Hendriks and Spenader (2006) prediction found experimental support in the study of Kuijper (2016), who demonstrated the existence of a correlation between pronoun interpretation and ToM skills in children. If, as we claim, also SI generation depends on bidirectional optimization, then we expect to find correlations between children's ability to generate SIs (in comprehension) and their ToM abilities. It is worth mentioning that, in contrast, we do not expect ToM to be correlated with the adult-like production of *some*, because production does not require complex inferential processes.

Studies on Autism Spectrum Disorders (ASD) can perhaps be useful in verifying whether, as we argue, the ability to generate SIs relies on ToM. Given that ToM deficits are considered core symptoms of ASD (Baron-Cohen et al., 1985; Frith, 2001), individuals with ASD are expected to show difficulties with SI generation. At first sight, the studies of Pijnacker et al. (2009) and Chevallier et al. (2010) fail to support this hypothesis. In both studies, the performance in SI generation of adult or adolescent participants with ASD did not differ from the one of neurotypicals. Nevertheless, two observations are in order. First of all, in these two studies, participants' ToM levels had not been assessed; hence, it is possible that a correlation between SI generation rate and ToM skills was present at the individual level. Besides, it is conceivable that individuals with ASD adopted a different strategy to generate implicatures, or better, to associate *some* with the SOME BUT NOT ALL meaning, without actually having to perform perspective-taking (cf. Hochstein et al., 2018, on epistemic reasoning in ASD).

Furthermore, some results pointing to a correlation between ToM and SI generation, at least in children, do exist. In Pastor-Cerezuela et al. (2018) the ability to generate different kinds of QBIs was assessed in TD children and children with ASD. The performance of children with ASD was significantly lower than the performance of age-matched and language-matched TD children (see also Surian et al., 1996, on the "deafness" of children with ASD to Gricean Maxims).

Even stronger evidence for the existence of an association between ToM and SIs in language acquisition comes from the recent study of Foppolo et al. (2020). Importantly, this study was the first to systematically assess in monolingual TD children the possible correlations between SI generation, on one hand, and linguistic and cognitive abilities (lexical and morphosyntactic comprehension, IQ, and first-order ToM), on the other. In the group of preschoolers (i.e., before the age of 6, when the ability to generate SIs is still feeble and ToM still developing), first-order ToM abilities (which were found to be independent from lexicon, morphosyntax, and IQ measures) correlated with the ability to generate SIs[2]. Hence, these experimental data speak in favor of an association between ToM and children's SI generation abilities.

This being said, concerning the relationship between perspective-taking and SIs, we should sound a note of caution. It is surely conceivable that bidirectional optimization as a kind of perspective-taking process could gradually become more automatic, not only in individuals with ASD, but also in the neurotypical adult language system (cf. Blutner's, 2006 hypothesis of fossilization). This could indeed speed up and hence facilitate adults' SI generation. As said, it is generally acknowledged that SIs are quite context-independent, precisely because they are rather lexicalized. After all, if "today's morphotactics is yesterday's syntax" (Givón, 1971, p. 25), then today's semantics is yesterday's pragmatics. Nonetheless, we are inclined to believe that SI generation remains a cognitively costly two-step process (see Hendriks, 2014, for the claim that bidirectional optimization can be conceived as a two-step mechanism). Several studies on adults' SI generation seem to support this idea (Huang and Snedeker, 2009; Tomlinson et al., 2013; cf. Chemla and Bott, 2014 and van Tiel et al., 2019). Moreover, processing effort is most likely to re-emerge whenever a SI is canceled. This and the related issue of recursive ToM in adults is the focus of the next section.

When Speakers Become Less Logical and More Pragmatic

We have argued that, for an implicature to emerge, hearers need to take the perspective of the speaker (through bidirectional optimization), whereas speakers can remain "logical." However, this does not necessarily mean that, in general, speakers do not intend implicatures to be generated. Surely, reasoning recursively about their interlocutor, adult speakers *can* intend an implicature.

Let us take Grice's (1989) famous example of a philosophy professor who is asked to provide a recommendation letter and, in describing the abilities of a student who is applying for a philosophy job, writes: "Dear Sir, Mr. X's command of English is excellent, and his attendance at tutorials has been regular. Yours, etc." (Grice, 1989, p. 33). Obviously, in this case, it is evident that the philosophy professor wants the reader to infer that the pupil is actually not suited for the job. Undoubtedly, then, the philosophy professor intends an implicature to be generated. In writing the letter, the professor is reasoning, in a recursive fashion, about the reader's reasoning about the professor's own words. Specifically, the professor wants the reader

[2] *Ad hoc* implicatures, on the other hand, did not correlate with ToM. This brings further support to the hypothesis that children make use of different strategies to generate scalar implicatures and *ad hoc* implicatures (see footnote 1).

to recognize that the professor wants the reader to think that the pupil is not suited for the job. In terms of ToM abilities, this corresponds at least to second-order ToM. Second-order ToM is the ability to understand other people's intentions/beliefs about other people's intentions/beliefs. Complex as it may be, this kind of recursive mind reading is within the reach of the average adult cognitive capacity. Human adults are talented mind readers, despite the fact that this talent appears to have a limit (see Franke and Degen, 2016, for a discussion of recursive reasoning in reference games). Either way, children develop first-order ToM (the ability to understand intention/beliefs) around the age of 4, whereas second-order ToM skills (the ability to understand intentions/beliefs about intentions/beliefs) require at least two more years to develop (Perner and Wimmer, 1985; Sullivan et al., 1994; Tager-Flusberg and Sullivan, 1994) if not more (cf. Flobbe et al., 2008). So, if second-order ToM is involved in those circumstances in which speakers actually intend an implicature to be generated (such as in the case of Grice's philosophy professor), we can predict that 5-year old children are not yet able to do so.

Moreover, it is worth observing that if a speaker intends an implicature to be generated, the speaker should also be able to consciously cancel the implicature. An implicature can be canceled by adding an expression that conveys the negation of what can be inferred via an implicature (Hirschberg, 1985; Grice, 1989; Mayol and Castroviejo, 2013, for discussion):

(14) Julia misses some of her friends. *In fact, not just some. She misses all of them.*

(15) Dear Sir, Mr. X's command of English is excellent, and his attendance at tutorials has been regular. *With this, I do not mean to say he has not great potential as a philosopher. He surely has.*

Canceling an implicature presupposes being able to grasp the difference between the semantic content of the sentence and what can be inferred from the utterance. Thus, as emerges from these examples, implicature cancelation seems impossible without considering the hearer's inference about the speaker's utterance. Because of the fact that this probably requires at least second-order ToM skills, we expect 5-year-old children not to be able to cancel implicatures. Future research could systematically explore this interesting issue.

The claim that the ability to cancel can be considered a litmus test for the ability to generate SIs is particularly relevant for the discussion concerning the division of labor between semantics and pragmatics in non-human animals. In thought-provoking work, Schlenker et al. (2013) argued that Campbell's monkey alarm calls can be analyzed as implicature-like phenomena. In another study, Schlenker et al. (2016a) hypothesized that the system of Putty-nosed monkey alarm sequences could be based on an informativity principle (see also Schlenker et al., 2016b). We believe that such claims are in harmony with the general perspective on SIs outlined in this paper. Adopting a weak scalar term to negate the stronger one does not require complex inferential processes; informativeness alone suffices. Interestingly, Schlenker et al.'s (2016a)' informativity principle appears directly comparable with our constraint Strength. However, turning back to implicature cancelation, it can be observed that human and monkey's "implicatures" are likely to differ substantially. Given their ToM skills, we could confidently claim that non-human primates are unlikely to be able to reach a level of perspective-shifting sophistication that would allow them to cancel implicatures (for an overview on ToM in non-human animals, see Penn et al., 2008).

Explaining Children's Variable Performance in Comprehension Studies

In this section, we will take a closer look at previous studies on children's SI generation and discuss children's performance in light of our Asymmetry Account.

One issue we raised at the beginning of this paper concerns the fact that the experimental manipulation adopted when testing children's comprehension of *some* has substantial influence on children's performance on the task. Most remarkably, at the same age at which children fail to generate SIs in Truth Value Judgment Tasks (TVJTs), they perform adult-like in Felicity Judgment Tasks (FJTs) (Foppolo et al., 2012). Our account straightforwardly explains the difference in results between these two tasks, and in particular the reason why SI generation is not necessary in FJTs. Let us start by considering the TVJT. For children who cannot shift their perspective, *some* is ambiguous according to our Bi-OT account and can be taken as referring to complete as well as non-complete sets. Consequently, they accept sentences such as "Some chipmunks are taking a shower" when shown a picture in which all five chipmunks are taking a shower. The non-adult-like overacceptance of this sentence in such a context stems from children's strict adherence to the constraints Strength and FaithAll. Children's adult-like performance in FJTs is also in line with our account. In this paradigm, children are presented with a visual scenario representing a complete set and asked which of two utterances better matches the visual scenario. One statement contains the quantifier *some*, the other (the most appropriate) the quantifier *all*. The reason why children perform well in this task is that the choice of form does not require perspective-taking and SI generation. When hearing the form *all*, children immediately associate this form with a complete set, thanks to the constraint FaithAll. On the other hand, when hearing the *some*-sentence, they are faced with an ambiguity. On the reasonable assumption that non-ambiguous forms should be preferred to ambiguous ones, it is natural for children to prefer *all* to *some*. In other words, to refer to complete sets, *all* is a better candidate, because it better predicts the complete set.

Katsos and Bishop's (2011) results can arguably be explained along similar lines. In this study, children were tested using both a classical binary TVJT and a ternary judgment task. Instead of rejecting or accepting sentences, in this second paradigm, participants have a middle answer option, and can reward a puppet who utters *some*- and *all*-sentences using a huge, big, or small strawberry. Notably, despite accepting underinformative *some*-sentences in the binary TVJT, in the ternary judgment task, children preferentially chose the middle option.

In light of our account, it can be argued that, because the comprehension of the form *some* results in two possible

meanings (see **Figure 2B**), children can grasp the ambiguity of *some* even without considering the form *all*, and without generating the implicature. Because of this, they are in fact expected to choose a medium-sized reward for *some*-sentence in such contexts. Note that, while we predict that children's SI generation abilities correlate with ToM abilities, we do not predict such a correlation between their performance in ternary judgment task and their ToM abilities. Future research could experimentally test this prediction.

Explaining why Pouscoulous et al.'s (2007) action-based task enhances children's ability to generate SIs is more challenging. However, we will propose a possible explanation. In Pouscoulous et al. 's (2007) action-based task, 4- to 7-year-olds were presented with different scenarios. In critical trials, children were shown an ALL-scenario (five open boxes containing a token each) and heard a puppet uttering "I would like some boxes to contain a token." Children's task was to act on the scenario to comply with the wish of the puppet. Obviously, removing tokens in this scenario means having generated a SI. Children performed extremely well in this AOT. Quite surprisingly, 68% of 4-year-olds and 73% of 5-year-olds demonstrated the ability to grasp the incongruity between the *some*-statement and the ALL-scenario. At first glance, our Bi-OT explanation cannot account for this result. The comprehension of the form *some*, as we have seen, leads to ambiguity, unless a perspective-taking operation takes place. Hence, children's choice should be simply to leave the scenario as it is. However, by carefully considering what children are required to do in the task, we could argue that this kind of AOT does not require simply a comprehension process. On the contrary, this task instead appears to trigger a production process. The reason is the following: children know that they have been asked to act on the scenario (leaving it as it is, or removing tokens, or adding tokens). Hence, in order to act, children need to actively focus on, reflect on, and act on the visual context. Thus, quite naturally, the elements and characteristics of the visual scenario become for them concepts or meanings. When presented with the ALL-scenario, children obviously recognize a complete set. From this, they can carry out a simple production process and choose the form *all*. In Bi-OT terms, it could be said that this kind of AOT does not consist simply in evaluating the optimal interpretation for a given form (as happens in a typical comprehension process); rather, this task consists in selecting the interpretation that gives rise to the *given form as the optimal form*. The task, thus, triggers a production-like process. Consequently, children's performance is enhanced. This is clearly in contrast to what happens in a TVJT, because in this case, children are asked to accept or reject a sentence, and there is nothing in the instructions they receive that can trigger a production process (from a given meaning to a potential form for optimally expressing that meaning).

In general, then, it can be argued that in order to explain children's good performance in particular comprehension tasks, we should always consider whether the instructions or the manipulation encourage children to take as the input a particular meaning, because this is likely to trigger a production process, or, conversely, a particular form, thus triggering a comprehension process. Perhaps the rather vague expression "task demand," often used to justify children's variable performance (e.g., Pouscoulous et al., 2007), reflects exactly this: the more production-like the task is, the less children struggle to generate the SOME BUT NOT ALL meaning.

One last remark, concerning the relationship between our account and other accounts of SI generation is in order here. First, as clearly emerges from Section "Translating Horn Scales in Constraints," the Asymmetry Account attributes a fundamental role to Horn scales. Children's knowledge of Horn scales (hence, of lexical alternatives) is seen as a prerequisite for children's ability to optimize bidirectionally over scalar elements. Therefore, it should be emphasized that our account and lexicalist accounts (Barner et al., 2011; but also Foppolo et al., 2012) are not mutually exclusive, but rather can be integrated in a broader perspective on children's SIs generation. Secondly, we would like to stress that our account, which focuses on children and shows how children's difficulties are to be related to their developing cognitive abilities (in particular, ToM) can complement other accounts of implicature generation in adults. In particular, the constraint-based, probabilistic approach proposed by Degen and Tanenhaus (2015) seems highly compatible with OT models (see the constraint-based, stochastic version of OT developed by Boersma, 1998).

Why Children Interpret *n* as EXACTLY *n*

According to some researchers (e.g., Horn, 1972; Levinson, 1983; cf. Horn, 1992), bare numerals are scalar items whose literal meaning is AT LEAST *n*. The EXACTLY *n* meaning of numerals is derived via the standard process of SI generation, through which scalar elements assume upper-bounded interpretations. The process can be exemplified as follows:

(16) Utterance: The boy caught three crabs.
(17) Non-pronounced alternative: The boy caught four crabs.
(18) Inference: The boy caught exactly three (and not four) crabs.

This neo-Gricean view has been challenged on theoretical grounds by proponents of a seemingly simpler approach to numeral interpretation, according to which numerals are not interpreted as other scalar elements: their primary meaning is EXACTLY *n* (e.g., Geurts, 2006; Breheny, 2008), and the other possible readings, such as the lower-bounded AT LEAST *n*, are secondary and derived in various ways (see Spector, 2013, for a discussion of different approaches).

Importantly, the acquisition literature shows that children's interpretation of numerals does not pattern with their comprehension of other scalar elements. Papafragou and Musolino (2003), for instance, showed that 5-year-old children, despite not being able to reject underinformative sentences containing the scalar term *some* ("Some of the horses jumped over the fence" when three out of three did), tend to reject underinformative sentences containing a numeral ("Two of the horses jumped over the fence" when three out of three did). This result was replicated in various studies. Hurewitz et al.

(2006) tested children aged 3;0 to 4;0 adopting a variation of the classical Picture Selection Task. In line with Papafragou and Musolino's (2003) findings, children performed at chance in the comprehension of *some*, but attributed the exact meaning to the numeral *two*. Even stronger evidence in this direction came from the study of Huang et al. (2013). Adults and children between the ages of 2;6 and 3;5 were tested using a clever paradigm, the Covered Box Task (see Huang et al., 2013, for details). Adults, but not children, demonstrated the ability to generate and also cancel the *some*-implicature. On the other hand, in interpreting the numeral *two*, both children and adults behaved as if an implicature could not be canceled. Both groups always interpreted *two* as only compatible with EXACTLY TWO. Thus, this study shows that it is not simply the case that children merely learn to draw numeral-inferences before the *some*-implicature, as it could have been hypothesized in light of previous studies. Specifically, Huang et al.'s (2013) results yield support for the claim that there exists a true difference between the interpretation of scalar elements such as *some* and the interpretation of numerals. In essence, unlike the former, the latter does not involve implicature generation.

Our account provides a clear explanation as to *why* numerals do not give rise to SIs. Recall that, in our Bi-OT account, Horn scales are defined as scales characterized by the presence of an apex, a culmination point that represents the maximization of the dimensions denoted by the scale. Every time we use a weaker term of a scale, the negated alternative corresponds to the apex. Clearly, if scales do not have apices, they cannot trigger the generation of SIs. Notably, the scale of numerals is unbounded, that is to say, it is a scale without apex. In fact, by definition, in this scale, there cannot be an upper bound: the set of natural numbers, which adults intuitively conceive as being based on the successor function $S(n) = n + 1$, is infinite. Consider again the sentence in (16). The numeral *four* is stronger in terms of logical entailment than *three*. However, so is the numeral *five*, the numeral *six*, and so forth, *ad infinitum*. In our terms, the dimension conveyed by this scale has no maximization point. Hence, given that a single, relevant, strongest alternative cannot be identified (because the alternatives are infinitely many), no single, relevant, strongest alternative can be negated. As a consequence, no comparison between the weaker term and its strongest alternative can take place, and we predict that the interpretation of numerals does not require a process akin to SI generation. In other words, it is because of the very semantics of the apex-less scale of numerals that numerals receive an exact semantics.

One final remark is in order. As an anonymous reviewer rightly points out, in the classical experimental setting adopted to test children's comprehension of numerals, a context-dependent apex can be identified and hence, in principle, an implicature *can* be generated. To give a concrete example, imagine a visual context featuring a boy catching four crabs. Given this scenario, in order to judge an utterance such as (16), "The boy caught three crabs," it is possible for participants to proceed as follows. First, they can consider the four crabs that are visually salient in the context and regard them as an *ad hoc* apex (i.e., an apex that is not based on a particular Horn scale, but that is contextually relevant). Secondly, participants can carry out a perspective-taking process (bidirectionalization). Finally, having generated an implicature and inferred THREE (AND NOT FOUR), they can reject the utterance (16).

Albeit possible, such a bidirectionalization process would require cognitive skills that are not fully developed in young children. Besides, as predicted by the Asymmetry Account, numerals already receive an EXACTLY *n* interpretation as their primary meaning. Consequently, carrying out such a complex inferential process would be not only cognitively costly but also unnecessary: the results of the bidirectionalization process would correspond to the primary meaning of the form.

Hence, despite arguing that the interpretation of numerals does not require implicature generation, our account sheds new light on children's comprehension of numerals. It shows that it is precisely because of the unboundedness of the scale of numerals that implicature generation does not naturally take place. If an implicature did arise, it would be superfluous, and, at least for young children with immature ToM skills, cognitively unfeasible.

CONCLUSION

In this paper, we proposed the Asymmetry Account, a novel account of SI generation in the framework of Bi-OT. The Asymmetry Account is able to explain the rather puzzling asymmetry between production and comprehension of SIs that emerges in language acquisition. Furthermore, it allows us to make a number of interesting predictions. A crucial feature of our hypothesis is that ToM plays a fundamental role in children's comprehension of implicatures, but not in their production. Because of this, children are expected to experience difficulties in comprehension, albeit being able to produce *some* with its upper-bounded meaning from a very young age. Furthermore, our account explains why an extremely variable performance emerges in studies testing children's implicature generation in comprehension: some tasks do not require perspective-taking, or inadvertently elicit a production-like process and enhance children's performance. Moreover, our Asymmetry Account demonstrates that SI generation is not necessary for the interpretation of numerals.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

IM and PH conceived the research questions, developed the theory, and wrote the first version of the manuscript. SAS and SJMK contributed in commenting on and revising the manuscript. All authors approved the submitted manuscript.

REFERENCES

Aloni, M. (2007). "Expressing ignorance or indifference," in *International Tbilisi Symposium on Logic, Language, and Computation*, Vol. 4363, eds B. D. ten Cate and H. W. Zeevat (Berlin: Springer), 1–20. doi: 10.1007/978-3-540-75144-1_1

Barbet, C., and Thierry, G. (2018). When some triggers a scalar inference out of the blue. An electrophysiological study of a Stroop-like conflict elicited by single words. *Cognition* 177, 58–68. doi: 10.1016/j.cognition.2018.03.013

Barner, D., Brooks, N., and Bale, A. (2011). Accessing the unsaid: the role of scalar alternatives in children's pragmatic inference. *Cognition* 118, 84–93. doi: 10.1016/j.cognition.2010.10.010

Baron-Cohen, S., Leslie, A. M., and Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition* 21, 37–46. doi: 10.1016/0010-0277(85)90022-8

Blutner, R. (1998). Lexical pragmatics. *J. Semant.* 15, 115–162. doi: 10.1093/jos/15.2.115

Blutner, R. (2000). Some aspects of optimality in natural language interpretation. *J. Semant.* 17, 189–216. doi: 10.1093/jos/17.3.189

Blutner, R. (2010). "Some experimental aspects of optimality-theoretic pragmatics," in *The Role of Data at the Semantics-Pragmatics Interface*, eds K. Bibok and E. T. Németh (Berlin: De Gruyter).

Blutner, R. (ed.). (2006). "Embedded implicatures and optimality theoretic pragmatics," in *A Festschrift for Kjell Johan Sæbø*, Vol. 2004, (Oslo: Oslo University), 11–29.

Boersma, P. (1998). *Functional Phonology*. Amsterdam: Netherlands Graduate School of Linguistics.

Breheny, R. (2008). A new look at the semantics and pragmatics of numerically quantified noun phrases. *J. Semant.* 25, 93–139. doi: 10.1093/jos/ffm016

Brooks, R., and Meltzoff, A. N. (2005). The development of gaze following and its relation to language. *Dev. Sci.* 8, 535–543. doi: 10.1111/j.1467-7687.2005.00445.x

Carpenter, M., Nagell, K., Tomasello, M., Butterworth, G., and Moore, C. (1998). Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monogr. Soc. Res. Child Dev.* 63, i–vi, 1–143. doi: 10.2307/1166214

Chemla, E., and Bott, L. (2014). Processing inferences at the semantics/pragmatics frontier: disjunctions and free choice. *Cognition* 130, 380–396. doi: 10.1016/j.cognition.2013.11.013

Chevallier, C., Wilson, D., Happé, F., and Noveck, I. A. (2010). Scalar inferences in autism spectrum disorders. *J. Autism. Dev. Disord.* 40, 1104–1117. doi: 10.1007/s10803-010-0960-8

Chierchia, G., Crain, S., Guasti, M. T., Gualmini, A., and Meroni, L. (2001). "The acquisition of disjunction: evidence for a grammatical view of scalar implicatures," in *Proceedings of the 25th Boston University Child Language Development Conference*, eds A. H.-J. Do, L. Dominguez, and A. Johansen (Somerville, MA: Cascadilla Press), 157–168.

Chierchia, G., Crain, S., Guasti, M. T., Thornton, R., and University, B. (1998). *"Some" and "Or": A Study on the Emergence of Logical Form*. Somerville, MA: Cascadilla Press.

Chierchia, G., Fox, D., and Spector, B. (2012). "The grammatical view of scalar implicatures and the relationship between semantics and pragmatics," in *Semantics: An International Handbook of Natural Language Meaning*, Vol. 3, eds P. Portner, C. Maienborn, and K. von Heusinger (Berlin: Mouton de Gruyter), 2297–2332.

Colonnesi, C., Stams, G. J. J. M., Koster, I., and Noom, M. J. (2010). The relation between pointing and language development: a meta-analysis. *Dev. Rev.* 30, 352–366. doi: 10.1016/j.dr.2010.10.001

Davies, C., and Katsos, N. (2010). Over-informative children: production/comprehension asymmetry or tolerance to pragmatic violations? *Lingua* 120, 1956–1972. doi: 10.1016/j.lingua.2010.02.005

de Carvalho, A., Reboul, A. C., Van der Henst, J.-B., Cheylus, A., and Nazir, T. (2016). Scalar implicatures: the psychological reality of scales. *Front. Psychol.* 7:1500. doi: 10.3389/fpsyg.2016.01500

De Villiers, J., Cahillane, J., and Altreuter, E. (2006). "What can production reveal about Principle B," in *Proceedings of the Inaugural Conference on Generative Approaches to Language Acquisition–North America*, Vol. 1, Honolulu, HI, 89–100.

Degen, J. (2013). *Alternatives in Pragmatic Reasoning*. Ph.D. thesis, University of Rochester, Rochester, NY.

Degen, J., and Tanenhaus, M. K. (2015). Processing scalar implicature A constraint-based approach. *Cogn. Sci.* 39, 667–710. doi: 10.1111/cogs.12171

Eiteljoerge, S. F. V., Pouscoulous, N., and Lieven, E. V. M. (2018). Some pieces are missing: implicature production in children. *Front. Psychol.* 9:1928. doi: 10.3389/fpsyg.2018.01928

Flobbe, L., Verbrugge, R., Hendriks, P., and Krämer, I. (2008). Children's application of theory of mind in reasoning and language. *J. Logic Lang. Inf.* 17, 417–442. doi: 10.1007/s10849-008-9064-7

Foppolo, F., and Guasti, M. T. (2005). "Children's comprehension of sentences involving scalar items," in *Contributions to the Thirtieth Incontro di Grammatica Generativa*, eds L. Brugè, G. Giusti, N. Munaro, W. Schweikert, and G. Turano (Venice: Libreria Editrice Cafoscarina), 13–24.

Foppolo, F., Guasti, M. T., and Chierchia, G. (2012). Scalar implicatures in child language: give children a chance. *Lang. Learn. Dev.* 8, 365–394. doi: 10.1080/15475441.2011.626386

Foppolo, F., Mazzaggio, G., Panzeri, F., and Surian, L. (2020). Scalar and ad-hoc pragmatic inferences in children: guess which one is easier. *J. Child Lang.* doi: 10.1017/S030500092000032X [Epub ahead of print].

Franke, M., and Degen, J. (2016). Reasoning in reference games: individual- vs. population-level probabilistic modeling. *PLoS One* 11:e0154854. doi: 10.1371/journal.pone.0154854

Frith, U. (2001). Mind blindness and the brain in autism. *Neuron* 32, 969–979. doi: 10.1016/S0896-6273(01)00552-9

Gazdar, G. (1979). Pragmatics and logical form. *J. Pragmat.* 4, 1–13. doi: 10.1016/0378-2166(80)90014-4

Gazdar, G. (1980). Pragmatics, implicature, presuposition and logical form. *Critica* 12, 113–122. doi: 10.22201/iifs.18704905e.1980.341

Geurts, B. (2006). "Take 'five': the meaning and use of a number word," in *Non-Definiteness and Plurality*, eds S. Vogeleer and L. Tasmowski (Amsterdam: Benjamins), 311–329. doi: 10.1075/la.95.16geu

Geurts, B. (2010). *Quantity Implicatures*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511975158

Givón, T. (1971). Historical syntax and synchronic morphology: an archaeologist's field trip. *Chic. Linguist. Soc.* 7, 394–415.

Grice, H. P. (1975). "Logic and conversation", in *Syntax and Semantics*, Vol. 3, eds P. Cole and J. Morgan (New York, NY: Academic Press), 41–58.

Grice, H. P. (1989). *Studies in the Way of Words*. Cambridge, MA: Harvard university press.

Grosse, G., Behne, T., Carpenter, M., and Tomasello, M. (2010). Infants communicate in order to be understood. *Dev. Psychol.* 46, 1710–1722. doi: 10.1037/a0020727

Guasti, M. T., Chierchia, G., Crain, S., Foppolo, F., Gualmini, A., and Meroni, L. (2005). Why children and adults sometimes (but not always) compute implicatures. *Lang. Cogn. Process.* 20, 667–696. doi: 10.1080/01690960444000250

Hamann, C. (2011). "Binding and coreference: views from child language," in *Handbook of Generative Approaches to Language Acquisition*, Vol. 41, eds J. de Villiers and T. Roeper (Dordrecht: Springer), 247–290. doi: 10.1007/978-94-007-1688-9_7

Hendriks, P., and Spenader, J. (2006). When production precedes comprehension: an optimization approach to the acquisition of pronouns. *Lang. Acquis.* 13, 319–348. doi: 10.1207/s15327817la1304_3

Hendriks, P. (2014). *Asymmetries between Language Production and Comprehension*. Dordrecht: Springer Netherlands. doi: 10.1007/978-94-007-6901-4

Hendriks, P. (2016). "Unfaithful conduct: a competence-based explanation of asymmetries between production and comprehension," in *Optimality-Theoretic Syntax, Semantics, and Pragmatics*, eds G. Legendre, M. T. Putnam, H. de Swart, and E. Zaroukian (Oxford: Oxford University Press), 300–324. doi: 10.1093/acprof:oso/9780198757115.003.0013

Hendriks, P., De Hoop, H., Krämer, I., De Swart, H., and Zwarts, J. (2010). *Conflicts in Interpretation*. London: Equinox Publishing.

Hirschberg, J. B. (1985). *A Theory of Scalar Implicature*. Ph.D. thesis, University of Pennsylvania, Philadelphia, PA.

Hochstein, L., Bale, A., and Barner, D. (2018). Scalar implicature in absence of epistemic reasoning? The case of autism spectrum disorder. *Lang. Learn. Dev.* 14, 224–240. doi: 10.1080/15475441.2017.1343670

Hogeweg, L. (2009). The meaning and interpretation of the Dutch particle wel. *J. Pragmat.* 41, 519–539. doi: 10.1016/j.pragma.2008.06.012

Horn, L. (1972). *On the Semantic Properties of the Logical Operators in English*. Ph.D. thesis, UCLA, Los Angeles, CA.

Horn, L. R. (1992). The said and the unsaid. *Semant. Linguist. Theory* 2, 163–192. doi: 10.3765/salt.v2i0.3039

Horn, L. R. (2006). "Implicature," in *The Handbook of Pragmatics*, eds L. R. Horn and G. Ward (Oxford: Blackwell Publishing Ltd), 2–28. doi: 10.1002/9780470756959.ch1

Horowitz, A. C., Schneider, R. M., and Frank, M. C. (2018). The trouble with quantifiers: exploring children's deficits in scalar implicature. *Child Dev.* 89, e572–e593. doi: 10.1111/cdev.13014

Huang, Y. T., and Snedeker, J. (2009). Semantic meaning and pragmatic interpretation in 5-year-olds: evidence from real-time spoken language comprehension. *Dev. Psychol.* 45, 1723–1739. doi: 10.1037/a0016704

Huang, Y. T., Spelke, E., and Snedeker, J. (2013). What exactly do numbers mean? *Lang. Learn. Dev.* 9, 105–129. doi: 10.1080/15475441.2012.658731

Hurewitz, F., Papafragou, A., Gleitman, L., and Gelman, R. (2006). Asymmetries in the acquisition of numbers and quantifiers. *Lang. Learn. Dev.* 2, 77–96. doi: 10.1207/s15473341lld0202_1

Katsos, N., and Bishop, D. V. M. (2011). Pragmatic tolerance: implications for the acquisition of informativeness and implicature. *Cognition* 120, 67–81. doi: 10.1016/j.cognition.2011.02.015

Katsos, N., and Smith, N. (2010). "Pragmatic tolerance or a speaker-comprehender asymmetry in the acquisition of informativeness?," in *Proceedings of the 34th Annual Boston University Conference on Language Development*, (Somerville, MA: Cascadilla Press), 221–232.

Katzir, R. (2007). Structurally-defined alternatives. *Linguist. Philos.* 30, 669–690. doi: 10.1007/s10988-008-9029-y

Krifka, M. (2007). "Approximate interpretations of number words: a case for strategic communication," in *Proceedings of the Colloquium*, Amsterdam, 27–28.

Krifka, M. (2010). "A note on an asymmetry in the hedonic implicatures of olfactory and gustatory terms," in *Between the Regular and the Particular in Speech and Language*, eds S. Fuchs, P. Hoole, C. Mooshammer, and M. Żygis (Frankfurt am Main: Peter Lang), 235–245.

Krifka, M. (2011). "Be brief and vague! And how bidirectional optimality theory allows for verbosity and precision," in *Sounds and Systems*, eds D. Restle and D. Zaefferer (Berlin: De Gruyter Mouton). doi: 10.1515/9783110894653.439

Kuijper, S. (2016). *Communication Abilities of Children with ASD and ADHD*. Ph.D. thesis, University of Groningen, Groningen.

Levinson, S. (1983). *Pragmatics*. Cambridge: Cambridge University Press.

Levinson, S. C. (2000). *Presumptive Meanings: The theory of Generalized Conversational Implicature*. Cambridge, MA: MIT press.

Lidz, J., and Musolino, J. (2002). Children's command of quantification. *Cognition* 84, 113–154. doi: 10.1016/S0010-0277(02)00013-6

Liebal, K., Behne, T., Carpenter, M., and Tomasello, M. (2009). Infants use shared experience to interpret pointing gestures. *Dev. Sci.* 12, 264–271. doi: 10.1111/j.1467-7687.2008.00758.x

Magri, G. (2009). A theory of individual-level predicates based on blind mandatory scalar implicatures. *Nat. Lang. Semant.* 17, 245–297. doi: 10.1007/s11050-009-9042-x

Matsumoto, Y. (1995). The conversational condition on horn scales. *Linguist. Philos.* 18, 21–60. doi: 10.1007/BF00984960

Matthews, D. (2014). *Pragmatic Development in First Language Acquisition*, Vol. 10. Amsterdam: John Benjamins Publishing Company.

Mayol, L., and Castroviejo, E. (2013). How to cancel an implicature. *J. Pragmat.* 50, 84–104. doi: 10.1016/j.pragma.2013.02.002

Mognon, I., Sprenger, S., Kuijper, S., and Hendriks, P. (in press). "Balancing the (Horn) scale: explaining the production-comprehension asymmetry for scalar implicatures," in *Proceedings of the University of Pennsylvania Working Papers in Linguistics*, Vol. 27, Philadelphia, PA .

Noveck, I. A. (2001). When children are more logical than adults: experimental investigations of scalar implicature. *Cognition* 78, 165–188. doi: 10.1016/S0010-0277(00)00114-1

Papafragou, A., and Musolino, J. (2003). Scalar implicatures: experiments at the semantics-pragmatics interface. *Cognition* 86, 253–282. doi: 10.1016/S0010-0277(02)00179-8

Paris, S. G. (1973). Comprehension of language connectives and propositional logical relationships. *J. Exp. Child Psychol.* 16, 278–291.

Pastor-Cerezuela, G., Tordera Yllescas, J. C., González-Sala, F., Montagut-Asunción, M., and Fernández-Andrés, M.-I. (2018). Comprehension of generalized conversational implicatures by children with and without autism spectrum disorder. *Front. Psychol.* 9:272. doi: 10.3389/fpsyg.2018.00272

Penn, D. C., Holyoak, K. J., and Povinelli, D. J. (2008). Darwin's mistake: explaining the discontinuity between human and nonhuman minds. *Behav. Brain Sci.* 31, 109–130; discussion 130–178. doi: 10.1017/S0140525X08003543

Perner, J., and Wimmer, H. (1985). "*John thinks* that Mary *thinks* that." attribution of second-order beliefs by 5- to 10-year-old children. *J. Exp. Child Psychol.* 39, 437–471. doi: 10.1016/0022-0965(85)90051-7

Pijnacker, J., Hagoort, P., Buitelaar, J., Teunisse, J. P., and Geurts, B. (2009). Pragmatic inferences in high-functioning adults with autism and Asperger syndrome. *J. Autism Dev. Disord.* 39, 607–618. doi: 10.1007/s10803-008-0661-8

Pouscoulous, N., Noveck, I. A., Politzer, G., and Bastide, A. (2007). A developmental investigation of processing costs in implicature production. *Lang. Acquis.* 14, 347–375. doi: 10.1080/10489220701600457

Prince, A., and Smolensky, P. (2004). *Optimality Theory: Constraint Interaction in Generative Grammar*. Malden, MA: Blackwell. doi: 10.1002/9780470759400

Schaeken, W., Haeren, M. V., and Bambini, V. (2018). The understanding of scalar implicatures in children with autism spectrum disorder: dichotomized responses to violations of informativeness. *Front. Psychol.* 9:1266. doi: 10.3389/fpsyg.2018.01266

Schlenker, P., Chemla, E., Arnold, K., Lemasson, A., Ouattara, K., Keenan, S., et al. (2013). Monkey semantics: two "dialects" of Campbell's monkey alarm calls. *Linguist. Philos.* 37, 439–501. doi: 10.1007/s10988-014-9155-7

Schlenker, P., Chemla, E., Arnold, K., and Zuberbühler, K. (2016a). Pyow-hack revisited: two analyses of Putty-nosed monkey alarm calls. *Lingua* 171, 1–23. doi: 10.1016/j.lingua.2015.10.002

Schlenker, P., Chemla, E., Schel, A. M., Fuller, J., Gautier, J. P., Kuhn, J., et al. (2016b). Formal monkey linguistics. *Theor. Linguist.* 42, 1–90. doi: 10.1515/tl-2016-0001

Schulz, K., and Van Rooij, R. (2006). Pragmatic meaning and non-monotonic reasoning: the case of exhaustive interpretation. *Linguist. Philos.* 29, 205–250. doi: 10.1007/s10988-005-3760-4

Skordos, D., and Papafragou, A. (2016). Children's derivation of scalar implicatures: alternatives and relevance. *Cognition* 153, 6–18. doi: 10.1016/j.cognition.2016.04.006

Smith, C. L. (1980). Quantifiers and question answering in young children. *J. Exp. Child* 30, 191–205. doi: 10.1016/0022-0965(80)90057-0

Smolensky, P. (1996). On the comprehension/production dilemma in child language. *Linguist. Inq.* 27, 720–731.

Spector, B. (2013). Bare numerals and scalar implicatures. *Lang. Linguist. Compass* 7, 273–294. doi: 10.1111/lnc3.12018

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

Sperber, D., and Wilson, D. (1986). *Relevance: Communication and Cognition*, Vol. 142. Cambridge, MA: Harvard University Press.

Stiller, A. J., Goodman, N. D., and Frank, M. C. (2015). Ad-hoc implicature in preschool children. *Lang. Learn. Dev.* 11, 176–190. doi: 10.1080/15475441.2014.927328

Sullivan, K., Zaitchik, D., and Tager-Flusberg, H. (1994). Preschoolers can attribute second-order beliefs. *Dev. Psychol.* 30, 395–402. doi: 10.1037/0012-1649.30.3.395

Sullivan, J., Davidson, K., Wade, S., and Barner, D. (2019). Differentiating scalar implicature from exclusion inferences in language acquisition. *J. Child Lang.* 46, 733–759. doi: 10.1017/S0305000919000096

Surian, L., Baron-Cohen, S., and Van Der Lely, H. (1996). Are children with autism deaf to Gricean maxims? *Cogn. Neuropsychiatry* 1, 55–72. doi: 10.1080/135468096396703

Tager-Flusberg, H., and Sullivan, K. (1994). A second look at second-order belief attribution in autism. *J. Autism. Dev. Disord.* 24, 577–586. doi: 10.1007/BF02172139

Tomlinson, J. M., Bailey, T. M., and Bott, L. (2013). Possibly all of that and then some: scalar implicatures are understood in two steps. *J. Mem. Lang.* 69, 18–35. doi: 10.1016/j.jml.2013.02.003

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 Tiel, B., Pankratz, E., and Sun, C. (2019). Scales and scalarity: processing scalar inferences. *J. Mem. Lang.* 105, 93–107. doi: 10.1016/j.jml.2018.12.002

Wellman, H. M. (2018). Theory of mind: the state of the art. *Eur. J. Dev. Psychol.* 15, 728–755. doi: 10.1080/17405629.2018.1435413

Zeevat, H. (2000). The asymmetry of optimality theoretic syntax and semantics. *J. Semant.* 17, 243–262. doi: 10.1093/jos/17.3.243

Conflict of Interest: 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 © 2021 Mognon, Sprenger, Kuijper and Hendriks. 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) and the copyright owner(s) 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 2021
doi: 10.3389/fpsyg.2020.556624

The Role of Literal Features During Processing of Novel Verbal Metaphors

Camilo R. Ronderos[1]*, *Ernesto Guerra*[2] and *Pia Knoeferle*[1,3,4]

[1] Institut für Deutsche Sprache und Linguistik, Humboldt-Universität zu Berlin, Berlin, Germany, [2] Center for Advanced Research in Education, Institute of Education, Universidad de Chile, Santiago, Chile, [3] Einstein Center for Neurosciences Berlin, Berlin, Germany, [4] Berlin School of Mind and Brain, Berlin, Germany

When a word is used metaphorically (for example "walrus" in the sentence "The president is a walrus"), some features of that word's meaning ("very fat," "slow-moving") are carried across to the metaphoric interpretation while other features ("has large tusks," "lives near the north pole") are not. What happens to these features that relate only to the literal meaning during processing of novel metaphors? In four experiments, the present study examined the role of the feature of physical containment during processing of verbs of physical containment. That feature is used metaphorically to signify difficulty, such as "fenced in" in the sentence "the journalist's opinion was fenced in after the change in regime." Results of a lexical decision task showed that video clips displaying a ball being trapped by a box facilitated comprehension of verbs of physical containment when the words were presented in isolation. However, when the verbs were embedded in sentences that rendered their interpretation metaphorical in a novel way, no such facilitation was found, as evidenced by two eye-tracking reading studies. We interpret this as suggesting that features that are critical for understanding the encoded meaning of verbs but are not part of the novel metaphoric interpretation are ignored during the construction of metaphorical meaning. Results and limitations of the paradigm are discussed in relation to previous findings in the literature both on metaphor comprehension and on the interaction between language comprehension and the visual world.

Keywords: **verbal metaphors, eye-tracking, experimental pragmatics, figurative language comprehension, metaphor processing**

OPEN ACCESS

Edited by:
Valentina Cuccio,
University of Messina, Italy

Reviewed by:
Carolina Gattei,
Universidad Torcuato Di Tella, Argentina
Francesca M. M. Citron,
Lancaster University, United Kingdom

***Correspondence:**
Camilo R. Ronderos
camilo.rodriguez.ronderos@gmail.com

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

Received: *07 May 2020*
Accepted: *15 December 2020*
Published: *26 January 2021*

Citation:
Ronderos CR, Guerra E and Knoeferle P (2021) The Role of Literal Features During Processing of Novel Verbal Metaphors. Front. Psychol. 11:556624. doi: 10.3389/fpsyg.2020.556624

1. INTRODUCTION

In conversation, speakers usually use words in a way that is close to the word's conventional meaning. When this is the case, listeners are assumed to retrieve this word from their mental lexicon in order to grasp the meaning intended by the speaker. But what happens in a listener's mind when words are used in a previously unheard sense that requires a rapid integration of context in order to be understood? Such is the case of novel metaphors:

1. It was difficult for the journalist to see his opinion fenced-in after the change in regime.

In (1), a verb of physical confinement (*fenced-in*) is used to predicate over an abstract noun which does not have a physical dimension (*the journalist's opinion*), yet the intended meaning can be readily derived: The journalist is no longer allowed to speak freely. In this example, the feature of "physical confinement" is not part of the metaphor and is even incompatible with the speaker's intended meaning.

The role ascribed to features that relate only to the literal meaning of a word and are incompatible with that word's novel metaphorical meaning (henceforth "literal features") during processing varies depending on the theoretical perspective (see Holyoak and Stamenković, 2018 for a systematic review of competing views). Some accounts see metaphor comprehension as a type of category inclusion: They claim that understanding a metaphor, such as *The president is a walrus* involves a contextual adjustment of the meaning of the metaphoric vehicle (*walrus*) on the basis of the dimensions provided by the metaphoric topic (*The president*). Language comprehenders thus create a new, occasion-specific category (McGlone and Manfredi, 2001; Glucksberg, 2003; Rubio Fernandez, 2007; Sperber and Wilson, 2008), an idea inspired by Barsalou's work on *ad hoc* categories (Barsalou, 1983). This type of meaning modulation unfolds via rapid suppression of incompatible literal features (e.g., "has large tusks," "lives near the north pole") and enhanced activation of only those features that are compatible with the dimensions provided by the metaphoric topic and are relevant for interpretation (e.g., "very fat," "slow-moving") (Gernsbacher et al., 2001).

A competing set of views sees metaphor understanding as a process of indirect comparison. When encountering a metaphor, we reason analogically about the conceptual structure of both topic and vehicle in order to reach a final utterance interpretation (Gentner and Holyoak, 1997; Wolff and Gentner, 2000, 2011; Coulson and Oakley, 2005; Gentner and Bowdle, 2008). A necessary first step in this process is that of structural alignment: Topic and vehicle are scanned for commonalities in their structures, and only after these commonalities have been established, inferences are projected from vehicle to topic. Here, metaphor-incompatible features of the vehicle are not immediately suppressed and can only be discarded after structural alignment has been achieved (McGlone and Manfredi, 2001). The "career of metaphor" hypothesis (Bowdle and Gentner, 2005), an extension of the indirect comparison view, claims that there is a difference in processing between novel and conventional metaphors. For conventional metaphors, they claim that meaning is not constructed via analogical reasoning but is instead retrieved via category selection. Researchers working within the framework of category inclusion, however, have argued against this providing evidence suggesting that not conventionality but aptness (i.e., how "good" a metaphor is) determines a metaphor's processing mode, meaning that there should not be an a priori difference in processing route between novel and conventional metaphors (Jones and Estes, 2006)[1].

Several studies have dealt with whether these literal features are activated or suppressed during processing (and if so, when). As a whole, the results do not unequivocally support one or the other set of accounts (e.g., Gernsbacher et al., 2001; McGlone and Manfredi, 2001; Rubio Fernandez, 2007; Weiland et al., 2014). We argue that three common features of these studies could be improved upon when striving for consensus. Firstly, these studies restricted their investigations to sentences, such as "Some lawyers are sharks" (known in the literature as nominal metaphors), in which both metaphoric topic and vehicle are nouns and they have the surface form of a category statement. Considering that metaphors in the wild can take a wide range of morphosyntactic forms (see for example Bambini et al., 2019), it is problematic for theory development to consider only a small subset of metaphors.

Secondly, these studies usually make use of materials in which the relation between the metaphors and the tested literal features varies for every item. For example, two of the metaphoric items from McGlone and Manfredi (2001) (one of the most prominent studies on the role of literal features during metaphor comprehension) were *some stomachs are barrels* and *some cats are princesses*. The study examined the relationship between these sentences and the literal features captured in the sentences *barrels can be wooden* and *cats can be siamese*, respectively. *Wooden* and *siamese* are very different types of properties that require different kinds of world knowledge from a listener, and it is unclear to what extent we can meaningfully compare the relationship of each of these literal sentences to its metaphoric counterpart. It could be the case that variation in the relationship between literal features and target metaphors across experimental items is (at least partially) responsible for some of the contradictory results in the literature. A similar argument was made by Thibodeau and Durgin (2008) with regards to the difference in results of their study (facilitation effect of conventional metaphors on processing subsequent related novel metaphors) when compared to the results of Keysar et al. (2000) (no facilitation effect of conventional metaphors on processing subsequent related novel metaphors).

Finally, the majority of experiments investigating the role of literal features of a metaphor have been conducted using sentence reading times or reaction times as the dependent measures (but see Weiland et al., 2014, for a notable exception). As a result, the timing of the activation of literal feature representations remains unclear and should be addressed with a finer-grained method. With the present set of studies we intend to make a contribution to the debate on the role of literal features during metaphor processing by improving on these three issues.

Concretely, we set out to study the role of conceptual features that are part of the encoded meaning of a verb but are incompatible with its novel metaphoric interpretation: We conducted a series of experiments investigating the role of the specific feature of physical containment during processing of novel verbal metaphors, such as (1). In these metaphors, the vehicle is always a verb of physical containment used to signify difficulty. This allowed us to use the same animated videos displaying physical containment as a visual representation of the same literal feature across items. We based our paradigm and hypotheses on insights coming from psycholinguistic accounts of metaphor comprehension (Glucksberg, 2003; Gentner and Bowdle, 2008), as well as from research on metaphor production (Sato et al., 2015). Crucially, we relied on the insights and on the methodology of research conducted on the interaction of

[1]Whether aptness or conventionality modulates the processing route is still a matter of debate and outside of the scope of the current investigation, which focuses on novel verbal metaphors only. For in-depth discussions on the role these factors might play during processing see Gentner and Bowdle (2008), Glucksberg (2008), Holyoak and Stamenković (2018), and Pouscoulous and Dulcinati (2019).

(written) language processing and the visual context (Guerra and Knoeferle, 2014, 2017) to create our experimental paradigm.

The paper is structured as follows: The next two subsections provide an overview of the different views on metaphor processing and their predictions and briefly introduce the literature on the interaction between language processing and the visual world. We then present two eye-tracking during reading studies, one self-paced reading experiment and one lexical decision task, all investigating to what extent a depiction of physical containment influences the processing of novel verbal metaphors. Results are discussed in light of the background presented in section 1.

1.1. Understanding Metaphors

An issue of importance for metaphor theories is the role of the literal meaning of a metaphoric vehicle during processing: In (1), the verb *fenced-in* entails the concept of physical containment; its direct object is something that is not allowed to physically move. However, when we hear that *the journalist's opinion has been fenced-in*, the feature of a physical barrier is not part of the final interpretation. What happens to this **literal feature** during comprehension?

From a category inclusion perspective, the noun *opinion* in (1) provides the dimension of [+ abstract]. This dimension, together with the relevant utterance context, determines the interpretation of the verb: relevant features are selected while irrelevant ones are actively discarded. Evidence for this view comes from priming experiments. Gernsbacher et al. (2001) showed participants either a metaphoric or a literal sentence as a prime (*That defense lawyer is a shark* or *That large hammerhead is a shark*) and then asked them to perform a verification task on a sentence describing a feature of the vehicle that was irrelevant or relevant for the construction of the metaphoric meaning (*sharks are good swimmers* or *sharks are tenacious*). They found that, after reading metaphorical primes, participants were faster at verifying sentences describing a relevant feature for the metaphoric interpretation compared to when they read a literal prime. They also found that verifying sentences about a metaphor-irrelevant property took longer after reading a metaphor than after reading a literal statement. They interpreted these results in terms of activation of relevant features and suppression of irrelevant ones: When the word *shark* is used metaphorically, features, such as "tenacious" are enhanced and features, such as "good swimmer" are inhibited.

Rubio Fernandez (2007) conducted a similar study with the key difference that the target was a single word and it was shown at varying intervals. She found that at early intervals (0 and 400 ms) irrelevant literal features were primed by the metaphor and only actively suppressed when presented 1,000 ms after the prime. McGlone and Manfredi (2001) deployed a reversed version of this paradigm and showed participants irrelevant or relevant features as primes and then metaphorical sentences as targets. They found that relevant features facilitated whereas irrelevant features hindered comprehension compared to a baseline condition without a prime, suggesting that irrelevant properties are suppressed early on during processing. Weiland et al. (2014) created an ERP version of this paradigm: they showed participants a masked prime consisting of a word representing an irrelevant feature (*furry*) of a metaphor (*my lawyer is a hyena*) followed by the metaphor itself. They found that the N400 effect (computed as the difference in stimulus-related average electrical responses between the metaphor and a literal equivalent) was reduced when participants saw the irrelevant prime compared to when they did not see any prime at all, suggesting that irrelevant features can indeed ease comprehension of a metaphor, a result which is in conflict with that of McGlone and Manfredi (2001).

From the perspective of indirect comparison, on the other hand, the activation of relevant and irrelevant features of the vehicle are not contingent upon dimensions provided by the topic. Gentner and Holyoak (1997), Gentner et al. (2001), Bowdle and Gentner (2005), Gentner and Bowdle (2008) have argued that, during initial stages of comprehension, the elements of a novel metaphor are scanned for structural similarities: listeners reason analogically about the relationship between vehicle and topic. This requires irrelevant features of the vehicle to be initially activated and only suppressed or ignored during later stages, once structural alignment has already taken place (Gentner and Bowdle, 2008). This view is compatible with the findings of Weiland et al. (2014) but incompatible with those of McGlone and Manfredi (2001). According to the indirect comparison view, it is also likely that literal features remain active after a metaphor has been understood, because the pattern of structural mappings between topic and vehicle can be used for subsequent processing, as has been shown to be the case for extended metaphors. For these, words belonging to the same semantic domain are used to "extend" a metaphoric expression beyond a single topic-vehicle pairing, as in the famous lines from Shakespeare's *As you like it*: "All the world's a stage and all the men and women merely players; they have their exits and their entrances, and one man in his time plays many parts." Support for this view comes from priming paradigms, where it has been shown that novel metaphors facilitate processing of subsequent novel metaphors that share the same conceptual mappings between domains (Keysar et al., 2000) and even that conventional metaphors can prime subsequent related novel metaphors (Thibodeau and Durgin, 2008).

Findings on extended metaphors are somewhat challenging to account for from the perspective of category inclusion, which seems to posit that metaphor comprehension occurs only locally: If the meaning of the metaphoric vehicle is altered so that irrelevant literal features are suppressed, how can these features be re-activated to prime subsequent related metaphors? One answer, coming from within Relevance Theory, is given by Carston (2010). She claims that, in an extended metaphor, the multiple related words that are semantically associated are mutually reinforcing, resulting in an enhanced activation of the literal meaning (which she calls the "lingering" of the literal meaning). This can lead to the entire literal meaning of the extended metaphor to be meta-represented and considered as a sort of "imaginary world," where the individual metaphors are understood literally. This activates a second processing route for extended metaphors where metaphoric meaning is only derived in later stages of processing (Rubio-Fernández et al., 2016).

Regarding the activation of literal features, the difference between indirect comparison views and Carston (2010) seems to be that Carston (2010) might predict a facilitation effect of metaphors on subsequent related metaphors based on semantic reinforcement of related words, whereas Gentner et al. (2001) predicts a general activation of structural mapping patterns after any metaphor has been activated. In other words, the indirect comparison view predicts that literal features of a metaphor remain active after a metaphor is understood because these are part of a complex network of mappings between the encoded meanings of the metaphoric topic and the metaphoric vehicle. The category inclusion view of Carston (2010), on the other hand, predicts activation of the encoded semantic features of a metaphoric vehicle (i.e., "lingering" of the literal meaning), not of a network of systematic mappings.

In short, there is a lack of consensus in the literature on the timing of suppression and activation of literal features during and after metaphor comprehension: whereas, it has been suggested that irrelevant literal features hinder processing (McGlone and Manfredi, 2001) and are immediately suppressed after comprehension (Gernsbacher et al., 2001), others claim that literal features can ease subsequent processing of a metaphor (Weiland et al., 2014), remain active for at least 400 ms after processing (Rubio Fernandez, 2007), and even facilitate processing of subsequent related metaphors (Thibodeau and Durgin, 2008). It is therefore important to seek out more evidence in this debate since it has repercussions for theory development, as highlighted above.

It is crucial to note that the activation of literal features during metaphor comprehension could be affected by item-specific factors, such as a metaphor's conventionality (i.e., the subjective frequency of exposure to a specific vehicle in its metaphoric meaning) (e.g., Blasko and Connine, 1993; Wolff and Gentner, 2000; Bowdle and Gentner, 2005). It could, for instance, be the case that literal features facilitate access for novel metaphors and hinder comprehension for more conventional metaphors (in line with the "career of metaphor" hypothesis, Bowdle and Gentner, 2005). McGlone and Manfredi (2001), for example, found that the metaphors in their study that were rated as less conventional displayed less interference from irrelevant literal features during processing compared to the metaphors that were rated as more conventional. The effect was nevertheless one of interference, for both novel and conventional metaphors separately (−8 and −143 ms, respectively) and when taken as a whole. Weiland et al. (2014), on the other hand, controlled for conventionality by selecting only metaphors that were rated as being halfway between highly novel and highly conventional for their experiments, and did not report any mediating effect of conventionality. Gernsbacher et al. (2001) operationalized conventionality as the percentage of comprehension errors for each of the metaphors in their study. They found that it did not correlate with the effect size of each of the items in their experiment, suggesting that conventionality did not modulate the way that literal features were suppressed after metaphor comprehension. Finally, Rubio Fernandez (2007) did not report having controlled for conventionality. This specific literature therefore does not strongly suggest that conventionality mediates the role of literal features during metaphor comprehension. Furthermore, it is still an open question whether conventionality actually modulates processing, or whether it only appears to do so because it tends to be correlated with aptness (i.e., the degree to which the figurative meaning of the vehicle captures relevant properties of the topic, or how "good" the metaphor is), which has been claimed by some to be the true underlying factor that mediates metaphor processing (Jones and Estes, 2006; Glucksberg, 2008). An investigation on the effect of conventionality on the activation of literal features is beyond the scope of our current investigation, which focuses on the processing of novel metaphors exclusively.

Specifically, our contribution to the debate on the activation of literal features is to examine the effect of pre-activating said features on the processing of subsequent verbal metaphors, which, unlike nominal metaphors, have been largely overlooked in the literature. We do this by showing participants short animated clips of the literal feature of containment prior to participants reading verbal metaphors that entail this feature as part of their literal meaning. Our study makes use of eye-tracking during reading and draws its inspiration from research on the relation between visual attention and language production and processing. We will now turn to a brief overview of this specific research field.

1.2. The Interaction of Visual and Linguistic Information During Sentence Processing

Given the lack of converging evidence coming from the studies described above, we turned to neighboring disciplines for inspiration. One possibility is to draw from research on language-vision interactions (see Knoeferle and Guerra, 2016 for an introduction to this field). The seminal work of Cooper (1974) showed that there is a close temporal adjacency between language understanding and the processing of visual stimuli. In the study, participants heard stories while simultaneously being presented with images of potential referents while their eye movements were monitored, something that years later came to be known as the Visual World Paradigm (for a review, see Huettig et al., 2011). The results of this study showed that participants looked at the visual representations of objects immediately after they were mentioned in a story, highlighting the rapid and automatic way in which language and visual processes interact.

Through eye-tracking technology it has also been shown that the processing of visual stimuli interacts with the processing of written abstract language. Guerra and Knoeferle (2014) showed participants a video of two playing cards that either moved closer together or further apart. Participants then read German sentences that dealt with semantic dissimilarity, such as *Frieden und Krieg sind bestimmt verschieden* ("Peace and war are certainly different") or similarity, such as *Kampf und Krieg sind freilich entsprechend* ("Battle and war are certainly similar"). Their results showed that when the motion of the cards was conceptually aligned with the direction of the semantic relation (close~similar; far~different), participants were faster at reading the second of the presented nouns (Experiment 3) as well as the adjective (Experiments 1 and 2) than when there was no such

conceptual alignment. The result was interpreted as evidence for an abstract co-indexing link between spatial distance and semantic similarity. One characteristic of the eye-tracking during reading method is that it allows for a rough mapping of the results onto different stages of language processing (Clifton et al., 2007; see Vasishth et al., 2013, for a counterpoint). The fact that Guerra and Knoeferle (2014) found effects in first-pass reading times (considered a measure of early stages of processing) can be interpreted as a sign of the early and rapid integration of language processing and the visual context.

It's important to note that Guerra and Knoeferle (2014) investigated the effects of the visual context on the processing of concepts that have been retrieved from memory, such as the meaning of the words "war" and "peace." But how does the visual world interact with processing concepts that are not retrieved from one's mental lexicon, but are instead constructed on the fly, such as novel metaphors? We might find an answer to this question if we look at how the visual world interacts with the production of metaphoric expressions. Sato et al. (2015) investigated whether showing participants images depicting spatial containment would encourage them to produce expressions in which spatial containment is used metaphorically to speak of abstract difficulty. They found that even when the sentences they produced were thematically unrelated to the images viewed, participants still produced more metaphors drawing from the domain of spatial containment than when they saw a neutral picture as prime. The authors, who work within the framework of Conceptual Metaphor Theory (Lakoff and Johnson, 2008), interpreted the result as evidence for an activation of the Conceptual Metaphor DIFFICULTY IS CONTAINMENT after having seen the pictures, leading to the production of individual linguistic metaphors derived from this specific Conceptual Metaphor.

It's possible that these results could translate to language comprehension: activating the feature of spatial containment could facilitate comprehension of novel metaphors of difficulty that have spatial containment as part of their encoded meanings. This would suggest that literal features of a metaphor are important for the construction of metaphoric meaning and would be broadly in line with the indirect comparison view. With this in mind, we now turn to the description of our investigation, in which we explore the role of literal features during comprehension of novel verbal metaphors.

2. THE PRESENT STUDY

The current set of studies seeks answers to the following questions raised in section 2: Does activating literal features hinder or ease processing of novel metaphors? And do said features remain active after a metaphor has been understood? We conducted four experiments to answer these questions. In Experiments 1 and 2 (eye-tracking during reading), participants saw short animated clips depicting physical containment. They then read sentences in which verbs of physical containment were metaphorically used to signify difficulty [such as in sentence (1)], and then answered questions about either the sentences or the videos. The animated clips showed a moving ball: In one video, the ball bounces freely while in the other the ball is trapped by a box.

The goal of these two experiments was to study how seeing a video depicting physical containment—which we assume to be a prominent feature of the encoded meaning of the verbs used in all our sentences, yet incompatible with the meaning of the individual metaphors—interacts with the processing of verbs of spatial containment used metaphorically. We compared this to how the same sentences are processed after seeing a video clip that does not share the conceptual feature of containment with the verbs. In these two experiments participants also answered questions about what they saw in the video after reading the sentence. This should provide insight on the role that literal features might play after a metaphor has been understood.

In Experiment 3 (self-paced reading), we examined how participants would naturally answer the same questions asked in experiments 1 and 2 (after sentence comprehension) when the video clips are followed by literal sentences instead of metaphors. Doing this gives us a baseline measure to interpret the results of the question-answering times of Experiments 1 and 2.

Finally, Experiment 4 (lexical-decision task) investigated how the same video clips of Experiments 1–3 interact with the processing of spatial containment verbs from Experiments 1 and 2 when these verbs are read in the absence of a context (i.e., when participants are expected to retrieve the literal meaning only).

3. EXPERIMENT 1

We began our investigation by asking the following question: Will watching video clips of spatial containment facilitate or hinder comprehension of metaphors made up by verbs of spatial containment? Additionally, how will the activation of spatial containment interact with processing the metaphorically used verbs after the metaphors have been understood? Experiment 1, an eye-tracking during reading study, was designed to answer these questions.

3.1. Participants

Forty-eight monolingual university students who were native speakers of German (ages 18–31, 30 female) were recruited and tested at the Humboldt-Universität zu Berlin. All participants were right handed and had normal or corrected-to-normal vision. They all gave their written informed consent and were payed 8 euros upon completing the experiment. This study was covered by the ethics vote granted to the psycholinguistics lab of the Humboldt-Universität zu Berlin by the German Linguistic Society (Deutsche Gesellschaft für Sprachwissenschaft, DGfS).

3.2. Materials and Design

We created 40 critical items consisting of German metaphorical sentences. All sentences had an identical syntactic structure, namely a main clause with an infinitive subject clause, as exemplified in (3). In the infinitive clause, a verb of physical containment, which always appeared in the same position, was used metaphorically to denote abstract difficulty. In the main clause, it was asserted that the situation described in the infinitive

clause was "difficult." All critical and filler sentences can be found in the **Supplementary Material**.

(3) *Es war für den Redakteur /schwierig* ADJ */, seine / Meinung* TARGET NOUN */ nach dem Regimewechsel / umgittert* VERB*/ zu sehen.*
It was for the journalist/ difficult ADJ */, his / opinion* TARGET NOUN */ after the change in regime/ fenced-in* VERB */ to see*

"It was difficult for the journalist to see his opinion be fenced-in after the change in regime."

3.2.1. Sentence Norming

Our goal when creating the materials was to use metaphors that were novel yet readily understandable. To make sure the metaphors could be understood, we conducted a norming study of the target sentences. A sample of 15 participants, who did not participate in the main study, were asked to rate 80 sentences on a scale of 1–7, with 1 being totally incomprehensible and 7 being totally comprehensible. The 80 sentences were made up of the critical 40 metaphoric sentences and 40 semantically incoherent filler sentences (e.g., *It was sad that Thomas drank the car so fast*). Order of presentation of the sentences was randomized. The goal of the norming task was to establish whether any of the critical metaphorical sentences would be rated as incomprehensible (meaning a rating of 3.5 or lower) and whether the metaphorical sentences were rated significantly higher than the semantically incoherent sentences.

3.2.2. Results of the Norming Task

Four of the forty critical sentences were rated lower than 3.5 on average and were dropped from the investigation. The remaining 36 sentences formed the base for all subsequent experiments.

To determine whether these 36 sentences were in fact understood, an ordered logistic regression model was fitted to the data (Gelman and Hill, 2006). The model was constructed to see whether our critical items and the semantically incoherent fillers could predict the 1–7 ratings. The results show that a change from level 0 (semantically incoherent) to level 1 (critical item) was associated with an increase of odds ratio of 7.96 ($t = 17.5$, $p < 0.001$) This means that for metaphorical sentences, the odds of being rated higher were 7.96 times those of incoherent sentences, holding constant all other variables. The data therefore strongly suggests that participants were able to determine a difference in meaning between the semantically incoherent sentences and the novel metaphoric sentences.

Finally, to confirm that the resulting 36 sentences were in fact perceived as novel, we asked a further 50 participants (who did not take part in the main experiment) to rate how familiar they thought the metaphoric sentences were on a scale from 1 (very novel) to 100 (very familiar). The mean familiarity score was 27.98 with a standard deviation of 10.2. We take this as confirmation that the metaphors created were indeed perceived to be fairly novel.

3.2.3. Filler Sentences

Seventy-two filler sentences were constructed to reduce the likelihood of strategic behavior and to mask the purpose of our investigation. We thus had 24 German idioms as fillers with similar syntactic structure to our critical items, as well as 24 novel metaphors different to the critical items. The remaining 24 filler sentences were literal statements.

3.2.4. Visual Primes

Two critical videos were created by animating individually created images with proprietary video editing software. Each video showed a ball bouncing with identical motion: In one of them (used in the "match" conditions) the ball was seen to be captured by a moving box, forcing the ball to a still stand. In the other (used in the "mismatch" conditions), the ball bounces freely and stops on its own. **Figure 1** shows a series of stills for each of the videos. The videos themselves can be seen in full length in the **Supplementary Material**.

Furthermore, inspired by Experiment 1 of Guerra and Knoeferle (2014), two versions of each video were created: One with a printed word from each critical sentence on the ball and one without any printed word. Participants thus saw, for example, a video of a box trapping a ball (or a ball bouncing freely) that had the word *opinion* written on it, and subsequently read sentence (3), in which an "opinion" is said to be *fenced in*. This was done to maximize the possibility that participants would establish a relation between the visual context and the written sentence.

For the filler trials, four other animated videos were created that were randomly paired with the 72 filler sentences. To prevent participants from identifying the critical videos, the filler videos presented the same objects as the critical ones, i.e., a combination of bouncing balls and boxes. In the filler videos a box lands next to a bouncing ball without trapping it (filler video 1); two balls cross each other diagonally and bounce toward each other (filler video 2) or away from each other (filler video 3); and two balls fall on top of a box but only one of the balls goes in the box (filler video 4).

3.2.5. Comprehension Questions

To investigate the role of literal features after a metaphor has been comprehended, we included a comprehension question after every trial. For critical trials, the question was always about the video, either (a) referring to the ball (*Was the ball in the box?*) or (b) to the metaphoric topic that may or may not have appeared written on the ball in the video (*Was the opinion in the box?*). Trials with incorrect answers were discarded from the analysis.

The idea of having these two different questions was that they might allow us to investigate different ways in which literal features could be activated after metaphor comprehension: It could be the case that literal features are simply activated because they are seen in the video and mentioned in the sentence, in which case question (a) should be easier to respond to when the video-prime seen prior to the metaphor activates the literal feature of containment. This would be compatible with indirect comparison views and with Carston's (2010) "lingering" of the literal meaning view. Alternatively, literal features could remain activated because they are part of a network of systematic mappings between topic and vehicle established during structural alignment, as suggested by Gentner and Boronat (1992), Gentner

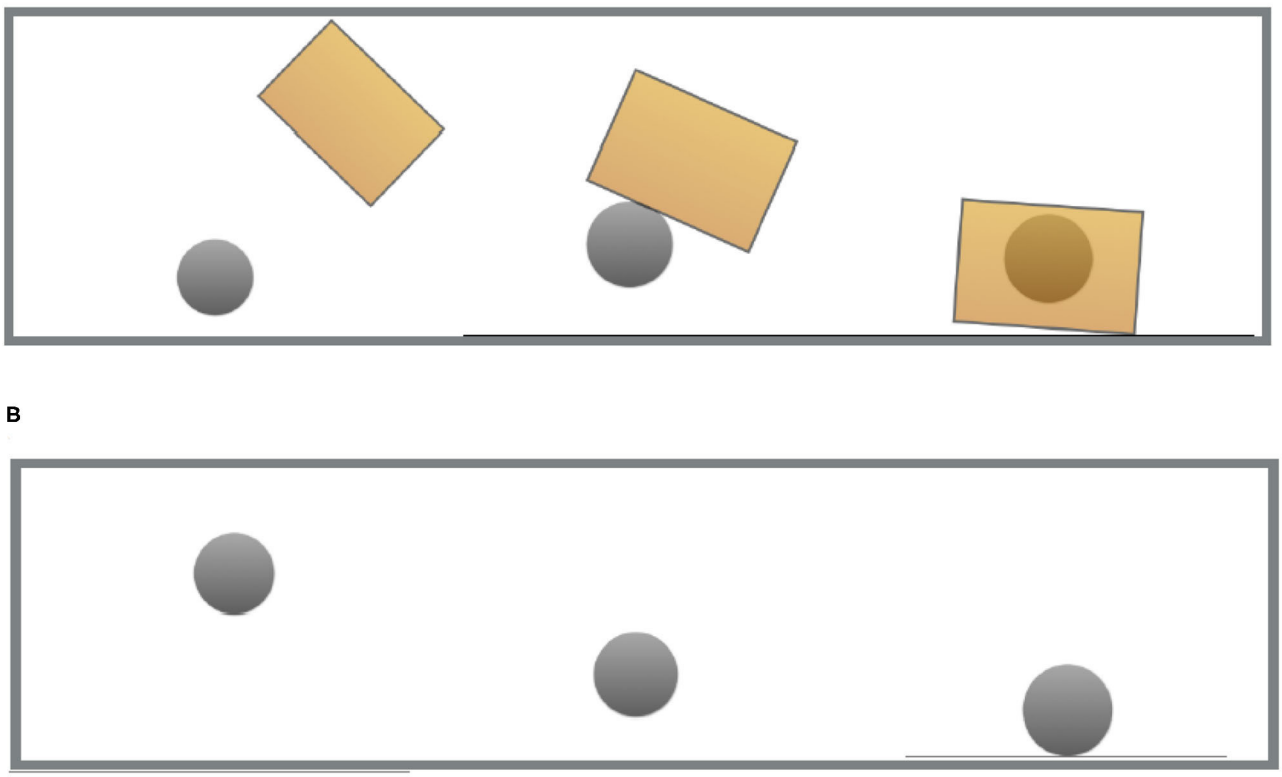

FIGURE 1 | Stills from the video used in the "match" and "mismatch" conditions of experiments 1–4. **(A)** Visual prime—Containment. **(B)** Visual prime—Non-containment.

et al. (2001), and Thibodeau and Durgin (2008). This would result in a facilitation effect when answering question (b), considering that it suggests a parallel in structure between video and sentence by effectively "blending" together both representations. Finally, it could be the case that literal features are always suppressed after metaphor comprehension, in which case neither type of question should be easier to answer when the video activates the literal feature of containment compared to when the video does not activate it. This would be compatible with the category inclusion view (Glucksberg, 2008). We return to these positions and how they relate to the experimental design when discussing the results of the question-response times.

3.3. Design

Experiment 1 had a 2 × 2 × 2 Latin square design with three factors: "containment" (match vs. mismatch), "question type" (video-question vs. noun-question), and "prime type" (animation-prime vs. mixed-prime). "Containment" refers to whether the video showed the ball bouncing freely (mismatch conditions) or being trapped by a box (match conditions) (see **Figure 1**). "Question type" refers to whether the comprehension question inquired about the video (video-question conditions) or about the metaphoric topic [*the opinion* in (3)] (noun-question conditions). Finally, "prime type" refers to whether the metaphoric topic was written on the ball (mixed-prime conditions) in the video prime or whether the video prime had no written language in it (animation-prime conditions).

We calculated three eye-tracking measures commonly associated with different temporal processing stages (see Rayner, 1998, 2009) for our three regions of interest (i.e., the adjective, the noun, and the verb region): First-pass reading times, defined as the duration of all fixations made in a region until the first time the region is abandoned either to a subsequent or to a prior word; regression path duration, defined as the duration of all fixations from the first fixation in a region up to (but excluding) the first fixation to the right of this region (but including the duration of all fixations made to the left of the critical region after the first fixation in the critical region); and total reading times, defined as the sum of the duration of all fixations in a critical region. These three measures were chosen since they can provide insight about the point in time in which effects might arise: If effects are found in first-pass reading times, it would suggest that they occur during the earliest stages of processing. If they are visible in regression path duration, it would likely point to it being related to the way in which a region is integrated into

the sentential context, whereas if they are found only in total reading times, it would suggest that such an effect might appear incrementally but only during later processing.

3.4. Predictions by Region

Our first set of predictions concerns the effect of the video on reading comprehension. We focused on three specific regions which we believed to be likely to interact with the visual prime: the adjective, noun, and verb regions.

3.4.1. Adjective Region

In Guerra and Knoeferle (2014) the authors found that visually depicted spatial distance facilitated reading comprehension of adjectives denoting abstract similarity. They reasoned that this facilitation effect might be due to an existing co-indexing link between spatial distance (close, far) and semantic distance (similar, dissimilar). They borrowed this idea from Conceptual Metaphor Theory, which hypothesizes the existence of such a link (Lakoff and Johnson, 2008). This theory also posits the existence of a link between the concepts of difficulty and containment. Thus, watching videos of spatial containment might ease processing of an adjective denoting difficulty. We therefore reasoned that if there is a link between difficulty and containment similarly to that found for the case of similarity and distance, we should find a main effect of containment in the adjective region, with shorter reading times in the match vs. mismatch conditions.

3.4.2. Noun Region

By adding the word in the noun region to the video (mixed prime type conditions), we expected a clear repetition priming effect to appear when participants encountered this word in the sentence. Concretely, if participants were able to integrate the written word from the video with the subsequently read sentence, we should observe a main effect of prime type in all dependent measures, with the mixed-prime conditions being overall faster to read than the animation-prime conditions.

3.4.3. Verb Region

Our predictions for this region are derived from the debate on metaphor processing presented in section 2. We expected a facilitation effect on an early measure, such as first-pass reading times, provided that the video relates to the literal meaning of the verb. This finding would suggest that features related to the literal meaning of a verb (in this case, physical containment) are initially active even though they might be absent from the intended metaphoric meaning. This would be in line with the results of Weiland et al. (2014), who observed that masked primes made up of irrelevant features of the metaphoric vehicle reduced the N400 effect found upon encountering the metaphoric vehicle, and would also generally support the indirect comparison view of metaphor understanding.

Alternatively, if activating the spatial representation of containment interferes with processing the metaphorically used verb, we should find longer reading times in the match vs. mismatch conditions. This would be more in line with the findings of McGlone and Manfredi (2001) and generally with category inclusion accounts that claim that literal features irrelevant for understanding the metaphor are actively suppressed during processing. Activating them should therefore interfere with the construction of metaphoric meaning.

3.5. Post-sentence Comprehension Question

A second set of predictions relates to how understanding each metaphor affects participants' response time patterns for questions related to the content of the video.

The main prediction for the response patterns to the post-comprehension questions was that if the feature of physical containment is active after participants have understood the sentence, it should be possible to find a main effect of containment on question-answering times, with overall shorter answering times in the match vs. mismatch conditions. This would suggest that the feature of containment activated in the match conditions (the ball is trapped by the box) was not suppressed after the metaphor was understood and facilitates answering both question (a) *Was the ball in the box?* and (b) *Was the opinion in the box?* If, on the other hand, the features activated by the video are suppressed after the metaphor has been understood, there should be either an interference or a null-effect of containment on response times.

However, given that there were two types of post-sentence comprehension questions, (a) and (b) above, it would be possible to observe different result patterns beyond the prediction of a main effect of Containment. Such patterns would bring about a more nuanced view on the activation of literal features after a metaphor has been comprehended, which could further inform theories of metaphor comprehension. **Table 1** presents a description of all conditions for the response times.

Of particular importance for a nuanced view on the role of literal features are the response times in the noun-question/animation-prime conditions. This is because, in these conditions, participants were asked a question that effectively "blended" the representations of video and sentence by asking whether the "opinion" was in the box when there was nothing written on the ball in the video but they had read about an opinion in the sentence.

If the feature of physical containment is activated after sentence comprehension, we would expect this feature to interfere with correctly answering the question in the noun-question/animation-prime conditions (because the correct response here would be NO and participants might want to answer YES if the feature of Containment is active), particularly in the match condition, where physical containment was seen in the video. This should in turn result in an interaction of question type and prime type, with the noun-question/animation-prime conditions showing longer reaction times than all other conditions. If the match level (of the noun-question/animation-prime conditions) is harder to respond to than the mismatch level, there should additionally be a three-way interaction between question type, prime type, and containment. If, on the other hand, the feature of physical containment is not active after

TABLE 1 | Description of all conditions for the question-response times in Experiments 1, 2, and 3.

Condition	Question asked	Description of video-prime	Correct response
Noun-question, animation-prime, match	Was the opinion in the box?	Ball is trapped in a box	No
Video-question, animation-prime, match	Was the ball in the box?	Ball is trapped in a box	Yes
Noun-question, mixed-prime, match	Was the opinion in the box?	Ball with word "opinion" on it is trapped in a box	Yes
Video-question, mixed-prime, match	Was the ball in the box?	Ball with word "opinion" on it is trapped in a box	Yes
Noun-question, animation-prime, mismatch	Was the opinion in the box?	Ball is bouncing freely	No
Video-question, animation-prime, mismatch	Was the ball in the box?	Ball is bouncing freely	No
Noun-question, mixed-prime, mismatch	Was the opinion in the box?	Ball with word "opinion" on it is bouncing freely	No
Video-question, mixed-prime, mismatch	Was the ball in the box?	Ball with word "opinion" on it is bouncing freely	No

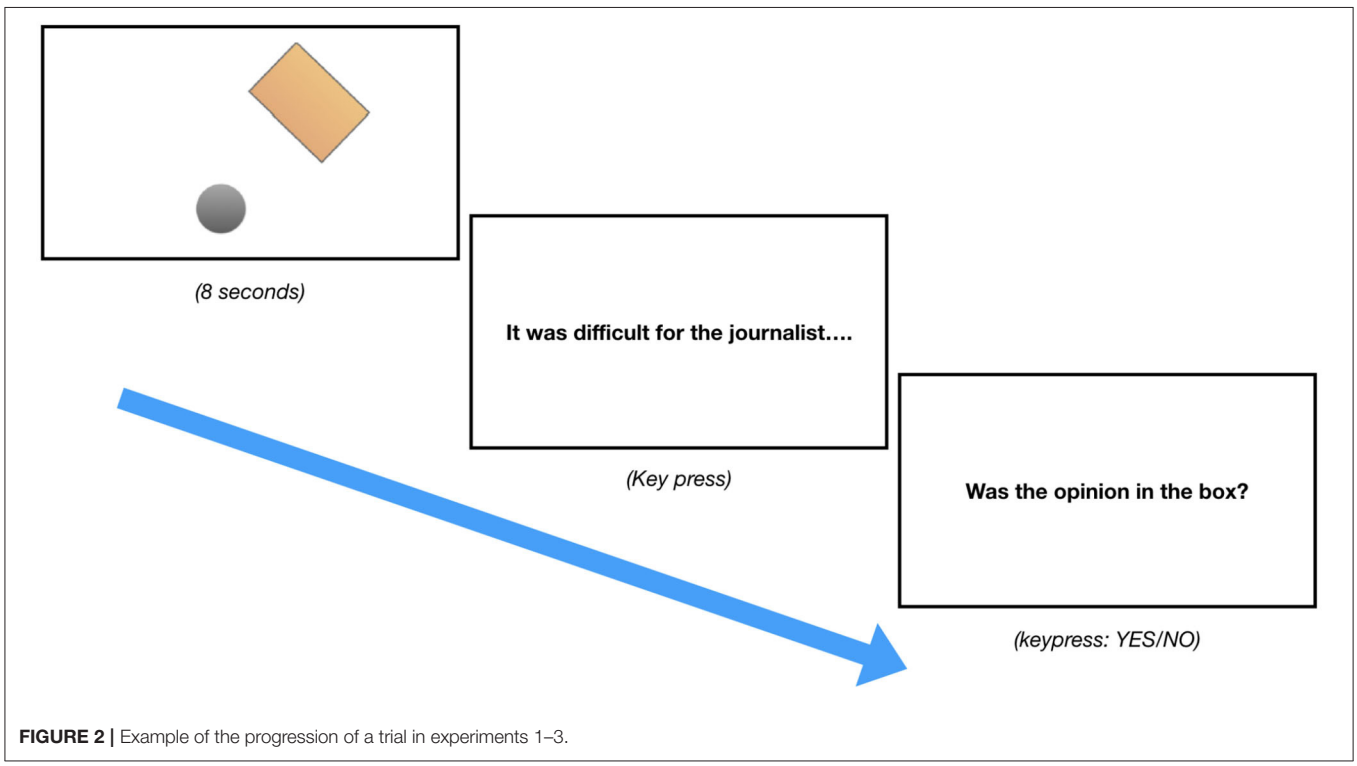

FIGURE 2 | Example of the progression of a trial in experiments 1–3.

participants have understood the sentence, we should expect the noun-question/animation-prime conditions to take just as long as the others, thus not resulting in a significant interaction of question type and prime type.

3.6. Procedure

Participants' eye movements were recorded using an Eyelink 1000 plus desktop head-stabilized tracker, produced by SR Research. At the beginning of each experimental session, the eye-tracker was calibrated with a 9-point calibration procedure to ensure accurate monitoring of the eyes. The procedure was performed and repeated until there was less than a maximum error of 0.5°. If it was not possible to meet this criterion, the experiment was aborted and participants were replaced. Re-calibration was performed after every block, i.e., twice more. After calibration, participants saw three practice trials before the experiment began. Each trial in the experiment consisted of three phases (see **Figure 2**): First, participants saw an animated video presented on the screen for 8 s. The video disappeared and a sentence appeared on the screen. Participants read the sentence and pressed a button on a Cedrus response pad that was in front of them when they had finished reading. The sentence then disappeared and a question appeared on the screen. Participants had to answer this question by pressing either the YES or NO button on the pad (position of YES and NO buttons was counterbalanced across participants). An entire experimental session lasted an average of around 50 min.

TABLE 2 | Random effects structure for models in every experiment.

	Models of eye-tracking data	Models of forced choice data
Experiment_1	(0 + prime type * containment ||item) + (0 + prime type * containment || subject)	(1 + type || subject) + (0 + type | item)
Experiment_2	(0 + prime type * containment || item)+ (0 + prime type * containment || subject)	(1 + type || subject) + (0 + type | item)
Experiment_3		(0 + question type * prime type * containment || item) + (0 + containment * question type * prime type ||subject)
Experiment_4		(1|subject) + (1|item)

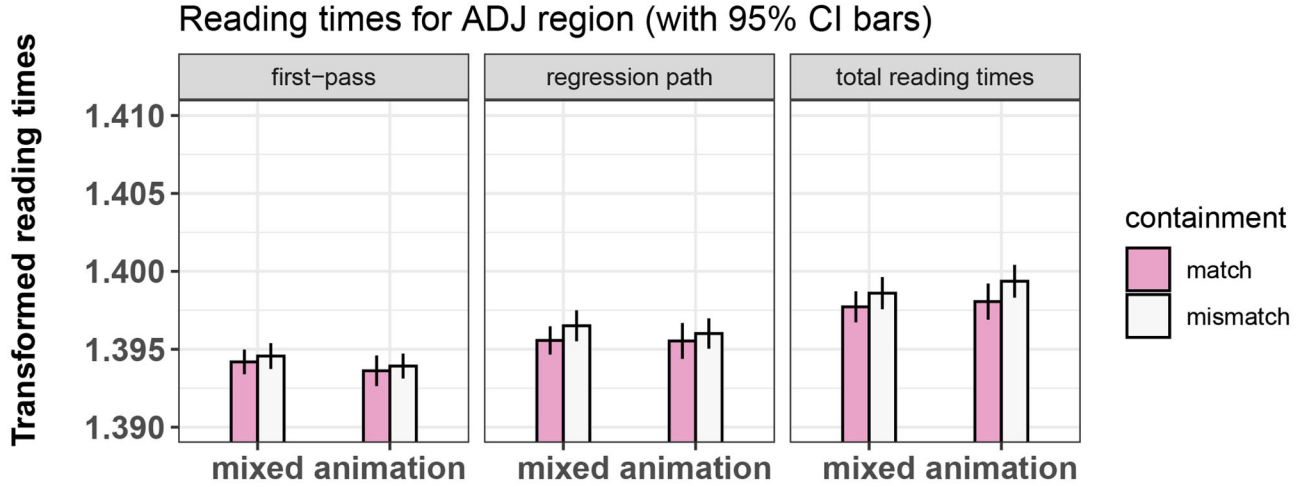

FIGURE 3 | Summary of results for the ADJ region, Experiment 1.

3.7. Analysis and Results

3.7.1. Analysis of Eye-Tracking Data

Prior to analysis, an intercepts-only regression model was fitted to the data in order to observe the distribution of the residuals. These were not normally distributed (which violates the assumptions of the linear model), and thus a box-cox test (Box and Cox, 1964) was performed. The test showed that the reading times measures needed to be transformed using a Lambda value of −0.7, which was used for transforming all eye-tracking measures and regions. Cases in which participants gave an incorrect answer to the comprehension question were also excluded from all analyses. This procedure was followed for all subsequent experiments. Accuracy for comprehension questions in experiment 1 was above 85% in all conditions.

We analyzed all data in our experiments using the R statistical programming environment and the LME4 package for regression analysis. To test our predictions, we fitted mixed-effects linear regression models to every measure and every region. For constructing the statistical models, we followed the recommendations of Barr et al. (2013). First, we tried fitting the largest possible random effects structure granted by our experimental design (in our case, random intercepts and slopes by items and subjects for both independent variables). If the model failed to converge, we reduced the random effects structure step-wise until a converging model was found by first removing the random correlations, then the random intercepts, followed by the interaction effects and the main effects. We used the same maximally converging random effects structure for all dependent measures in every region.

All models included trial order as a fixed effect, since it significantly improved the model fit. The models were fitted using a sum-contrast coding scheme (unless stated otherwise). Alpha thresholds for assessing statistical significance for eye-tracking data were Bonferroni-corrected, following the recommendations of von der Malsburg and Angele (2017).

The final random effects structure used for every model is shown in **Table 2**. **Figures 3–5** show bar-plots of the results in the adjective, noun, and verb region respectively. The output of the respective statistical models can be seen in **Tables 3–5**. **Figure 6** shows the results of the post-sentence comprehension question response times.

3.7.2. Results of Eye-Tracking, Adjective Region

No significant main effects or interactions were found in any measure for this region.

3.7.3. Results of Eye-Tracking, Noun Region

As predicted, we observed a significant main effect of prime-type in all three measures, with shorter reading times in the mixed-prime vs. animation-prime conditions. This confirms that our

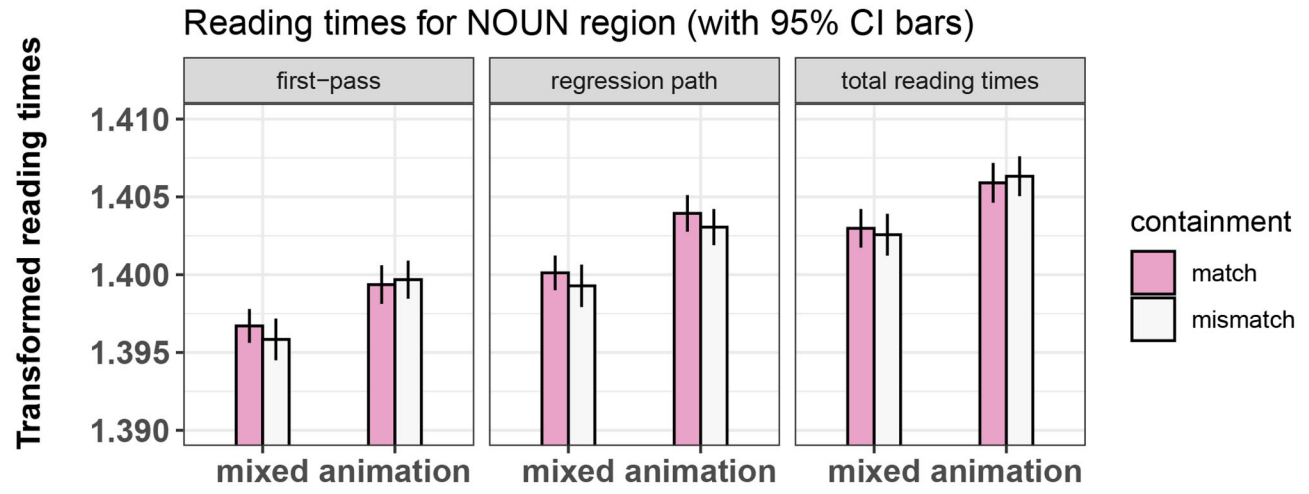

FIGURE 4 | Summary of results for the NOUN region, Experiment 1.

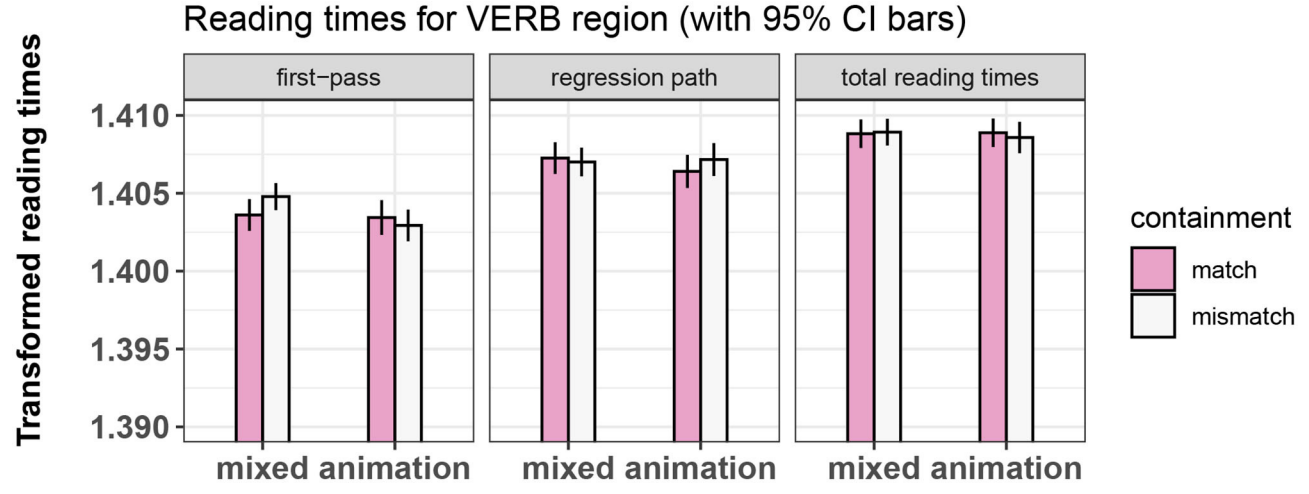

FIGURE 5 | Summary of results for the VERB region, Experiment 1.

experimental paradigm was sensitive enough to detect identity priming effects, and that participants were actively integrating the information processed during the video with the information from the sentence.

3.7.4. Results of Eye-Tracking, Verb Region

No significant main effects or interactions of our manipulated variables were found in any measure for this region.

3.7.5. Analysis and Results of Question Response Times

A box-cox test determined that the response times needed to be log-transformed. We thus fitted a linear mixed-effects regression model to the log-transformed reaction times. This model was fitted only to correct responses, which were over 92% of all trials. The results pattern can be seen in **Figure 6** and the output of the model is summarized in **Table 6**.

There was a main effect of question type, showing that participants were significantly slower at answering questions in the noun vs. video-question conditions. There was also a main effect of prime type, indicating that participants were faster to answer questions in the mixed-prime compared to the animation-prime condition, and a main effect of containment, showing that there was an overall facilitation in the match vs. mismatch conditions. There were also significant interactions between question type and prime type and containment and prime type, reflecting in particular that the noun-question/animation-prime conditions displayed a different pattern than all others (see **Figure 6**). The three-way interaction was not significant.

TABLE 3 | Regression analysis of reading times in the ADJECTIVE region of Experiment 1.

	Dependent variable		
	First-pass	**Regression path**	**Total reading times**
Prime Type	0.0003	0.0001	−0.0002
	t = 1.444	t = 0.581	t = −0.795
Containment	−0.0002	−0.0003	−0.0005
	t = −0.744	t = −1.389	t = −1.956
Trial order	−0.00000	−0.00001	−0.00003
	t = −0.293	t = −1.512	t = −3.105**
Containment/Prime Type interaction	−0.00002	−0.0001	0.0001
	t = −0.107	t = −0.494	t = 0.338
Constant	1.394	1.397	1.400
	t = 3,233.003***	t = 2,739.009***	t = 2,625.450***
Observations	1,180	1,180	1,180
Log Likelihood	4,162.329	3,965.666	3,910.930
Akaike Inf. Crit.	−8,300.659	−7,907.333	−7,797.859
Bayesian Inf. Crit.	−8,239.780	−7,846.453	−7,736.980

*$p < 0.017$; **$p < 0.0033$; ***$p < 0.00033$.
Significance thresholds are Bonferroni-corrected (alpha/3).

TABLE 4 | Regression analysis of reading times in the NOUN region of Experiment 1.

	Dependent variable		
	First-pass (1)	**Regression path (2)**	**Total reading times (3)**
Prime Type	−0.002	−0.002	−0.002
	t = −5.490***	t = −6.144***	t = −5.230***
Containment	0.0002	0.0004	0.00003
	t = 0.430	t = 1.221	t = 0.075
Trial order	−0.00000	−0.00001	−0.00004
	t = −0.072	t = −1.265	t = −3.687***
Containment/Prime Type interaction	0.0003	−0.00002	0.0002
	t = 0.800	t = −0.054	t = 0.507
Constant	1.398	1.402	1.407
	t = 2,294.932***	t = 2,344.835***	t = 2,192.578***
Observations	1,111	1,111	1,111
Log Likelihood	3,559.108	3,578.840	3,505.177
Akaike Inf. Crit.	−7,094.215	−7,133.679	−6,986.355
Bayesian Inf. Crit.	−7,034.059	−7,073.523	−6,926.198

*$p < 0.017$; **$p < 0.0033$; ***$p < 0.00033$.
Significance thresholds are Bonferroni-corrected (alpha/3).

A potential response bias was discovered after running the experiment: The correct answer to the question asked was always NO in the mismatch conditions and YES in the match conditions (see **Table 1**). It is therefore not possible to tell whether the effect of containment was caused by the difference in the conditions (match vs. mismatch) or by the differences in correct answer (YES vs. NO).

The noun-question/animation-prime was the only exception to this: Here, the correct response was NO in both match and mismatch levels. Because of this, we re-fitted the statistical model for the question-response times using a treatment contrast coding scheme in order to look at the noun-question/animation-prime condition exclusively. This was important because both match and mismatch levels of this condition were the only ones where both the question ("Was the NOUN in the box?") and the correct answer (NO) were the same. This type of contrast coding allows for direct comparisons between the condition set as the intercept of the model and the other individual conditions. This model showed no significant difference between match and mismatch levels

TABLE 5 | Regression analysis of reading times in the VERB region of Experiment 1.

	Dependent variable		
	First-pass (1)	Regression path (2)	Total reading times (3)
Prime Type	0.001	0.0002	0.0002
	t = 2.180	t = 0.760	t = 0.792
Containment	−0.0001	−0.0001	0.0001
	t = −0.496	t = −0.459	t = 0.611
Trial order	−0.00002	−0.00001	−0.0001
	t = −2.234	t = −1.383	t = −9.256***
Containment/Prime Type interaction	−0.0004	0.0002	−0.0002
	t = −1.683	t = 0.891	t = −0.736
Constant	1.405	1.408	1.413
	t = 2,726.656***	t = 2,704.316***	t = 3,084.624***
Observations	1,148	1,148	1,148
Log Likelihood	3,877.586	3,866.024	4,012.204
Akaike Inf. Crit.	−7,731.172	−7,708.049	−8,000.409
Bayesian Inf. Crit.	−7,670.623	−7,647.499	−7,939.859

*$p < 0.017$; **$p < 0.0033$; ***$p < 0.00033$.
Significance thresholds are Bonferroni-corrected (alpha/3).

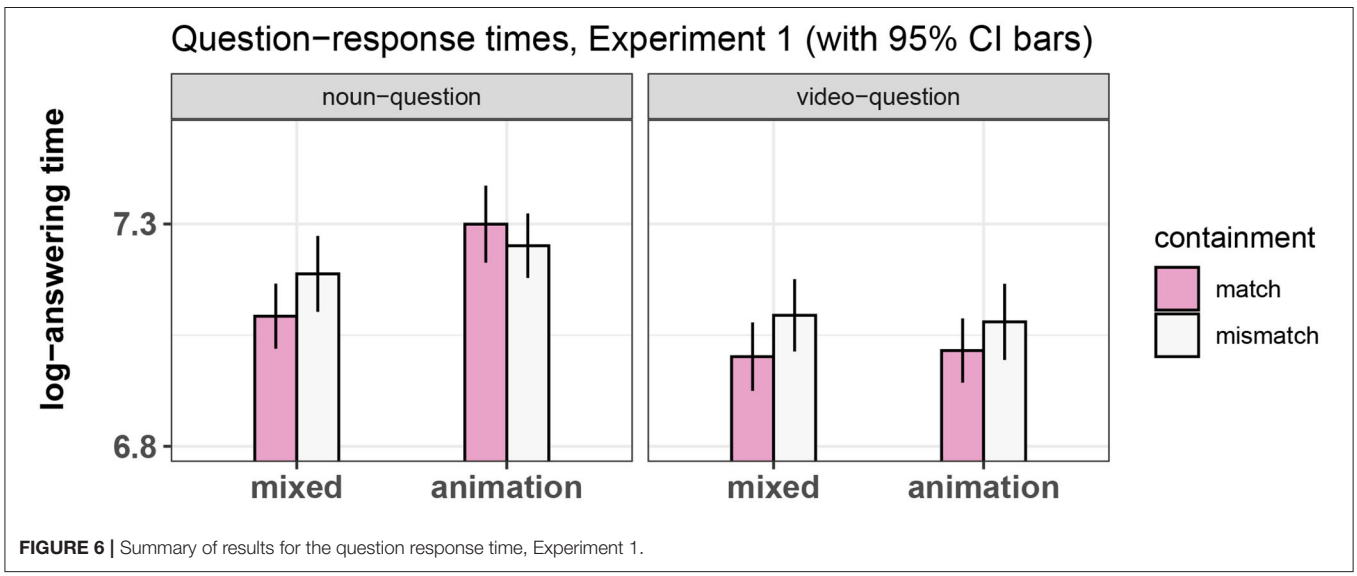

FIGURE 6 | Summary of results for the question response time, Experiment 1.

of the noun-question/animation-prime. This model is shown in **Table 12**.

3.8. Discussion

In Experiment 1, we failed to find a difference in reading times between conditions in the adjective region. More importantly, we found no differences in the verb region, the main interest region of the experiment. However, the presence of the effect of priming type in the NOUN region suggests that the absence of an effect of containment might be interpreted meaningfully: It could be the case that we did not find an effect of containment on reading times of the verb because the feature of containment is not relevant for the construction of the metaphoric meaning and it is thus ignored during processing, exerting neither facilitation nor interference. This interpretation would be broadly compatible with views that ascribe an insignificant role to features related exclusively to the encoded meaning of the metaphoric vehicle during processing.

However, it might also be possible that no effect was found given the temporal distance between presentation of the visual prime and reading of the metaphorically used verb. Perhaps this distance masked a true facilitation or interference effect that the video would have otherwise exerted on processing the verbs. This lays the groundwork for Experiment 2, in which we changed the sentence structure so that the verb could be temporally closer to the video prime.

TABLE 6 | Regression analysis of response-times in Experiment 1.

	Dependent variable
	Response times (in log-milliseconds)
Containment	−0.021
	t = −1.813
Prime Type	−0.027
	t = −2.325*
Question Type	0.081
	t = 6.754***
Trial order	−0.004
	t = −9.385***
Containment/Prime Type interaction	−0.027
	t = −2.267*
Containment/Question Type interaction	0.012
	t = 1.032
Question Type/Prime Type interaction	−0.038
	t = −3.269**
Three-way interaction	−0.016
	t = −1.345
Constant	7.327
	t = 215.365***
Observations	1,111
Log Likelihood	−559.837
Akaike Inf. Crit.	1,145.674
Bayesian Inf. Crit.	1,210.843

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

Results from the post-sentence comprehension questions present an intricate pattern. The results showed a main effect of question type, with longer response times in the noun-question conditions than in the video-question conditions. There was a main effect of containment, with shorter response times in the match compared to the mismatch conditions in all but the noun-question/animation-prime conditions (as evidenced by the interaction effect between containment and question type).

To better understand this pattern, it is useful to think about how the results might possibly be linked to the theoretical debate on the activation of literal features following metaphor comprehension. Indirect comparison views suggest that after a metaphor is understood, literal features remain active because they are part of the network of established mappings between topic [in this case, the target noun *opinion* in sentence (3)] and vehicle [the verb *fenced in* in (3)], which can be used to reason analogically about subsequent linguistic input (see for example Gentner et al., 2001). If this holds, it would accommodate a facilitation effect of match vs. mismatch levels in the video-question conditions, signifying a sustained activation of the feature "containment." It would also account for an interference effect of match vs. mismatch levels in the noun-question/animation-prime conditions, which could be explained as a sustained activation of established mappings between different conceptual domains which interferes with answering a question about an "opinion" being in the video. This is because question (a) is a reference to the video alone, requiring only information about the feature of containment in order to answer it. If the feature is active, this should result in a facilitation effect compared to when containment was not presented (i.e., the mismatch condition). Question (b), on the other hand, is a complex combination of information about the sentence (given the presence of the target noun) and the video (given the reference to the box, which could have only been seen in the video). In this case, an interference effect for answering question (b) in the match vs. mismatch conditions would suggest that not only the feature of containment has been activated (as would be the case in the video-question conditions), but also its relationship with the metaphoric topic (the target noun). This should cause difficulty when negatively answering a question about an "opinion" being in the box. Carston (2010) suggests that literal features might "linger" after a metaphor has been understood. However, her theory seems to suggest that they "linger" only as semantic features, not as part of a network of systematic associations between topic and vehicle. That being the case, it would explain a facilitation effect of match vs. mismatch video on the question-response times in the video-question/animation-prime condition, but there should not be an effect on the response times in the noun-question/animation-prime conditions.

At first glance then, the pattern of results found in Experiment 1 seems to be in line with the idea that when the conceptual feature of containment was activated by the verb, it generally facilitated responses, resulting in shorter response times in the contained vs. not-contained conditions in all but the noun-question/animation-prime conditions.

This could suggest that the feature of containment was activated after the metaphor was understood, but not as part of a complex mapping between containment and the metaphoric topic (which would have caused a difference in the noun-question/animation-prime conditions), compatible with Carston's (2010) view on the "lingering" of the literal meaning, but incompatible with the stronger view of Gentner et al. (2001), according to which the pattern of mappings should remain available for further processing and potentially cause interference with the answering of the question.

There is, however, a simpler explanation for the current pattern of results. As mentioned in the results section, the correct responses were confounded with the match and mismatch conditions, with match conditions always requiring a YES response and mismatch conditions a NO response in all but the noun-question/no-label conditions, were the correct response was NO in both levels of containment. It is therefore likely that it was simply easier for participants to answer YES than to answer NO, explaining the main effect of containment. Additionally, the effect of question type could be due to the fact that questions in the "noun" conditions (which varied according to the target noun in every trial, 33 characters on average) were on average longer than the questions in the "video" conditions (which were always the same, i.e., *Was the ball in the box?*, 30 characters in German). It is possible that participants just took longer to read the questions in the noun compared to the video conditions and thus took longer to answer the question.

The only comparison not affected by these two issues was that between match and mismatch levels of the noun-question/animation-prime condition. For these two levels, the question and correct response remained the same. We found no significant difference between these two conditions. It's important to note, however, that the YES/NO confound affected only the question response times and not the eye-tracking data. We address the issue of the interpretation of question-response times in Experiment 3, where we examine the response patterns to the same questions in the absence of metaphorical verbs. For now, we turn to Experiment 2, where we attempted to replicate the pattern of reading times displayed in **Figure 5** using sentences with a different syntactic structure.

4. EXPERIMENT 2

The goal of Experiment 2 was to determine the robustness of the results of Experiment 1. First, we altered the sentence structure in order to minimize the temporal distance between prime and verb. We did this because we thought it was likely that participants were not able to use the information extracted from the visual prime to facilitate processing of the metaphoric verb due to working memory constraints. This possibility finds some support in the literature on working memory, where it has been noted that people have a relatively low average number of sequentially presented meaningful units that they can remember (somewhere between 3 and 7, Miller, 1956; Chen and Cowan, 2005). We also increased the number of participants, from 48 to 64, to obtain higher statistical power. We did this following a power analysis via simulation using the R package SimR (Green and MacLeod, 2016). For the power analysis, we took the model of the total reading times for the verb region as starting point. The simulations suggested that with 64 participants we would have over 80% power to detect a main effect of containment, assuming a true effect size of containment of Cohen's $d = 0.15$, i.e., somewhat smaller than the rule of thumb for a "small" effect size (Sawilowsky, 2009). By doing this we aimed to either detect a small effect that we were not able to find in the previous experiment, or to replicate the pattern of results of Experiment 1 with more validity.

4.1. Participants

Sixty-four native speakers of German (ages 18–31, 39 female) with normal or corrected-to-normal vision were recruited and tested at the Humboldt-Universität zu Berlin. None of them had participated in Experiment 1. They gave their informed consent and received 8 euros as compensation upon finishing the experiment. Experiment 2 was covered by the ethics vote granted to the psycholinguistics lab of the Humboldt-Universität zu Berlin by the German Linguistic Society (Deutsche Gesellschaft für Sprachwissenschaft, DGfS).

4.2. Materials, Design, and Procedure

The materials, design, and procedure were identical to those in Experiment 1, except for the syntactic structure of the critical sentence, which now displayed a leftward movement of the subject clause. This allowed for the verb to appear as the fourth word in the sentence, making it temporally closer to the video prime. The structure of the sentences was as follows:

(4) *Dass seine / Meinung* TARGET NOUN */ umgittert* VERB*/ wurde nach dem Regimewechsel, war / schwierig* ADJ */ für den Redakteur.*
"That his / opinion TARGET NOUN */ fenced-in* VERB */ was after the change in regime, was /difficult* ADJ*/ for the journalist"*

"The fact that his opinion was fenced-in after the change in regime was difficult for the journalist."

4.3. Predictions

Our predictions were motivated by the results of Experiment 1: If the absence of an effect of containment on the verb region was due to the temporal distance between verb and video, moving the verb closer to the video should correct this. Specifically, if priming physical containment facilitates processing of verbs of spatial containment used metaphorically, we should find shorter reading times in the match vs. mismatch conditions in the VERB region.

With regards to the question-answering times: The overall facilitation effect of match vs. mismatch in Experiment 2 was confounded with the type of response ("YES" for matches and "NO" for mismatches) in all but one relevant comparison: The noun-question/animation-prime conditions. We did not find a significant difference between these two conditions. In Experiment 2 we hoped to replicate the question-answering pattern in general, and the results of the noun-question/animation-prime conditions in particular.

4.4. Results

4.4.1. Eye-Tracking

Results for all regions and measures are shown in **Figures 7–9**. The output of the statistical models can be seen in **Tables 7–9**.

4.4.1.1. Adjective

No significant effects of containment or of prime type were found in this region.

4.4.1.2. Noun

We replicated the main effect of prime type on all measures, with the mixed-prime conditions showing overall shorter reading times than the animation-prime conditions. This shows that our participants were in fact relating video to sentence, leading to a reliable priming effect.

4.4.1.3. Verb

We failed to find an effect of containment on any measure, as was the case in Experiment 1. There was also no effect of prime type and no significant interaction of containment and prime type.

4.4.2. Question-Response Times

Question-response times were analyzed in the same way as in Experiment 1. As can be seen in **Figure 10**, the results are very similar to those of Experiment 1. We replicated all previous findings with the exception of the main effect of containment: There was a main effect of question type and of prime type. There

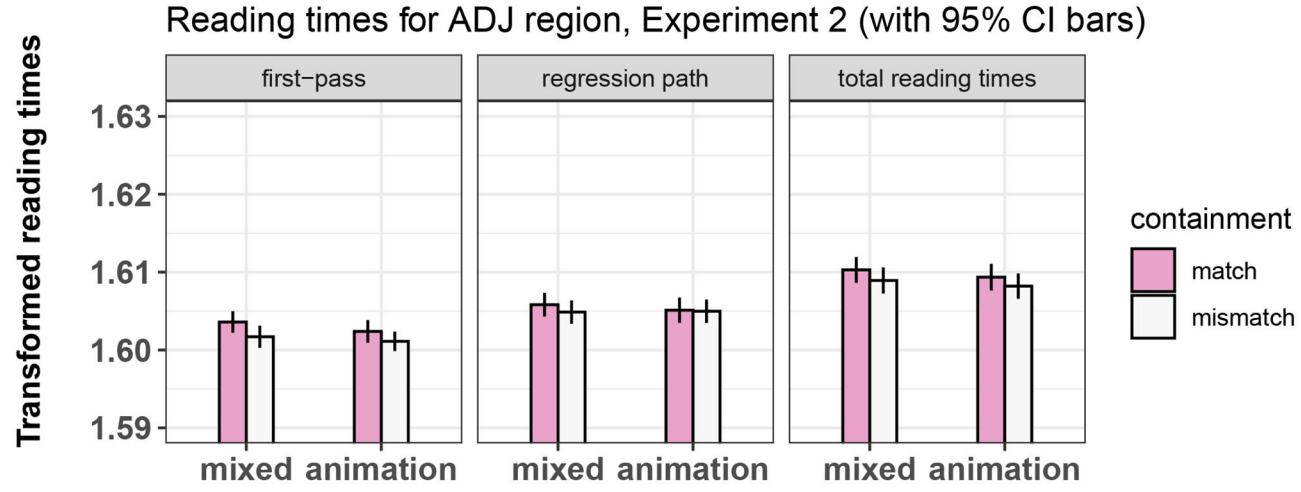

FIGURE 7 | Summary of results for the ADJ region, Experiment 2.

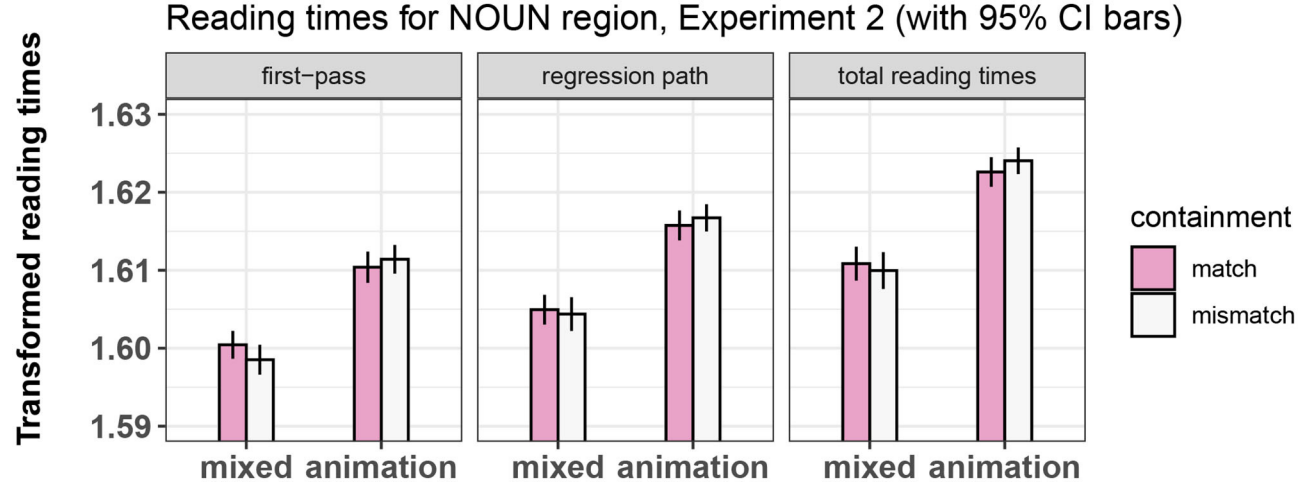

FIGURE 8 | Summary of results for the NOUN region, Experiment 2.

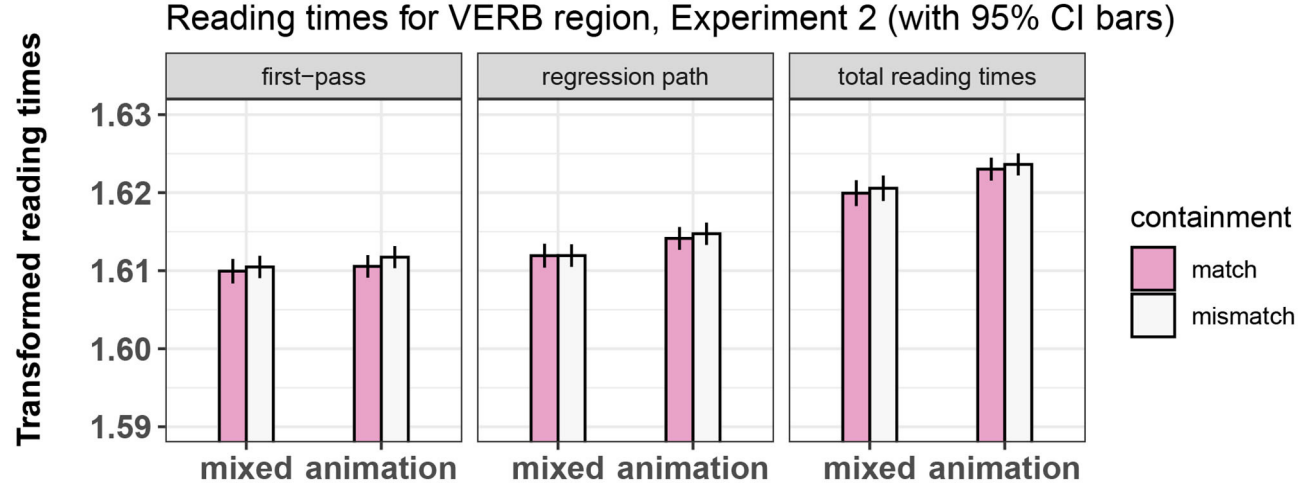

FIGURE 9 | Summary of results for the VERB region, Experiment 2.

TABLE 7 | Regression analysis of reading times in the ADJECTIVE region of Experiment 2.

	Dependent variable		
	First-pass	**Regression path**	**Total reading times**
	(1)	**(2)**	**(3)**
Prime Type	0.0005	0.0002	0.0005
	t = 1.356	t = 0.445	t = 1.029
Containment	0.001	0.0002	0.001
	t = 2.329	t = 0.660	t = 1.389
Trial order	−0.00001	−0.00002	−0.0001
	t = −0.619	t = −1.480	t = −3.907***
Containment/Prime Type interaction	0.0001	0.0002	0.00001
	t = 0.433	t = 0.479	t = 0.026
Constant	1.603	1.606	1.612
	t = 2,303.281***	t = 2,069.854***	t = 1,926.601***
Observations	1,634	1,634	1,634
Log Likelihood	4,709.178	4,530.882	4,406.515
Akaike Inf. Crit.	−9,394.356	−9,037.763	−8,789.029
Bayesian Inf. Crit.	−9,329.571	−8,972.978	−8,724.244

*$p < 0.017$; **$p < 0.0033$; ***$p < 0.00033$.
Significance thresholds are Bonferroni-corrected (alpha/3).

TABLE 8 | Regression analysis of reading times in the NOUN region of Experiment 2.

	Dependent variable		
	First-pass	**Regression path**	**Total reading times**
	(1)	**(2)**	**(3)**
Prime Type	−0.006	−0.006	−0.006
	t = −9.825***	t = −10.309***	t = −12.163***
Containment	0.0002	−0.0002	−0.0002
	t = 0.465	t = −0.334	t = −0.469
Trial order	−0.00002	−0.00005	−0.0001
	t = −1.618	t = −3.415**	t = −4.296***
Containment/Prime Type interaction	0.001	0.0002	0.0004
	t = 1.021	t = 0.294	t = 0.575
Constant	1.607	1.613	1.621
	t = 1,744.954***	t = 1,728.147***	t = 1,625.652***
Observations	1,491	1,491	1,491
Log Likelihood	3,908.996	3,884.676	3,796.362
Akaike Inf. Crit.	−7,793.993	−7,745.352	−7,568.724
Bayesian Inf. Crit.	−7,730.306	−7,681.666	−7,505.037

*$p < 0.017$; **$p < 0.0033$; ***$p < 0.00033$.
Significance thresholds are Bonferroni-corrected (alpha/3).

was an interaction between containment and question type and an interaction between question type and prime type. This model can be seen in **Table 10**.

As in the previous experiment, we re-fitted the model using a treatment-contrast scheme in order to directly compare match and mismatch levels of the noun-question/animation-prime condition. This model showed no significant difference between these conditions, replicating the result found in Experiment 1 (see **Table 12**).

4.5. Discussion

In Experiment 2 we tried to facilitate the interaction between video prime and metaphoric verb by increasing statistical power and decreasing the temporal distance between verb and video. We again failed to find an effect of containment in the verb region. Besides this, we replicated the effect of prime type on all measures in the noun region: Seeing the word *opinion* written on the ball in the video facilitated reading times of that same word once it appeared in the sentence. This confirms that participants

TABLE 9 | Regression analysis of reading times in the VERB region of Experiment 2.

	Dependent variable		
	First-pass	**Regression path**	**Total reading times**
	(1)	**(2)**	**(3)**
Prime Type	−0.0004	−0.001	−0.001
	t = −1.165	t = −3.061**	t = −3.779***
Containment	−0.0005	−0.0002	−0.0004
	t = −1.153	t = −0.550	t = −1.129
Trial order	−0.00004	−0.00005	−0.0001
	t = −3.886***	t = −4.105***	t = −8.117***
Containment/Prime Type interaction	0.0001	0.0001	−0.0002
	t = 0.270	t = 0.191	t = −0.370
Constant	1.613	1.616	1.627
	t = 2,206.844***	t = 2,209.841***	t = 2,151.451***
Observations	1,566	1,566	1,566
Log Likelihood	4,459.908	4,456.508	4,402.584
Akaike Inf. Crit.	−8,895.816	−8,889.016	−8,781.167
Bayesian Inf. Crit.	−8,831.541	−8,824.741	−8,716.892

*$p < 0.017$; **$p < 0.0033$; ***$p < 0.00033$.
Significance thresholds are Bonferroni-corrected (alpha/3).

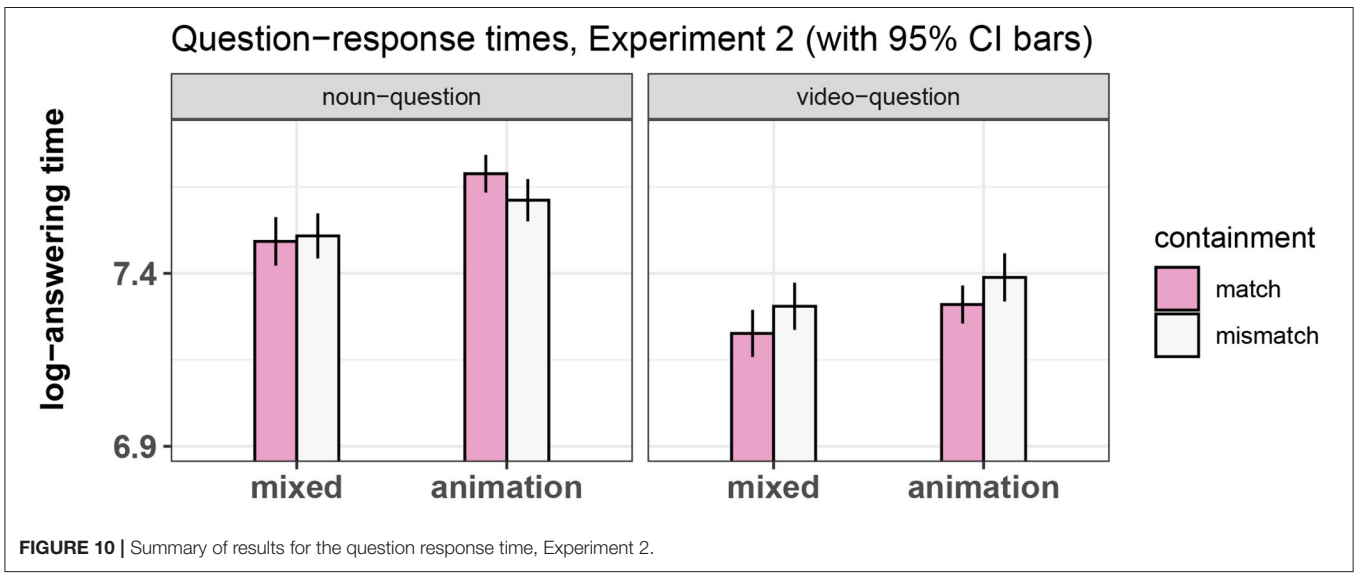

FIGURE 10 | Summary of results for the question response time, Experiment 2.

were able to use the information presented in the video to ease processing of the noun, and were nevertheless unable to use the feature of "containment" presented in the video to speed up (or slow-down) reading times in the verb region. This suggests that during processing of the metaphoric verb, participants largely ignored the feature of physical containment, seeing as it neither interfered with nor facilitated processing. This is consistent with a category inclusion view of metaphor comprehension that states that literal features are not initially activated if they are not necessary for the construction of the appropriate *ad hoc* category during metaphor processing.

However, it could also be the case that the lack of effects in the verb region is caused by inadequate materials: Activating the feature of spatial containment could indeed facilitate or hinder processing, but our video primes were simply not able to activate this feature. It is thus necessary to assess whether these videos could modulate processing in an environment in which they would be expected to do so reliably, namely when the verbs are processed in their encoded, literal meaning only. If the videos facilitate access to the literal meaning of the verbs, the current interpretation of the results of Experiments 1 and 2 becomes more plausible. We addressed this issue in Experiment 4.

The results of the question response task broadly replicated the findings of Experiment 1. It was easier for participants to answer the question in the match vs. mismatch levels of the video-question conditions. In the noun-question conditions, there was

TABLE 10 | Regression analysis of response-times in Experiment 2.

	Dependent variable
	Response times (in log-milliseconds)
Containment	−0.017
	t = −1.817
Prime Type	−0.060
	t = −6.551***
Question Type	0.134
	t = 14.741***
Trial order	−0.004
	t = −13.198***
Containment/Prime Type interaction	−0.018
	t = −1.963*
Containment/Question Type interaction	0.029
	t = 3.178**
Question Type/Prime Type interaction	−0.013
	t = −1.439
Three-way interaction	−0.007
	t = −0.722
Constant	7.663
	t = 246.117***
Observations	1,491
Log Likelihood	−613.253
Akaike Inf. Crit.	1,252.506
Bayesian Inf. Crit.	1,321.499

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

an effect of prime type, with the animation-prime conditions showing slower response times than the mixed-prime conditions.

The noun-question/animation-prime conditions did not show a significant difference between match and mismatch levels, just as in Experiment 1. This finding is important because the noun-question/animation-prime conditions were the only ones without a confound between condition and correct answer. Furthermore, there was an effect of prime type in the noun-question conditions, with the "animation" conditions showing longer response times than the "mixed" conditions.

As mentioned in the discussion of Experiment 1, these results could be interpreted as meaning that when reading the sentence, the conceptual feature of containment is activated, facilitating responses in the match vs. mismatch conditions and interfering with the responses in the noun-question/animation-prime conditions.

This interpretation, however, is contingent upon the assumption that the response patterns were caused by the interaction of processing video and metaphor and not by the YES/NO response confound or by other external factors. We sought to test this assumption in Experiment 3.

5. EXPERIMENT 3

Question-response times in Experiments 1 and 2 show an overall facilitation effect for match vs. mismatch conditions, except for the noun-question/animation-prime conditions, which showed no difference between match and mismatch levels. In Experiment 3, we set out to test whether these results were caused by the interaction of video, metaphor and question, or whether they could be explained by the interaction of video and question only. To do this, we ran a version of Experiment 2 in which the sentences read by participants did not contain any metaphors whatsoever: If the same pattern of results as in the previous two experiments is visible, it would suggest that the results are not related to the processing of verbal metaphors. Since we were not interested in the reading patterns of these sentences, but only in the question-response times, Experiment 3 was not run as an eye-tracking study. Instead, it was implemented as a self-paced reading reaction time task: Participants first watched the video-prime and then read the (non-metaphoric) sentence. When they were done reading, they pushed a button in front of them and were presented with the comprehension questions, which they answered by pushing either a YES or NO button. We measured only the response times to the comprehension questions.

5.1. Participants

Sixty-four native speakers of German (ages 18–31, 34 female) with normal or corrected-to-normal vision were recruited and tested at the Humboldt-Universität zu Berlin. None of them had participated in Experiments 1, or 2. They gave their informed consent and received 8 euros as compensation after completing the experiment. Experiment 3 was covered by the ethics vote granted to the psycholinguistics lab of the Humboldt-Universität zu Berlin by the German Linguistic Society (Deutsche Gesellschaft für Sprachwissenschaft, DGfS).

5.2. Materials and Design

To construct the materials in Experiment 3, we modified the sentences from Experiment 2 by replacing the verb with a non-metaphorical one that did not have the feature of spatial containment as part of its literal meaning, as presented in (5):

(5) *"Dass seine Meinung ignoriert wurde nach dem Regimewechsel, war für den Redakteur schwierig"*
"The fact that his opinion was ignored after the change in regime was difficult for the journalist"

The design was identical to that of the previous experiments, with the factors containment, question type and prime type. The experiment was programmed using the open source software Open Sesame and was run on a PC computer. The only dependent measure in this experiment was question response time.

5.3. Procedure

Participants were instructed to wear noise-reducing headphones throughout the experiment to avoid being distracted by the other participants. Each trial consisted of three phases: First, participants saw the same animated video presented in experiments 1–3. They then read a sentence and pressed the space bar on the keyboards that was in front of them. The sentence then disappeared and a question appeared on the screen. They had to answer this question by pressing either the letter F or J,

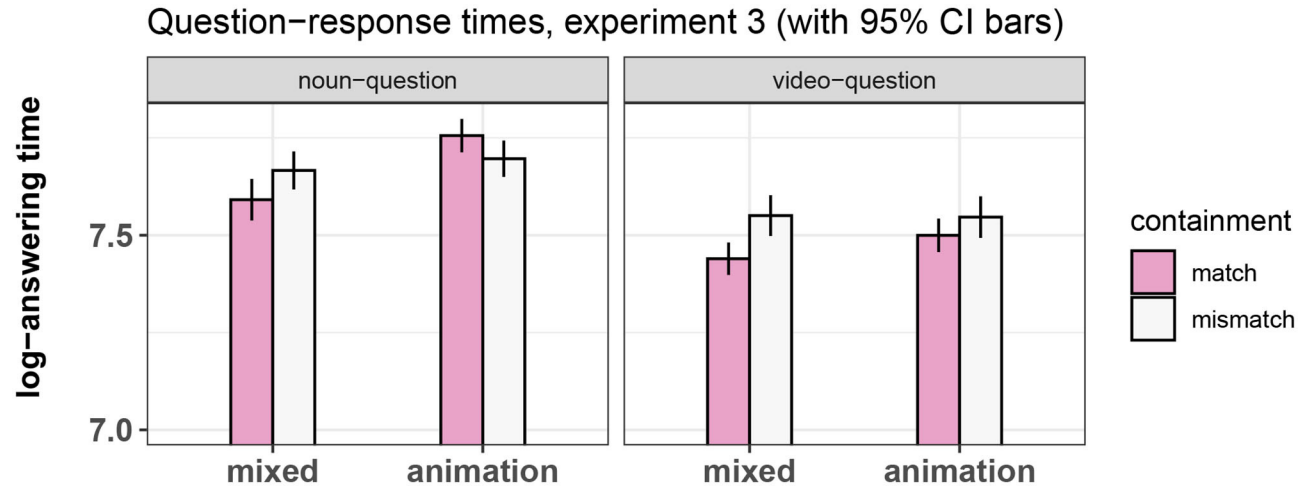

FIGURE 11 | Summary of results for the question response time, Experiment 3.

which were counterbalanced across participants to stand for either YES or NO.

5.4. Predictions

Our predictions are derived from the results of Experiments 1 and 2: If we find the same pattern of results in Experiment 3 as in the previous two iterations, it would suggest that the results were not driven by the interaction of video, metaphor and question, but just by the interaction of video and question, given that there are no metaphors in Experiment 3. If we find a different pattern than this, it would suggest that the results found in Experiments 1 and 2 were (at least partially) caused by the way participants processed the verbal metaphors. In this sense, Experiment 3 serves as a baseline against which we can interpret the results of the question-response times of Experiments 1 and 2. Of particular interest are again the noun-question/animation-prime conditions: These are the only match/mismatch pair where both the question asked and the correct response remained constant.

5.5. Results

We fitted a linear mixed effects regression model to the log-transformed reaction times. We found a main effect of containment, prime type and question type. We also found significant interactions of containment and question type, containment and prime type, question type and prime type and question type, prime type and containment. The results are shown in **Figure 11** and the model details are given in **Table 11**.

Re-fitting the model with treatment contrasts, as we did for the previous experiments, showed a significant difference between match and mismatch levels of the noun-question/animation-prime conditions, with the match condition showing significantly faster responses than the mismatch condition. The details of this model are shown in **Table 12**.

TABLE 11 | Regression analysis of response-times in Experiment 3.

	Dependent variable
	Response times (in log-milliseconds)
Containment	−0.043
	t = −2.281*
Prime Type	0.168
	t = 9.631***
Question Type	−0.062
	t = −3.881***
Trial order	−0.001
	t = −1.307
Containment/Prime Type interaction	0.071
	t = 2.219*
Containment/Question Type interaction	−0.099
	t = −2.936**
Question Type/Prime Type interaction	−0.067
	t = −2.105*
Three-way interaction	−0.069
	t = −1.088
Constant	7.612
	t = 466.819***
Observations	2,113
Log Likelihood	−887.236
Akaike Inf. Crit.	1,822.472
Bayesian Inf. Crit.	1,958.213

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

5.6. Discussion

The pattern of results is very similar to that found in Experiments 1 and 2. This suggests that the response times found in those experiments were mostly modulated by factors independent of the metaphorical verb, since there was no metaphorical verb in Experiment 3. This confirms the simple explanation that the response time results follow from a general response bias (Easier to answer YES than NO and easier to answer to

TABLE 12 | Model fitted with treatment-contrast coding for response times of Experiments 1–3.

	Dependent variable		
	Response times per Experiment (in log-milliseconds)		
	(1)	(2)	(3)
Containment	−0.066	−0.074	−0.064
	t = −1.436	t = −1.803	t = −2.185*
Constant	7.506	7.913	7.947
	t = 162.766***	t = 188.551***	t = 235.400***
Observations	1,111	1,491	2,113
Log Likelihood	−559.837	−602.232	−549.023
Akaike Inf. Crit.	1,145.674	1,256.464	1,150.046
Bayesian Inf. Crit.	1,210.843	1,394.452	1,297.098

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.
"Containment" shows effect in noun-question/animation-prime conditions only.

shorter than to longer questions), and are not a product of metaphoric interpretation.

However, the results of the noun-question/animation-prime conditions require further explanation. In Experiment 3, the match vs. mismatch conditions were significantly different from one another, whereas in Experiments 1 and 2, no significant difference was found. It is thus likely that this difference between experiments is the only one that is contingent on the presence of the metaphorical sentences in Experiments 1 and 2: If in the absence of a metaphor there are shorter response times in the mismatch compared to the match level of the noun-question/animation-prime condition (our baseline result), then the lack of a difference between conditions in the presence of a metaphor (Experiments 1 and 2) could actually be interpreted as a facilitation effect of the match compared to the mismatch condition relative to the baseline result of Experiment 3.

This interpretation, as well as the interpretation of the results of the gaze record of Experiments 1 and 2, relies on the assumption that participants can indeed derive the conceptual feature of containment from our prime videos and that this feature interacts with the way the verbs are processed. Experiment 4 directly addresses this issue.

6. EXPERIMENT 4

In this experiment we dealt with the question of whether or not the videos used in Experiments 1–3 can activate a mental representation of containment that leads participants to process verbs of physical containment more readily than when they first see a video that does not depict containment.

6.1. Participants

A sample of 259 German native speakers (ages 18–31, 120 female) were recruited online via the platform "clickworker." They gave their informed consent and received 50 cents as compensation upon finishing the experiment. Experiment 4 was covered by the data protection policy of the Humboldt-Universität zu Berlin.

6.2. Materials and Design

Experiment 4 was a web-based lexical decision task in which participants saw the same video clips from Experiments 1–3 as primes and then read the same verbs from Experiments 1 and 2, which were presented here without context. The experiment thus only had the factor containment with the levels match and mismatch.

6.3. Procedure

The experiment was designed and run using an instance of the IBEX farm (created by Alex Drummond) coupled with the Penncontroller extension (Zehr and Schwarz, 2018), which allows for a simple integration of video and linguistic stimuli. On each trial, participants first saw a video prime and then a target word in the middle of the screen, and had a total of 5 s to decide whether the word was a real word by either pressing F ("not a real word") or J ("real word"). After one practice item, participants were presented with six experimental trials (two critical, four fillers). There was a 1 s pause in-between trials. One experimental session lasted around 4–5 min.

6.4. Predictions

If the video in the "match" condition is not capable of eliciting a mental representation of "containment" that can aid lexical recognition of verbs of physical containment, there should be no difference in reaction times between conditions. If, on the other hand, the video in the "match" condition is indeed capable of eliciting a mental representation of "containment" that can ease lexical recognition of verbs of physical containment, we expect shorter reaction times in the match condition compared to the mismatch condition.

6.5. Analysis and Results

Prior to the analysis, participants who got <4/6 correct responses were excluded ($n = 9$), leaving the total number of participants at 250. Reaction times were log-transformed following the results of a box-cox test (Box and Cox, 1964).

A linear mixed effects model was then fitted to the data. The results showed a significant difference between the two conditions, with the match condition displaying shorter reaction times compared to the mismatch condition. The effect size had a value of Cohen's $d = 0.21$ (i.e., a "small" effect size according to Cohen, 1992). The results are presented in **Figure 12** and the model summary in **Table 13**.

6.6. Discussion

Experiment 4 showed that the video-clip primes used in Experiments 1–3 facilitated the retrieval of the encoded, literal meaning of different verbs of physical containment. This finding suggests that participants were able to derive the conceptual feature of physical containment from the videos in the match conditions, since this is the key feature we believe the videos share with the verbs.

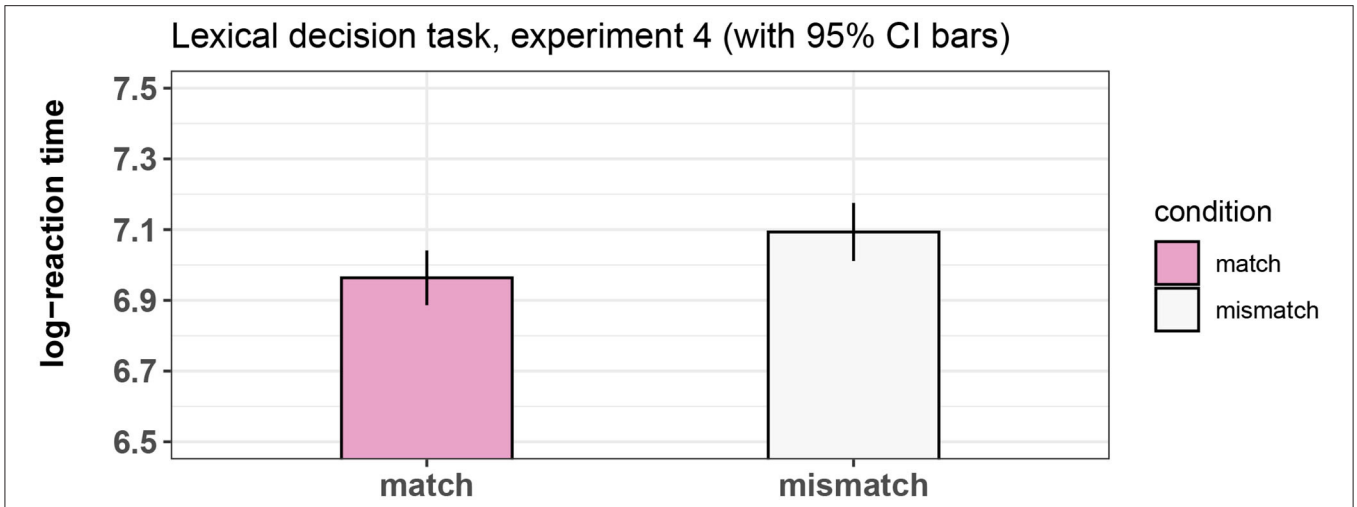

FIGURE 12 | Summary of results for the lexical decision task, Experiment 4.

TABLE 13 | Regression analysis of response-times in Experiment 4.

	Dependent variable
	Response times (in log-milliseconds)
Containment	0.133
	t = 2.993**
Constant	6.974
	t = 153.213***
Observations	465
Log Likelihood	−415.286
Akaike Inf. Crit.	840.571
Bayesian Inf. Crit.	861.282

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

7. GENERAL DISCUSSION AND CONCLUSION

Theories of metaphor processing make different predictions regarding the role of conceptual features related only to the literal meaning during and immediately after processing of (novel) metaphors. Category inclusion views believe that these literal features should not play a role during processing and might even hinder comprehension (McGlone and Manfredi, 2001). Furthermore, they should be rapidly suppressed after the metaphor has been understood (Gernsbacher et al., 2001; Rubio Fernandez, 2007). Indirect comparison views, instead, claim that features related to the literal meaning of a metaphor are initially active. This is caused by an alignment stage in which encoded meanings are fully retrieved prior to the projection of inferences (Gentner et al., 2001; Bowdle and Gentner, 2005, i.a.). This means that literal features should facilitate early stages of processing, as shown by Weiland et al. (2014), and can remain active after comprehension, easing understanding of subsequent, related novel, or conventional metaphors (Thibodeau and Durgin, 2008).

In our investigation, we looked at how priming the conceptual feature of spatial containment would interact with the processing of verbal metaphors in which physical containment is a crucial part of the literal meaning but (arguably) not of the metaphoric interpretation. The results of two eye-tracking experiments showed that the videos neither facilitated nor hindered processing of the verbs used (e.g., *fenced-in*), regardless of whether the verb appeared early on or late in the sentence (Experiments 1 and 2). This absence of an effect was accompanied by a reliable priming effect of the noun that appeared in both video and sentence, suggesting that participants were actively integrating the input of the video with the input of the sentence. Furthermore, we showed that the videos did elicit a priming effect on those same verbs in a de-contextualized lexical decision task (Experiment 4).

Data from the question-response times showed that participants were overall faster answering questions in the match vs. mismatch conditions. They were also overall slower to answer questions about the interaction between video and sentence (*Was the opinion in the box?*) than about just the video. Since these effects were present in both the experiments with a metaphoric verb (Experiments 1 and 2) and our baseline experiment without a metaphoric verb (Experiment 3) they do not tell us much about how the metaphors interacted with video and question type during processing. However, in the absence of a metaphor (Experiment 3), participants were significantly faster at correctly answering the question in the noun-question/animation-prime mismatch condition (*Was the opinion in the box?* When there was no word written on the ball and the ball bounced freely) compared to the noun-question/animation-prime match condition (*Was the opinion in the box?* When there was no word written on the ball and the ball was trapped by the box). In Experiments 1 and 2, there was no difference between these

conditions. This suggests that in the presence of a metaphor there could be a facilitation effect of the match compared to the mismatch noun-question/animation-prime conditions, which might mean that the metaphor itself activated the feature of spatial containment which later facilitated response times to the post-sentence questions. However, the evidence for this is very tenuous since the overall question-response pattern in all three experiments was similar.

We interpret the data as showing that the feature of physical containment is ignored during comprehension of novel verbal metaphors of containment and neither facilitates nor hinders processing. Failing to find a significant difference between conditions is not equivalent to finding that there is no difference between them. However, given the results of Experiment 4 and the fact that in Experiments 1 and 2 there was a significant effect of prime type (showing that some aspects of the prime were indeed integrated with the sentence), we believe that the absence of an effect of containment in Experiments 1 and 2 can be interpreted as meaningful.

We see this as being in line with a metaphor processing view that does not ascribe an important role to literal features of the metaphoric vehicle during initial stages of processing. Such is the case of category membership views (Glucksberg, 2001; Sperber and Wilson, 2008), which claim that the meaning of the vehicle is quickly modulated given the dimensions provided by the topic. In this process, features of the literal meaning that are not compatible with the dimensions provided by the topic do not need to be activated. However, pre-activating these features does not interfere with the lexical modulation of the metaphoric vehicle either.

It is important to note that the goal of the current set of studies was to investigate novel verbal metaphors only. Given that other factors, such as conventionality (Bowdle and Gentner, 2005), aptness (Jones and Estes, 2006), and familiarity (Thibodeau and Durgin, 2011) can modulate metaphor processing, it would be interesting to observe whether the current results would hold when examining metaphors that varied along those three dimensions. We leave this specific point for future research to examine.

Furthermore, it could be that metaphor processing varies according to syntactic class such that nominal metaphors are processed differently than verbal metaphors. This would mean that nominal metaphors could be understood via indirect comparison (following Gentner and Bowdle, 2008) and verbal metaphors via lexical modulation (as posited by category inclusion views). However, neuroimaging evidence suggests that the mechanisms for different types of metaphors might be the same. Cardillo et al. (2012) investigated processing of both nominal and verbal metaphors using functional magnetic resonance. Their results show that the neural processes associated with both of these types of metaphors do not differ significantly, suggesting that the underlying cognitive mechanisms are likely the same. We therefore believe that our results generalize beyond the case of verbal metaphors.

In terms of how our results relate to the literature on the interaction between language and the visual world we can draw the following conclusions: Guerra and Knoeferle (2014) found a facilitation effect of visual primes of distance on processing of semantic similarity. They argued that this was indicative of an abstract co-indexing link between distance and similarity. In Experiments 1 and 2 of the current investigation we failed to find such a link between videos of containment and adjectives of difficulty. It could be the case that these co-indexing links are constructed and stored in memory via repeated, conventional use: Perhaps speaking of semantic similarity in terms of distance is a more common occurrence than speaking of difficulty in terms of containment, leading to facilitation effects in the former but not in the latter case.

In a production study, Sato et al. (2015) found a priming effect of metaphors of difficulty after participants saw images of physical containment, an effect which we failed to find in the present language comprehension study. This difference in results could be explained by a difference in conventionality of the types of metaphors used: Sato et al. (2015) counted the production of spatial prepositions, such as *in* and *out* (e.g., *Bobbie fell in love working in the potato factory*) and of idiomatic expressions (*Nick said time is full of shit*) as instances of a containment-as-difficulty metaphor. These types of conventional, "fossilized" metaphoric expressions are likely to be processed differently than novel metaphors (Keysar et al., 2000; Bowdle and Gentner, 2005) making the results difficult to compare, given that the materials in our study were all novel verbal metaphors (It is not clear whether participants in the study by Sato and collaborators even produced any novel metaphors at all).

There are some caveats with our interpretation of the results: First, in Experiment 4 each participant saw only two critical items, whereas in Experiments 1 and 2 participants saw the full set of 36 items. It could therefore be the case that repeated exposure to the video primes interfered with an underlying true priming effect that our experimental set-up in Experiments 1 and 2 could not detect. To assess this possibility we conducted *post-hoc* analyses examining the pattern of results of Experiments 1 and 2 in the first third of the Experiment (i.e., after 36 trials). These showed the exact same pattern found for the entire experiment (i.e., no effect of video-prime on reading measures). It is thus not likely that a repetition effect is solely responsible for the differences in effect found between Experiments 1, 2, and 4.

It is also possible that the lack of an effect was due to the verbs being embedded in a sentence, regardless of whether the context encourages a literal or metaphoric interpretation of the verb. This is unlikely, considering that in Experiment 2 the Video-Prime and the verb were almost as temporally adjacent as in Experiment 4, but it cannot be ruled out completely. Further research is necessary in order to determine the exact nature of the prime-verb relation and the different contexts under which a priming effect could arise. We nevertheless see our set of experiments as a step forward in understanding how metaphors are processed outside of the narrow realm of nominal metaphors.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ethics committee of the Humboldt-Universität zu Berlin. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

CR, EG, and PK planned and designed the experiments, edited the paper, and produced the final version. CR conducted the experiments and wrote the first draft of the paper. CR and EG analyzed the data. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by ANID/ PIA/ Basal Funds for Centers of Excellence FB0003 (EG) and FONDECYT individual grants 11171074 (EG) by ANID (National Agency for Research and Development, Government of Chile) are gratefully acknowledged.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.556624/full#supplementary-material

Video 1 | Example prime video for a critical item in the containment-mismatch conditions used in Experiments 1–4 (no text was written on the ball in Experiments 1 and 4).

Video 2 | Example prime video for a critical item in the containment-match conditions used in Experiments 1–4 (no text was written on the ball in Experiments 1 and 4).

REFERENCES

Bambini, V., Canal, P., Resta, D., and Grimaldi, M. (2019). Time course and neurophysiological underpinnings of metaphor in literary context. *Discour. Process.* 56, 77–97. doi: 10.1080/0163853X.2017.1401876

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

Barsalou, L. W. (1983). *Ad hoc* categories. *Mem. Cogn.* 11, 211–227. doi: 10.3758/BF03196968

Blasko, D. G., and Connine, C. M. (1993). Effects of familiarity and aptness on metaphor processing. *J. Exp. Psychol. Learn. Mem. Cogn.* 19, 295–308. doi: 10.1037/0278-7393.19.2.295

Bowdle, B., and Gentner, D. (2005). The career of metaphor. *Psychol. Rev.* 112, 193–216. doi: 10.1037/0033-295X.112.1.193

Box, G. E. P., and Cox, D. R. (1964). An analysis of transformations. *J. R. Stat. Soc. B Methodol.* 26, 211–252.

Cardillo, E. R., Watson, C. E., Schmidt, G. L., Kranjec, A., and Chatterjee, A. (2012). From novel to familiar: tuning the brain for metaphors. *Neuroimage* 59, 3212–3221. doi: 10.1016/j.neuroimage.2011.11.079

Carston, R. (2010). XIII-metaphor: *ad hoc* concepts, literal meaning and mental images. *Proc. Aristot. Soc.* 110, 295–321. doi: 10.1111/j.1467-9264.2010.00288.x

Chen, Z., and Cowan, N. (2005). Chunk limits and length limits in immediate recall: a reconciliation. *J. Exp. Psychol. Learn. Mem. Cogn.* 31, 1235–1249. doi: 10.1037/0278-7393.31.6.1235

Clifton, C., Staub, A., and Rayner, K. (2007). "Eye movements in reading words and sentences," in *Eye Movements: A Window on Mind and Brain*, eds R. van Gompel, M. H. Fischer, W. S. Murray, and R. L. Hill (Oxford: Elsevier), 341–372. doi: 10.1016/B978-008044980-7/50017-3

Cohen, J. (1992). A power primer. *Psychol. Bull.* 112:155.

Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language: a new methodology for the real-time investigation of speech perception, memory, and language processing. *Cogn. Psychol.* 6, 84–107. doi: 10.1016/0010-0285(74)90005-X

Coulson, S., and Oakley, T. (2005). Blending and coded meaning: literal and figurative meaning in cognitive semantics. *J. Pragmat.* 37, 1510–1536. doi: 10.1016/j.pragma.2004.09.010

Gelman, A., and Hill, J. (2006). *Data Analysis Using Regression And Multilevel/Hierarchical Models*, Vol. 3. Cambridge: Cambridge University Press.

Gentner, D., and Boronat, C. (1992). "Metaphor as mapping," in *Workshop on Metaphor* (Tel Aviv).

Gentner, D., and Bowdle, B. (2008). "Metaphor as structure-mapping," in *The Cambridge Handbook of Metaphor and Thought*, ed R. Gibbs (New York, NY: Cambridge University Press), 109–128. doi: 10.1017/CBO9780511816802.008

Gentner, D., Bowdle, B., and Wolff, P. (2001). "Metaphor is like analogy," in *The Analogical Mind: Theory and Phenomena*, eds D. Gentner, K. J. Holyoak, and B. Kokinov (Cambridge, MA: MIT Press), 199–253.

Gentner, D., and Holyoak, K. J. (1997). Reasoning and learning by analogy: introduction. *Am. Psychol.* 52, 32–34. doi: 10.1037/0003-066X.52.1.32

Gernsbacher, M. A., Keysar, B., Robertson, R. R. W., and Werner, N. K. (2001). The role of suppression and enhancement in understanding metaphors. *J. Mem. Lang.* 45, 433–450. doi: 10.1006/jmla.2000.2782

Glucksberg, S. (2001). *Understanding Figurative Language*. Oxford: Oxford University Press.

Glucksberg, S. (2003). The psycholinguistics of metaphor. *Trends Cogn. Sci.* 7, 92–96. doi: 10.1016/S1364-6613(02)00040-2

Glucksberg, S. (2008). "How metaphors create categories–quickly," in *The Cambridge Handbook of Metaphor and Thought*, ed R. Gibbs (New York, NY: Cambridge University Press), 67–83. doi: 10.1017/CBO9780511816802.006

Green, P., and MacLeod, C. J. (2016). SIMR: an R package for power analysis of generalized linear mixed models by simulation. *Methods Ecol. Evol.* 7, 493–498. doi: 10.1111/2041-210X.12504

Guerra, E., and Knoeferle, P. (2014). Spatial distance effects on incremental semantic interpretation of abstract sentences: evidence from eye tracking. *Cognition* 133, 535–552. doi: 10.1016/j.cognition.2014.07.007

Guerra, E., and Knoeferle, P. (2017). Visually perceived spatial distance affects the interpretation of linguistically mediated social meaning during online language comprehension: an eye tracking reading study. *J. Mem. Lang.* 92, 43–56. doi: 10.1016/j.jml.2016.05.004

Holyoak, K. J., and Stamenković, D. (2018). Metaphor comprehension: a critical review of theories and evidence. *Psychol. Bull.* 144, 641–671. doi: 10.1037/bul0000145

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

Jones, L., and Estes, Z. (2006). Roosters, robins, and alarm clocks: aptness and conventionality in metaphor comprehension. *J. Mem. Lang.* 55, 18–32. doi: 10.1016/j.jml.2006.02.004

Keysar, B., Shen, Y., Glucksberg, S., and Horton, W. S. (2000). Conventional language: how metaphorical is it? *J. Mem. Lang.* 43, 576–593. doi: 10.1006/jmla.2000.2711

Knoeferle, P., and Guerra, E. (2016). Visually situated language comprehension. *Lang. Linguist. Compass* 10, 66–82. doi: 10.1111/lnc3.12177

Lakoff, G., and Johnson, M. (2008). *Metaphors We Live By*. Chicago, IL: University of Chicago Press.

McGlone, M. S., and Manfredi, D. A. (2001). TopicVehicle interaction in metaphor comprehension. *Mem. Cogn.* 29, 1209–1219. doi: 10.3758/BF03206390

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychol. Rev.* 63:81.

Pouscoulous, N., and Dulcinati, G. (2019). "Metaphor," in *Oxford Handbook of Experimental Semantics and Pragmatics*, eds C. Cummins and N. Katsos (Oxford: Oxford University Press), 298–315. doi: 10.1093/oxfordhb/9780198791768.013.19

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychol. Bull.* 124:372.

Rayner, K. (2009). The 35th Sir Frederick Bartlett lecture: eye movements and attention in reading, scene perception, and visual search. *Q. J. Exp. Psychol.* 62, 1457–1506. doi: 10.1080/17470210902816461

Rubio Fernandez, P. (2007). Suppression in metaphor interpretation: differences between meaning selection and meaning construction. *J. Semant.* 24, 345–371. doi: 10.1093/jos/ffm006

Rubio-Fernández, P., Cummins, C., and Tian, Y. (2016). Are single and extended metaphors processed differently? A test of two relevance-theoretic accounts. *J. Pragmat.* 94, 15–28. doi: 10.1016/j.pragma.2016.01.005

Sato, M., Schafer, A. J., and Bergen, B. K. (2015). Metaphor priming in sentence production: concrete pictures affect abstract language production. *Acta Psychol.* 156, 136–142. doi: 10.1016/j.actpsy.2014.09.010

Sawilowsky, S. S. (2009). New effect size rules of thumb. *J. Mod. Appl. Stat. Methods* 8:26. doi: 10.22237/jmasm/1257035100

Sperber, D., and Wilson, D. (2008). "A deflationary account of metaphors," in *The Cambridge Handbook of Metaphor and Thought*, ed R. Gibbs (New York, NY: Cambridge University Press), 84–105. doi: 10.1017/CBO9780511816802.007

Thibodeau, P., and Durgin, F. (2008). Productive figurative communication: conventional metaphors facilitate the comprehension of related novel metaphors. *J. Mem. Lang.* 58, 521–540. doi: 10.1016/j.jml.2007.05.001

Thibodeau, P., and Durgin, F. (2011). Metaphor aptness and conventionality: a processing fluency account. *Metaphor Symbol* 26, 206–226. doi: 10.1080/10926488.2011.583196

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

von der Malsburg, T., and Angele, B. (2017). False positives and other statistical errors in standard analyses of eye movements in reading. *J. Mem. Lang.* 94, 119–133. doi: 10.1016/j.jml.2016.10.003

Weiland, H., Bambini, V., and Schumacher, P. B. (2014). The role of literal meaning in figurative language comprehension: evidence from masked priming ERP. *Front. Hum. Neurosci.* 8:583. doi: 10.3389/fnhum.2014.00583

Wolff, P., and Gentner, D. (2000). Evidence for role-neutral initial processing of metaphors. *J. Exp. Psychol. Learn. Mem. Cogn.* 26, 529–541. doi: 10.1037//0278-7393.26.2.529

Wolff, P., and Gentner, D. (2011). Structure-mapping in metaphor comprehension. *Cogn. Sci.* 35, 1456–1488. doi: 10.1111/j.1551-6709.2011.01194.x

Zehr, J., and Schwarz, F. (2018). *PennController for Internet Based Experiments (IBEX)*. OSF.

Conflict of Interest: 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 © 2021 Ronderos, Guerra and Knoeferle. 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) and the copyright owner(s) 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: 03 March 2021
doi: 10.3389/fpsyg.2021.624628

Children's and Adults' Sensitivity to Gricean Maxims and to the Maximize Presupposition Principle

Francesca Panzeri and Francesca Foppolo*

Department of Psychology, University of Milano-Bicocca, Milan, Italy

Up to age 5, children are known to experience difficulties in the derivation of implicitly conveyed content, sticking to literally true, even if underinformative, interpretation of sentences. The computation of implicated meanings is connected to the (apparent or manifest) violation of Gricean conversational maxims. We present a study that tests unmotivated violations of the maxims of Quantity, Relevance, and Manner and of the Maximize Presupposition principle, with a Truth Value Judgment task with three options of response. We tested pre-schoolers and school-aged children, with adults as controls, to verify at which age these pragmatic rules are recognized and to see whether there is a difference among these tenets. We found an evolutionary trend and that, in all age groups, violations of the maxims of Quantity and of Relation are sanctioned to a higher degree compared to infringements of the Maim of Manner and of the Maximize Presupposition principle. We conjecture that this relates to the effects that the violation of a certain maxim or principle has on the goals of the exchange: listeners are less tolerant with statements that transmit inaccurate or incomplete information, while being more tolerant with those that still permit to understand what has happened.

Keywords: maxims of conversation, acquisition of pragmatics, experimental pragmatics, maximize presuppositions, maxim of manner, pragmatic tolerance

OPEN ACCESS

Edited by:
Pietro Perconti,
University of Messina, Italy

Reviewed by:
Likan Zhan,
Beijing Language and Culture University, China
Paul Edward Engelhardt,
University of East Anglia, United Kingdom

****Correspondence:***
Francesca Panzeri
francesca.panzeri@unimib.it

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

Received: *31 October 2020*
Accepted: *01 February 2021*
Published: *03 March 2021*

Citation:
Panzeri F and Foppolo F (2021) Children's and Adults' Sensitivity to Gricean Maxims and to the Maximize Presupposition Principle. Front. Psychol. 12:624628. doi: 10.3389/fpsyg.2021.624628

INTRODUCTION

In his seminal work, Grice (1975) proposed an account of how speakers can communicate more than what they literally say based on the assumption that rational interlocutors collaborate to reach a common objective and that the most efficient way to accomplish this goal is to follow the maxims of conversation. These maxims regulate both the content of what is said—that has to be true (maxim of Quality), enough informative (maxim of Quantity), and relevant (maxim of Relation)—and its form: statements are required to be clear, unambiguous, concise, and with the events reported orderly (maxim of Manner). In interpreting speakers' remarks, then, the hearers assume that these general principles are obeyed; faced with apparent or manifest violations of these maxims, they will integrate new assumptions to save the cooperativity of the speakers. These implicit propositions, the implicatures, integrate what a speaker has said and permit the recognition of speakers' communicative intent. For instance, a speaker saying "Leo drew a dog in his notebook" will be normally taken to implicate that the dog was the only thing Leo drew, since if he had drawn something else (also a cat, for instance), a cooperative speaker who complies with the maxim of Quantity should have mentioned it. Or, if one utters "Lawyers are sharks," a statement that is obviously false if taken at face value, interlocutors will reinterpret it as a metaphoric comment

under the assumption that the speaker is cooperative at a deeper level. In other words, within the pragmatic perspective that views the recognition of the speaker meaning as an inferential process that starts from what a speaker has said and integrates it with additional assumptions, the conversational maxims play a key role in the derivation of these implicitly transmitted propositions, the implicatures. The apparent or manifest violations of the maxims, as in the cases discussed above, trigger the derivation of additional premises to save the appropriateness of the remark.

Numerous studies found that children experience many difficulties in the correct detection of the speaker intended meaning: they tend to stick to a literal interpretation of what was said without integrating the implicitly conveyed assumptions. Children, for instance, struggle with instances of figurative language for a long period: the correct interpretation of metaphors and ironic comments is a late developing skill (see Winner, 1997, for a review). Moreover, children up to at least 5 years of age do not compute the generalized conversational implicatures that ensue from the assumption that the speaker is obeying the maxim of Quantity: when presented with a scenario in which, for instance, Leo ate all the five cookies that were on a plate, if they are asked to evaluate the appropriateness of a sentence such as "Leo ate *some of* the cookies," preschoolers accept it, whereas adults reject it. The rejection of that sentence is couched on the derivation of the Quantity implicature: the utterance of "Leo ate some of the cookies" implicates that stronger statements (such as "Leo ate all of the cookies") are not true because the maxim of Quantity dictates to utter the most informative true statement.

The fact that younger children tend not to derive implicatures has received a lot of attention, and different proposals have been put forth to account for this non-adult-like behavior. Within a Relevance theory perspective that assumes that the derivation of implicit content is guided by the balance between the costs required by the activation of the inferential process and the cognitive gains that permit the strengthening of what is said, children might not be able to optimize this balance yet: they would be satisfied with a literal, unenriched interpretation of the statement because they do not realize that the cost of deriving implicit additional assumptions would result in a strengthened, and thus more informative, interpretation (Pouscoulous et al., 2007). Other scholars suggest that children might encounter difficulties in the identification of the salient alternatives that are necessary to trigger the inferential mechanism (Skordos and Papafragou, 2016): in the Quantity implicatures discussed above, children might not be able to understand that the statement "Leo drew a dog" or "Leo ate some of the cookies" is not felicitous in a scenario in which Leo drew both a dog and a cat or in which Leo ate all of the cookies because they are not able to retrieve the relevant alternatives that the speaker should have uttered to adhere to the maxim of Quantity. Finally, children might simply be more tolerant than adults with respect to pragmatically inappropriate statements: presented with a statement that omits important details to describe a given situation (e.g., "Leo drew a dog" when in fact he drew both a dog and a cat), children might accept it, being satisfied with the semantic truth of the statement and not sanctioning its pragmatic inappropriateness. To test this hypothesis, Katsos and Bishop (2011) proposed a Truth Value Judgment (TVJ) task, in which participants are presented with a statement that is true but underinformative in a given scenario in two different versions. In the first version, participants had to evaluate a puppet's statements using two options of response; in line with previous findings, children accepted, and adults rejected, true but underinformative statements. In a second version, though, participants were asked to reward the puppet with three different sized strawberries: in this case, children did not differ from adults and preferred the middle-sized reward to judge true but pragmatically inappropriate (i.e., underinformative) statements. Katsos and Bishop (2011) thus concluded that children are indeed sensitive to violations of pragmatic appropriateness, but they do not sanction it as adults do.

The studies discussed above focused on children's failure to derive implicatures under the assumption that the speaker is complying with the maxim of Quantity and hypothesized that children might exhibit a non-adult-like behavior, accepting true but underinformative statements, either because they are simply more tolerant with respect to pragmatic inappropriateness or because they experience difficulties in the inferential process, in particular in the identification of the relevant, more informative, statements that a cooperative speaker should have chosen. A different but related question is whether this pragmatic immaturity is specific to the maxim of Quantity or whether it involves other Gricean conversational maxims.

This question has been tackled by Surian et al. (1996): they designed a Felicity Judgment (FJ) task, the Conversational Violations Test (CVT), in which children were presented with two puppets that answered to various questions, and the child's task was to identify who was giving silly answers: the target puppy was the one that was violating one of the maxims of conversation. In particular, besides testing violations of the Principle of Politeness (with one puppet providing rude remarks), the CVT tested violations of the maxim of Quality (with one puppet providing an impossible-to-be-true answer, e.g., "I live on the moon"), of the maxim of Relation (with a puppet providing completely irrelevant answers, such as "My trousers were blue," to the question "What did you do on holiday?"), and of the two submaxims of Quantity. The first submaxim (Quantity I) prescribes to provide as much information as is required; in the CVT, for instance, when asked "What did you receive for your birthday?" the silly puppet replied "A present," a clearly underinformative answer. The second submaxim of Quantity (Quantity II), on the other hand, invites speakers not to provide more information than what is required. Surian et al. (1996) exemplified violations of this submaxim, which they labeled "Avoid redundant information," with statements such as "[For breakfast, I had] a hard-boiled egg cooked in hot water in a sauce pan."

We actually believe that these items exemplify violations of the Be brief submaxim of Manner and not of Quantity II. Grice (1975) admitted that the second submaxim of Quantity was "disputable" and discussed, with many cautionary remarks, a case in which it gets violated: A asks B whether p is the case, and B "volunteers not only the information that, but information

to the effect that it is certain that p, and that the evidence for being the case that p is so-and-so and such-and-such" (Grice, 1975, p. 52). He then discusses a violation of Be brief: the utterance of "produced a series of sounds that corresponded closely with the score of 'Home sweet home,"' instead of the more concise "sang 'Home sweet home"' (Grice, 1975, p. 55). The items tested in the CVT constitute, in our view, unnecessary lengthy descriptions (a "rigmarole," in Grice's words) that do not add any other information compared to more concise expressions, and we therefore consider them as violations of Be brief and not of Quantity II.

The CVT has been used to test atypical populations (children with autism spectrum disorders and with specific language impairment, Surian et al., 1996; hearing-impaired children, Surian et al., 2010) and bilingual children (Siegal et al., 2009, 2010), with typically developing (TD) monolingual children serving solely as the control group. Indeed, the evolutionary trajectory of TD children in the mastering of Gricean maxims has not been attested yet, except for one study that used a revised CVT in Japanese-speaking children, aged 4–6 years (Okanda et al., 2015). In this study, moreover, Okanda et al. (2015) highlighted how the CVT did not test violations of the maxim of Manner, and they thus decided to test cases in which the question was "Which do you like, tea or milk?" and the puppets answered either "I like milk" or "Maybe tea or maybe milk," an answer that they consider a violation of the submaxim Avoid ambiguity.

Besides the maxims of conversation, another pragmatic principle has been proposed: the Maximize Presupposition principle (Heim, 1991). Analogously to what happens with conversational implicatures, also this principle requires the evaluation of semantically equivalent alternatives that differ in their pragmatic appropriateness. Following Heim's example, the statement "*A* (biological) father of the victim arrived at the scene" sounds anomalous compared to "*The* (biological) father of the victim arrived at the scene." The determiner phrases "a D" and "the D" differ in that the latter presupposes the existence of a unique object that satisfies the description D. Since every person has a unique biological father, this presupposition of (existence and) uniqueness is indeed satisfied. Maximize Presupposition accounts for the infelicity of using the determiner "a" instead of "the" by stating that it is more pragmatically appropriate to use the alternative that activates presuppositional requirements that are satisfied in the context.

Sauerland (2008) explicitly drew a parallelism between the application of this Maximize Presupposition principle and the derivation of (Quantity I) implicatures: in both cases, there are alternative statements that differ in "strength" because one element is more informative or activates presuppositions; if a speaker utters the weaker statement, an implicature or implicated presupposition can be drawn that the stronger statement does not hold. Or, conversely, the utterance of the weaker element when the stronger one could have been used is pragmatically inappropriate.

To test whether children are sensitive to the Maximize Presupposition principle, Yatsushiro (2008) tested children aged 6–9 years and adults by means of an FJ task: participants were presented, for instance, with a picture of a single girl playing soccer, and they were asked to indicate which of the statements "The girl is playing soccer" and "Every girl is playing soccer" best described the situation. Despite the fact that the second statement is semantically true (under its logical reading, it simply requires that all the individuals who are girls in a given context are playing soccer), the statement with the definite description *the girl* should be preferred, given that the uniqueness presupposition associated to *the* is satisfied. Yatsushiro (2008) found an evolutionary trend, with 6-year-olds performing worse (albeit showing an accuracy above 70%) than 7-year-olds. At 8 years of age, though, children were adult-like, with an accuracy above 90%. FJ tasks such as the CVT and the one employed by Yatsushiro (2008) tap the ability of children to recognize which one of the two presented statements violates one of the Gricean maxims or complies with the Maximize Presupposition principle. Nevertheless, choosing the correct answer in an FJ task does not necessarily indicate that the child is aware that an answer that violates the Maximize Presupposition principle or a maxim is pragmatically inadequate: when presented with two statements, one that violates and one that complies with pragmatic principles, the child could be simply identifying the (more) appropriate answer. Indeed, Foppolo et al. (2012) tested two groups of monolingual 5-year-old children who failed to derive scalar implicatures in a classical TVJT with the CVT adopted from Surian et al. (1996) and with an FJ task in which children had to compare an underinformative vs. an optimal alternative description of a situation. They found that, despite accepting underinformative statements violating the maxim of Quantity in the TVJ task, most of the children performed at ceiling in the CVT and in the FJ task, suggesting that these tasks might overestimate children's pragmatic competence.

To verify whether children appreciate Gricean maxims and the Maximize Presupposition principle, that is, whether they recognize when a statement is not appropriate in a given situation, we should resort to tasks such as the TVJ task, where the child is asked to evaluate the felicity of single statements against a scenario. Only if children correctly reject utterances that do not conform to Gricean maxims and that do not follow the Maximize Presupposition principle we can safely conclude that they are aware of these pragmatic principles.

Building on previous work, we designed a task that aims at testing children's sensitivity to pragmatic principles using a ternary TVJ task in which participants have to judge the appropriateness of a single statement in a given scenario, evaluating it with three options (bronze, silver, or gold medal). On the one hand, this should enhance children's performance compared to binary tasks, as already observed by Katsos and Bishop (2011); on the other, differently from FJ tasks, in this task, children have to evaluate one single statement at a time, without being provided with an alternative, i.e., a pragmatically appropriate description, rendering the task more apt to capture children's real competence. Our main goal was to verify at which age children realize that, if a speaker violates a conversational tenet (a Gricean maxim or the Maximize Presupposition principle) for no clear purpose, then the resulting utterance is infelicitous. Target items comprise violations of the Maximize Presupposition principle, of the maxim of Relation, of Quantity I, and of two submaxims of Manner, Be brief and Be orderly.

Control items constitute literally true and false statements: notice that false statements can also be seen as unmotivated violations of the maxim of Quality.

Since Grice himself stated that "the observance of some of these maxims is a matter of less urgency that is the observance of others" (Grice, 1975, p. 46) and that, in particular, a speaker who uses undue prolixity (violating the submaxim of Manner Be brief) is more cooperative than one who lies (violating Quality), we may expect differences in the rejection rates across the maxims. Another goal of the present study, then, is to compare the relative impact of maxims' violations and the consequent sanctioning of those violations across different maxims and age groups. In particular, since many studies employing a TVT task found that children up to 5 years of age tend to accept statements that violate Quantity I, not deriving scalar implicatures, we aimed at using the same task to verify whether unmotivated violations of other maxims are sanctioned at the same level.

MATERIALS AND METHODS

Participants

We tested a total of 163 Italian monolingual TD children, 45 (22 F) were preschoolers, with a mean age of 5 years and 2 months (age range: 3.7–6.2), and 118 (68 F) were school-aged children enrolled in the first 3 years of primary school (40 first graders; 30 second graders; 48 third graders), with a mean age of 7.5 (age range 6–9). A group of 36 adults (18 F, mean age 36 years) served as control.

Materials and Procedure

In the task, children were asked to evaluate the appropriateness of statements uttered by a boy, Bruno, who answers to the questions of a puppet, Elm. In the warm-up sessions, children were introduced to the two characters: they were told that Bruno is a boy who does many things and that Elm is very curious to know what happens, but he is blindfolded, and therefore he poses a lot of questions to Bruno. Children were warned that Bruno always answers Elm's questions, but sometimes Bruno's answers are wrong, or at least not completely adequate: in those cases, children should warn Elm, and tell him what has really happened. For each item, participants were first provided with a scenario about what Bruno did, then Elm comes in and poses a question to Bruno, and he answers with a fully informative, underinformative, or false statement about what he did. The child task was to judge the appropriateness of Bruno's statement relative to the given context. Following Katsos and Bishop (2011), they had to select one of three options: gold medal/smiley face for really appropriate answers, bronze medal/sad face for completely wrong ones, and silver medal/blank face for so-and-so answers (i.e., true but somehow misleading descriptions of the context). For example, they were presented with the scenario in **Figure 1** (in which Bruno drew a cat and a dog), and they were told, "Here is what Bruno has drawn," then Elm comes in and poses his question to Bruno, "What have you drawn?" Bruno answers, "I drew a dog," thus violating the maxim of Quantity. Finally, the child is asked to judge Bruno's statement by selecting the bronze, silver, or golden medal.

The task comprises 24 items: 12 critical items and 12 control sentences. The critical items constitute unmotivated violations of the maxim of Quantity (four items), Relation (two items), and Manner (four items) and of the Maximize Presupposition principle (two items). As for the Quantity maxim, the critical statements are underinformative with respect to the given context, thus violating Quantity I ("Make your contribution as informative as is required"), either because Bruno mentions only one conjunct instead of two (as in **Figure 1**) or because he mentions the superordinate term instead of the basic one (e.g., instead of answering "I've eaten chicken," he says "I've eaten food"). The two items that involve unmotivated violations of the maxim of Relation constitute irrelevant answers to a question: for instance, participants are shown Bruno's favorite shirt (red, with an image of a monkey) that Bruno describes as a shirt "that has two sleeves and a hole for the head." Infringements of the maxim of Manner are tested with two items that violate the submaxim Be orderly (e.g., "I went to the bed and I brushed my teeth") and two items that violate the submaxim Be brief (e.g., Bruno said that his snack was "A fruit with yellow peel and that monkeys really like," instead of simply saying "banana"). As discussed in the *Introduction*, these items are lengthy descriptions that do not add any information to more concise terms, and we thus consider them as violations of Be brief, whereas in the CVT, they were viewed as involving Quantity II.

Notwithstanding, since it is admittedly disputable what maxim is involved, in the *Discussion* section, we take into account also the hypothesis that the second submaxim of Quantity is implicated.

Finally, two critical items were violations of the Maximize Presupposition principle, in which the indefinite determiner *a* was used instead of the stronger presuppositional trigger *the* (e.g., "A sun is setting"). The control statements were eight clearly true statements and four clearly false. Notice that the false statements can also be viewed as unmotivated violations

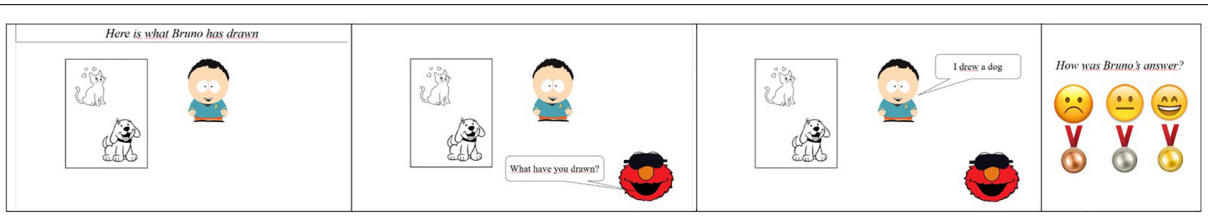

FIGURE 1 | Exemplification of a critical trial involving the violation of the maxim of Quantity.

of the maxim of Quality, in particular, of the first submaxim "Do not say what you believe to be false." Since the critical items that constitute violations of Be brief were long statements, some of the control items were lengthy descriptions, some of which were true (for instance, "I like to have tea with lemon and sugar") and some false ("On the desk, there is a sheet of paper, colored pencils, and a book"—when there were watercolors instead of a book).

In an initial warm-up session, participants were familiarized with the task. This session comprised a clearly true and a clearly false statement (to be rewarded with gold and bronze medal, respectively) and an instance of a so-and-so type of answer: participants were first shown an image of Bruno's favorite pizza (with sausages and French fries), then Elm asked Bruno how he liked to eat pizza and he answered that he likes the pizza on a plate.

The task was implemented on Microsoft PowerPoint, with all the statements prerecorded and presented auditorily to children or presented in written form on the screen (for adults). A researcher annotated participants' answers on a sheet of paper. Children were tested in a quiet room of their schools after parents signed a consent form. All of them completed the task. Adult participants were recruited on a voluntary basis.

RESULTS

The distribution (in percentages) of children's and adults' responses on the ternary scale for True and False controls and critical items is summarized in **Table 1**.

All participants responded with the fully positive option in True controls (>95% of gold medals in all groups); in general, they also rejected False controls above 90% of the time by selecting either the bronze or the silver medal. Specifically, they selected the bronze medal above 60% of the time. The high percentage of silver medals for control statements that were literally false was somehow unexpected; however, this was mainly due to the items that constituted partially true descriptions of the situation, like the long statement discussed before, in which two out of three objects that were indeed present on the desk were mentioned. In the target conditions, i.e., those that involved a violation of pragmatic principles, the older children and adults selected the middle option in the majority of the cases, as expected, while the younger children chose the gold medal in the majority of the cases, although about one fourth of the time, they selected the silver medal. All groups tended to be more polarized in their answers in the control conditions compared to the target conditions.

1To evaluate the pattern of responses between target and control conditions, and among age groups, we implemented a mixed-effects ordinal regression model with a logit link function using the *clmm*() function in the *ordinal* package (Christensen, 2018). This is a statistical model specifically designed to treat ordinal-dependent measures that cannot be assumed to represent an interval scale, as it is the case with the ternary option used in our study. The maximal model that converged included Condition (control vs. target) and age Group (preschoolers, primary-school children, and adults) as fixed effects, and their interactions, as well as participants and items as random intercepts. We used dummy coding for Condition and age Group so that control items and primary-school children served as the baselines in the contrasts. The model (**Table 2**) reported a difference in the distribution of medals between the younger and the older children, while no difference was revealed between primary-school children and adults. This suggests a developmental trend that was further explored in a second model. No fully significant interactions were reported. However, we have to take this result with some caution due to the relatively small number of data points considered in the analyses.

We then ran a second model in which we contrasted the experimental items (settled as the baseline) with True and False controls considered separately in the three age groups. This model revealed a significant difference of primary-school children both with the preschool children and adults, as well as significant interactions between item Type and age Group (**Table 3**). The interaction between Type and age Group was significant in the comparison between the younger and the older children in the case of False controls ($p = 0.002$), while it only approached significance in the case of True controls ($p = 0.057$), again suggesting a developmental trend in children's ability to conform to the task and to detect the violations of conversational maxims, which improved with age. A significant interaction of Type and Group was also revealed in the case of True controls between primary-school children and adults, suggesting that the older children, despite being more pragmatically mature than the younger children, were not fully adult-like yet in the treatment of statements that violated a maxim, accepting them at a higher rate compared to adults.

To investigate whether a different pattern was revealed across different types of violations of maxims, and across different age groups, we focused on the target items only. The distribution of the response options across maxims violations and age groups is plotted in **Figure 2**; the mean percentages are reported in **Table 4**.

TABLE 1 | Distribution of response types (bronze/silver/gold medals) in the experimental target conditions (violations of maxims and of Maximize Presupposition) and in the True/False control conditions.

	Preschool children			Primary-school children			Adults		
	Bronze medal	Silver medal	Gold medal	Bronze medal	Silver medal	Gold medal	Bronze medal	Silver medal	Gold medal
True	0%	2.2%	97.8%	0.5%	6.4%	93.1%	0.3%	4.5%	95.1%
False	66.1%	25.6%	8.3%	63.3%	34.1%	2.5%	71.5%	27.8%	0.7%
Target	15.7%	24.4%	59.8%	11.3%	51.1%	37.6%	9.5%	69.4%	21.1%

TABLE 2 | Output of the model with type of Medal as the dependent variable, Condition and age Group as independent variables (dummy coded with control and primary-school children as baselines), and subjects and items as random intercepts.

	Estimate	Std. Err.	*z*-value	*p*-value
Cond (target vs. control)	−1.386	1.203	−1.152	0.249
Group (adults vs. primary-school children)	−0.201	0.252	−0.798	0.425
Group (preschool vs. primary-school children)	0.536	0.239	2.245	0.025
Cond: Group (adult)	−0.388	0.217	−1.783	0.075
Cond: Group (preschool)	0.312	0.214	1.460	0.144

TABLE 3 | Output of the model with type of Medal as the dependent variable, Type (target items vs. False/True controls) and age Group as independent variables (with primary-school children as the baseline), and subjects and items as random intercepts.

	Estimate	Std. Err.	*z*-value	*p*-value
Type (false vs. target)	−3741	0.698	−5.361	<0.001
Type (true vs. target)	3.747	0.566	6.619	<0.001
Group (adults vs. primary-school children)	−0.589	0.206	−2.861	0.004
Group (preschool vs. primary-school children)	0.859	0.198	4.343	<0.001
Type (false): Group (adult)	0.021	0.273	079	0.937
Type (true): Group (adult)	0.930	0.333	2.792	0.005
Type (false): Group (preschool children)	−0.817	0.257	−3.175	0.002
Type (true): Group (preschool children)	0.794	0.416	1.907	0.057

We ran four separate models to compare the different maxims' violations, as well as the interaction for each of the age groups, by changing the level of the variable to serve as the baseline for each of the contrasts. For better readability, we discuss the main results in the paper and refer to the Appendix for the full model outputs (**Appendix Tables A–D**). In general, all maxim violations differ from the others (all $p < 0.001$) except for the maxims of Quantity and Relation, which do not differ ($p = 0.783$). As for a general effect of age Group, primary-school children differ from preschool children in detecting the violations of the maxims of Manner, Quantity, and Relation (all ps < 0.05). In these conditions, the preschool children tended to be more tolerant than the older children, selecting the gold medal most of the time, compared to the older children, who selected the silver medal in most of the cases. Comparing primary-school children with adults, the only significant difference is in the detection of the violation of the Maximize Presupposition principle: in this case, the primary-school children tended to be significantly more tolerant than adults, selecting the gold medal at a higher rate compared to adults ($p < 0.001$). This is also captured by the three significant interactions of Group (adults vs. primary-school children) and type of Violation in the contrasts in which Maximize Presupposition is set as the baseline (all ps < 0.001).

The maxim of Manner was tested with two items that violated the submaxim Be brief and two items that violated the submaxim Be orderly. As discussed above, unnecessary lengthy descriptions were viewed by Surian et al. (1996) as involving the second submaxim of Quantity, whereas we considered them as related to Be brief. It is therefore relevant to further inspect these items and in particular to verify whether they are treated like Be orderly or like Quantity I violations. In **Table 4**, we provide the distribution of medals for all pragmatic violations, splitting the maxim of Manner in its two submaxims, Be brief and Be orderly, in all age groups.

We ran three additional models, one for each age group separately, in which we considered submaxims as the independent variable, setting Be brief as the baseline. In this way, we could evaluate the comparison between violations of the two submaxims of Manner, as well as the difference, if any, between Be brief and the first submaxim of Quantity in each age group. In all age groups, no difference was revealed in the distribution of medals between Be brief and Be orderly (all ps > 0.05). As for the comparison between Be brief and the first submaxim of Quantity, the models revealed a significant difference in adults and primary-school children (all ps < 0.001); in preschool children, instead, no difference was revealed between any of these conditions, as in all cases, young children seem to equally tolerate such violations ($p = 0.258$, see **Appendix Tables E–G** for the full output of the models).

DISCUSSION

In this study, we aimed at assessing whether children are sensitive to violations of pragmatic principles, testing unmotivated violations of the Gricean maxims of Quantity, Relation, and Manner, and of the Maximize Presupposition principle (violations of the maxim of Quality were controlled by means of False controls). In particular, we wanted to determine whether children's pragmatic competence improves with age and whether there is a difference in the sanctioning of violations of different maxims and of the Maximize Presupposition principle. We used a TVJ task with three options of response, since Katsos and Bishop (2011) found that a binary task could blur children's ability to recognize the infelicity of underinformative sentences. All the critical items constituted literally true statements that were pragmatically inappropriate in the given scenario because they were infringing one of the pragmatic tenets. The expected answer, then, was a rejection of these remarks, assigning to Bruno, the speaker, the silver or the bronze medal.

Taking into account the contrast between control and target items, the first comparison across the age groups revealed an evolutionary trend, with primary-school children differing from the younger children in their choices, but not from adults. In the case of a violation of a maxim, the younger children selected the gold medal almost 60% of the time, while about 37 and 21% of the older children and adults did so, respectively, showing that pragmatic violations are sanctioned more with age. In fact, the majority of the older children and the adults sanctioned the infelicitous statement by selecting the silver or the bronze medal, although the older children overall accepted the violations of the maxims more than the adults, as emerged in the second

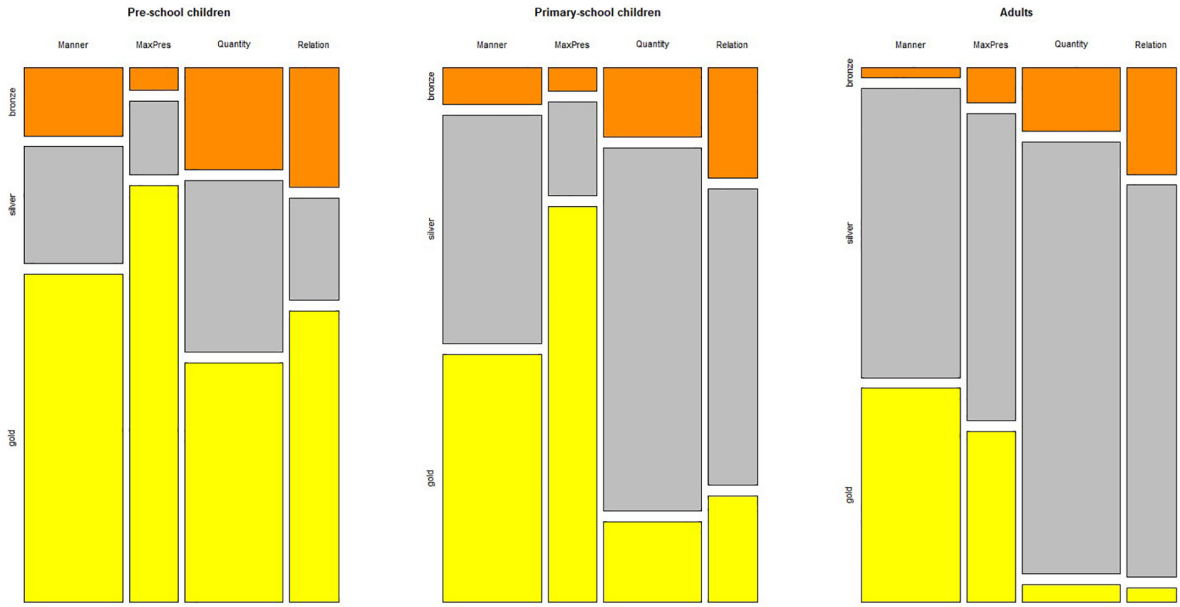

FIGURE 2 | Distribution of response options across pragmatic violations and age groups.

TABLE 4 | Distribution of medals (in percentages) for all the types of violations (with the submaxims of Manner, Be brief and Be orderly, provided separately) and all age groups.

	Preschool children			Primary-school children			Adults		
	Bronze medal	**Silver medal**	**Gold medal**	**Bronze medal**	**Silver medal**	**Gold medal**	**Bronze medal**	**Silver medal**	**Gold medal**
MaxPres	4.4%	14.4%	81.1%	4.7%	18.2%	77.1%	6.9%	59.7%	33.3%
Manner: Be brief	16.7%	23.3%	60%	2.1%	53%	45%	2.8%	66.7%	30.6%
Manner: Be orderly	10%	22.2%	67.8%	12.3%	36%	51.7%	1.4%	45.8%	52.8%
Quantity I	20%	33.3%	46.7%	13.6%	70.8%	15.7%	12.5%	84.0%	3.5%
Relation	23.3%	20.0%	56.7%	21.6%	57.6%	20.8%	20.8%	76.4%	2.8%

model. Notice that, following Katsos and Bishop (2011), the TVJ task contemplated three options of response; still for younger children, the gold medal constitutes the preferred choice for all pragmatic violations.

Focusing on the pragmatic principles, we found a different degree of tolerance depending on the kind of violations. First of all, the items in which the maxims of Quantity or Relation were violated behaved similarly and were significantly less accepted than all the other violations across all age groups. In these cases, adults rejected the statements 97% of the time. Similarly, these violations were the most sanctioned by children. Related to this point, it is worth mentioning that all participants sanctioned False controls, which are indeed violations of the maxim of Quality, above 90% of the time.

Second, a difference was revealed both in adults and primary-school children between their reaction to the violation of the submaxim of Be brief (that in the CVT was regarded as a violation of Quantity II) and of Quantity I, speaking in favor of a different treatment of these two types of violations.

Third, it is interesting to notice that adults behaved somehow unexpectedly in the case of violations of the maxim of Manner and of the Maximize Presupposition principle. As is evident from **Figure 2**, they often chose to assign the gold medal to items that infringe these tenets, accepting unnecessary long answers or statements in which events are reported in reverse order (violations of the maxim of Manner received a gold medal 42% of the time) and in which the indefinite determiner *a* is used instead of the stronger, presuppositional triggering determiner *the* (violations of the Maximize Presupposition principle received a gold medal 33% of the time). Indeed, children accepted statements that violated the Maximize Presupposition principle even more than the adults, suggesting that this type of violation is fully accepted in most cases (77 and 81% of gold medals for older and younger children, respectively).

We will discuss the implications of these findings in turn. First of all, the fact that violations of Quantity and Relation differ from all the others highlights the different statuses of pragmatic principles. As already alluded to, Grice himself suggested that the observance of the maxims could be a matter of more or less "urgency." The maxim of Quality has always been considered to be special, since its violation does not to lead to simply "an

inferior kind of information; it just is not information" (Grice, 1989, p. 371). Our results corroborate this, showing a high rate of sanctioning of False controls by adults and children. In our study, though, we also found evidence for a clear differentiation also among the maxims of Quantity and Relation, on the one hand, and of the maxim of Manner and of the Maximize Presupposition principle on the other hand: while adults (and children) always sanction the maxims of Quantity and Relation, they are more tolerant with violations of the maxim of Manner and the Maximize Presupposition principle, accepting them to a higher degree.

To account for this difference, we can notice how violations of these maxims result in different outcomes: if a speaker violates the maxim of Quantity or Relation, the hearer cannot really understand what has happened. In our example, for instance, when Bruno says that he drew a dog, the blindfolded Elm will come to believe that the dog was the only thing that Bruno drew, under the assumption that Bruno is cooperative, and he will thus form an incorrect belief about the situation, since Bruno also drew a cat. Again, when Bruno answers that he ate food for lunch or that his favorite shirt has two sleeves and a hole for the head, Elm does not have the necessary and relevant information that would permit to understand what Bruno ate (a chicken) and what his favorite shirt looks like (red, with a monkey on it). The statements that infringe the maxims of Quantity (at least the first submaxim, Be enough informative) and of Relation, in other words, transmit inaccurate or incomplete information. The two submaxims of Manner Be brief and Be orderly and the Maximize Presupposition principle, on the other hand, regulate the form, and not the content, of the statements. If speakers provide unnecessary long descriptions, or if they report the events in a reverse order, or if they fail to use the stronger, presuppositions triggering expressions, they still transmit a piece of information that enables interlocutors to understand what has really happened. Thus, the rigmarole "a fruit with yellow peel and that monkeys really like" enables interlocutors to understand that the speaker is referring to a banana; hearing "I went to the bed and I brushed my teeth" permits Elm to understand which events took place, even if they are mentioned in the wrong order; similarly, the anomalous statement "A sun is setting" correctly depicts the situation of a sunset. We therefore hypothesize that the interlocutors sanction infringements of the maxims in different ways, rejecting those that transmit inaccurate or incomplete information and being more tolerant with those that still permit to understand what has happened.

To further speculate on this finding, we discuss an interesting parallelism that comes from the literature on referential expressions. Engelhardt et al. (2006) tackled the question whether adults are sensitive to the first (Be enough informative) and to the second (Do not be too informative) submaxim of Quantity. They focused on descriptions of a target referent that were underinformative (e.g., "the apple" when two different apples were present) or overinformative (e.g., "the apple on the towel" when only one apple was present) in tasks that required to produce or comprehend commands such as "Put the apple (on the towel) in the box." They found an asymmetry: referential expressions that did not provide enough information for the correct identification of the target (violating Quantity I) were never produced and were penalized in comprehension; descriptions that provided more information than what was strictly required to identify the referent (violating Quantity II) were spontaneously produced 30% of the time and were not rated significantly lower than optimally concise descriptions. Quite interestingly, though, in a third eye-tracking experiment, they found that overinformative descriptions did cause momentary confusion, indicating that unnecessary modifications are costlier to process. Even if Davies and Katsos (2013) argued that the "only moderately Gricean" behavior of participants in Engelhardt et al. (2006) studies might be due to methodological confounds linked to the complexity and visual salience of the array of stimuli, it is interesting to comment on these data.

Analogously to what was observed by Engelhardt et al. (2006), we found that violations of Quantity I are always sanctioned; following their explanation, we argue that this might be so because a lack of necessary information "can compromise communication, as an under-described utterance will not permit a listener to identify the correct referent from a set" (Engelhardt et al., 2006, p. 563); being overinformative, on the other hand, may cause a temporal ambiguity but does not block the identification of the referent and, for this reason, might be less sanctioned and even be spontaneously produced.

As already discussed in the *Introduction*, it is "disputable" whether unnecessary long descriptions constitute violations of Quantity II or of Be brief. What these kinds of "harmless" violations of the maxims have in common is that they do not lead to a communication failure, as much as violations of Quantity I (and of Relation) would do, and they are therefore less sanctioned. We argue that violations of Quantity II, "on the assumption that the existence of such a maxim should be admitted" (Grice, 1975, p. 52), are analogous to infringements of the submaxims of Manner Be brief and Be orderly and of the Maximize Presupposition principle.

With respect to the final point listed above, our results for the Maximize Presupposition principle limit the scope of Sauerland (2008)'s claims: even if Maximize Presupposition and Quantity I behave similarly, in that they both demand the choice of the stronger alternative statement, they also differ because violations of the former are viewed as less detrimental compared to statements that do not transmit enough information. In fact, the extremely low rejection of statements that violate Maximize Presupposition contrast with the results in Yatsushiro (2008), who found that children already reached 70% accuracy at age 6, even if they performed worse than the 7-year-olds, and only performed adult-like at age 8. In our study, we found that primary-school children, aged 6–9, rewarded the violations of the Maximize Presupposition principle with a gold medal 77% of the time. This difference might be due to the type of Maximize Presupposition violation presented in the two studies or to the type of task used. We presented statements with the indefinite determiner *a* instead of the more appropriate *the*, whereas Yatsushiro (2008) contrasted the definite description with the determiner *every*. Moreover, we used a TVJ task with

ternary options, whereas Yatsushiro (2008) employed an FJ task. As already discussed, FJ tasks might overestimate children's actual pragmatic competence (see also Foppolo et al., 2021): choosing the more appropriate alternative does not necessarily mean that when presented with the pragmatic infelicitous statement, children would reject it, as already discussed by Foppolo et al. (2012).

To conclude, the results of the present study offer potentially interesting lines of research. On the one hand, the question whether interlocutors are only "moderately" (Engelhardt et al., 2006) or "fully" (Davies and Katsos, 2013) Gricean remains open and should be further explored. We found that adult participants always penalized statements that violated the maxims of Quality, Quantity I, and Relation, whereas they were more tolerant with regard to infringements of the two submaxims of Manner Be brief and Be orderly. Even if this observation is in our view coherent with Grice's own view, since he explicitly recognized that the observation of maxims is indeed a matter of more or less urgency, it is worth exploring the conditions under which interlocutors might attribute more or less importance to the maxims of conversation. As already alluded to, Davies and Katsos (2013) argued that the apparent tolerance of overinformative statements in Engelhardt et al. (2006) might be imputed to contextual factors: being presented with a complex array of objects, participants might prefer overdescriptions to more concise expressions to avoid ambiguity. Still, as Engelhardt (2013) highlighted, the conditions that might push interlocutors to prefer avoiding possible ambiguity instead of choosing optimally concise expressions need to be further explored. Also, consider that the visual display in our task was extremely simple, and the excessively verbose description was referred to a single object.

Another factor that might have sharpened the difference between the violations of Quantity I and Relation, on the one hand, and of violations of Be brief, Be orderly, and Maximize Presupposition, on the other, is the fact that in our task, participants had to evaluate a single speaker, Bruno, who uttered all the sentences, including blatantly false ones. For this reason, he could be considered unreliable as a speaker. As Grodner and Sedivy (2011) have shown, when the speaker is presented as unreliable, adult participants do not interpret restrictive modifiers contrastively because they are aware of the fact that the speaker consistently produces overinformative utterances. We might then hypothesize that our participants—possibly also the younger ones—realized that Bruno was not reliable; thus, they decided to sanction the more harmful violations, those that effectively led to communication failures, and overlooked those that were considered to be more innocuous for the purpose of understanding what has happened, which, ultimately, was what they were asked to do.

In the end, we show that if you are asked what you have eaten, and you ate a kiwifruit, the answer "a fruit with a hairy brown skin and a green flesh," albeit long and weird, is preferable compared to "a banana" (false answer), or "food" (underinformative), or "something with the spoon" (irrelevant).

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the ethical committee of the University of Milano-Bicocca. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

FP conceived the experimental question, developed the tasks, recruited the children, and supervised the testing. FF performed the statistical analyses. FP drafted the manuscript, which was critically revised by FF. Both authors contributed to the article and approved the submitted version.

FUNDING

Funding was partially from funds provided by the Department of Psychology (University of Milano-Bicocca) for open access publication fees and partly from the authors' personal research funds.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.624628/full#supplementary-material

REFERENCES

Christensen, R. H. B. (2018). *Cumulative Link Models for Ordinal Regression with the R Package Ordinal. Submitted in J. Stat. Software.*

Davies, C., and Katsos, N. (2013). Are speakers and listeners 'only moderately Gricean'? An empirical response to Engelhardt et al.(2006). *J. Pragmat.* 49, 78–106. doi: 10.1016/j.pragma.2013.01.004

Engelhardt, P. E. (2013). Are listeners 'fully' Gricean?: a comment on Davies and Katsos (2013). *J. Pragmat.* 49, 107–113. doi: 10.1016/j.pragma.2013.01.001

Engelhardt, P. E., Bailey, K. G., and Ferreira, F. (2006). Do speakers and listeners observe the Gricean Maxim of Quantity? *J. Memory Lang.* 54, 554–573. doi: 10.1016/j.jml.2005.12.009

Foppolo, F., Guasti, M. T., and Chierchia, G. (2012). Scalar implicatures in child language: give children a chance. *Lang. Learn. Dev.* 8, 365–394. doi: 10.1080/15475441.2011.626386

Foppolo, F., Mazzaggio, G., Panzeri, F., and Surian, L. (2021). Scalar and ad-hoc pragmatic inferences in children: guess which one is easier. *J. Child Lang.* 48, 350–372. doi: 10.1017/S030500092000032X

Grice, H. P. (1975). "Logic and conversation," in *Syntax and Semantics*, Vol. 3, eds P. Cole and J. L. Morgan (New York, NY: Academic Press), 41–58. doi: 10.1163/9789004368811_003

Grice, H. P. (1989). "Retrospective epilogue," in *Studies in the Way of Words*, ed. P. Grice (Cambridge, MA: Harvard University Press), 339–386.

Grodner, D., and Sedivy, J. C. (2011). "The effect of speaker-specific information on pragmatic inferences," in *The Processing and Acquisition of Reference*, eds E. A. Gibson and N. J. Pearlmutter (Cambridge, MA: MIT Press), 239–271. doi: 10.7551/mitpress/9780262015127.003.0010

Heim, I. (1991). "Artikel und Definitheit," in *Semantik: Ein Internationales Handbuch der Zeitgenössischen Forschung*, eds A. von Stechow and D. Wunderlich (Berlin: De Gruyter), 487–535.

Katsos, N., and Bishop, D. V. (2011). Pragmatic tolerance: implications for the acquisition of informativeness and implicature. *Cognition* 120, 67–81. doi: 10.1016/j.cognition.2011.02.015

Okanda, M., Asada, K., Moriguchi, Y., and Itakura, S. (2015). Understanding violations of Gricean maxims in preschoolers and adults. *Front. Psychol.* 6:901. doi: 10.3389/fpsyg.2015.00901

Pouscoulous, N., Noveck, I. A., Politzer, G., and Bastide, A. (2007). A developmental investigation of processing costs in implicature production. *Lang. Acquis.* 14, 347–375. doi: 10.1080/10489220701600457

Sauerland, U. (2008). "Implicated presuppositions," in *The Discourse Potential of Underspecified Structures*, ed. A. Steube (Berlin: Mouton de Gruyter), 581–600. doi: 10.1515/9783110209303.4.581

Siegal, M., Iozzi, L., and Surian, L. (2009). Bilingualism and conversational understanding in young children. *Cognition* 110, 115–122. doi: 10.1016/j.cognition.2008.11.002

Siegal, M., Surian, L., Matsuo, A., Geraci, A., Iozzi, L., Okumura, Y., et al. (2010). Bilingualism accentuates children's conversational understanding. *PLoS One* 5:e9004. doi: 10.1371/journal.pone.0009004

Skordos, D., and Papafragou, A. (2016). Children's derivation of scalar implicatures: alternatives and relevance. *Cognition* 153, 6–18. doi: 10.1016/j.cognition.2016.04.006

Surian, L., Baron-Cohen, S., and Van der Lely, H. (1996). Are children with autism deaf to Gricean maxims? *Cogn. Neuropsychiatry* 1, 55–72. doi: 10.1080/135468096396703

Surian, L., Tedoldi, M., and Siegal, M. (2010). Sensitivity to conversational maxims in deaf and hearing children. *J. Child Lang.* 37:929. doi: 10.1017/S0305000909990043

Winner, E. (1997). *The Point of Words: Children's Understanding of Metaphor and Irony*. Cambridge, MA: Harvard University Press.

Yatsushiro, K. (2008). "Quantifier acquisition: presuppositions of 'every," in *Proceedings of Sinn und Bedeutung 12*, ed. A. Grønn (Oslo: ILOS), 663–676.

Conflict of Interest: 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 © 2021 Panzeri and Foppolo. 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) and the copyright owner(s) 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 March 2021
doi: 10.3389/fpsyg.2021.616065

Context Breeds False Memories for Indeterminate Sentences

*Levi Riven and Roberto G. de Almeida**

Department of Psychology, Concordia University, Montreal, QC, Canada

What are the roles of semantic and pragmatic processes in the interpretation of sentences in context? And how do we attain such interpretations when sentences are deemed indeterminate? Consider a sentence such as "*Lisa began the book*" which does not overtly express the activity that Lisa began doing with the book. Although it is believed that individuals compute a specified event to enrich the sentential representation – yielding, e.g., "*began [reading] the book*" – there is no evidence that a default event meaning is attained. Moreover, if indeterminate sentences are enriched, it is not clear where the information required to generate enriched interpretations come from. Experiment 1 showed that, in isolation, there is no default interpretation for indeterminate sentences. The experiment also showed that biasing contexts constrain event interpretations and improve plausibility judgments, suggesting that event representations for indeterminate sentences are generated by context. In Experiment 2, participants heard biasing discourse contexts and later falsely recognized foil sentences containing the biased events ("*Lisa began reading the book*") at the same proportion and with the same confidence as the original indeterminate sentence ("*Lisa began the book*"). We suggest that indeterminate sentences trigger event-enriching inferences but only in sufficiently constraining contexts. We also suggest that indeterminate sentences create two memory traces, one for the proposition consistent with the denotational, compositional meaning, and another for the proposition that is enriched pragmatically over time.

Keywords: indeterminate sentences, compositionality, false memory, pragmatics, inferences, semantic coercion, propositional representation, sentence comprehension

OPEN ACCESS

Edited by:
Valentina Cuccio, University of Messina, Italy

Reviewed by:
Alessandro Lenci, University of Pisa, Italy
Nikole Patson, The Ohio State University, United States

***Correspondence:**
Roberto G. de Almeida
roberto.dealmeida@concordia.ca

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

Received: *10 October 2020*
Accepted: *10 February 2021*
Published: *12 March 2021*

Citation:
Riven L and de Almeida RG (2021) Context Breeds False Memories for Indeterminate Sentences. Front. Psychol. 12:616065. doi: 10.3389/fpsyg.2021.616065

INTRODUCTION

We are rarely faced with the task of understanding sentences in isolation. Most often, linguistic expressions are understood and produced in rich utterance contexts, allowing us to interpret rather easily a variety of incomplete or anomalous expressions, such as disfluencies (*Well… uh…she…left!*) and metaphors (*He's a pig!*) – which are literally false but invite us to seek alternative interpretations. An arguably more subtle case is that of sentences deemed indeterminate or underspecified. For instance, if you were told upon breaking into a sneezing fit, *Do not worry, you are not going to die*, you would not likely interpret this sentence as declaring your immortality, nor would you think it is false. Rather, you would take the sentence to convey that you are not going to die *as a consequence of* your sneezing. Such indeterminacies are ubiquitous in natural language and are generally taken to be resolvable by simply filling-in the "blanks" – or what Perry (1986) called "unarticulated constituents" – with information supplied by context or by some default semantic operation.

Here, we report on two experiments investigating the interpretations that participants assign to indeterminate sentences in isolation and in context. The phenomenon we investigated, more specifically, involves sentences such as (1):

1. Lisa began the book.

While it is clear that Lisa began doing something with a book, the sentence is indeterminate with regards to what exactly Lisa began doing. This kind of construction has been the object of investigation in theoretical linguistics (Pustejovsky, 1995, 2011; de Almeida and Dwivedi, 2008; de Almeida and Riven, 2012; Asher, 2015; see also de Swart, 2011, for review), psycholinguistics (e.g., McElree et al., 2001; Traxler et al., 2002; de Almeida, 2004), and neuroimaging (Pylkkänen and McElree, 2007; Husband et al., 2011; de Almeida et al., 2016). What motivates general interest in indeterminate sentences is that, although they are grammatical and semantically felicitous – for a truth value judgment *can* be made – they appear to convey more information than what is explicitly said. Specifically, a sentence such as (1) in isolation is compatible with a wide range of events – such as *reading*, *writing*, *burning*, or even *eating* – each of which might serve as a suitable interpretation. However, it is not clear what exactly might be attained, nor which linguistic and cognitive systems are deployed in attempting to resolve indeterminacy.

Thus far, there have been several proposals on how these sentences are interpreted in isolation and in context. These proposals vary along several dimensions, including the degree to which the sentence is enriched, the role of context in the enrichment process, the formal mechanisms for this enrichment, and the source of information employed in enriching indeterminate sentences (e.g., Pustejovsky, 1995, 2011; de Almeida and Dwivedi, 2008; Pylkkänen, 2008; Asher, 2015). Some proposals appeal to *type-shifting* rules which work to *shift* the semantic types of constituents to allow for semantic composition – thus making the verb and noun-phrase complement "fit" together. Other proposals go beyond, appealing to a form of lexical-semantic *interpolation* – viz., by augmenting the proposition that a sentence such as (1) conveys, yielding a particular event in the resulting proposition, similar to what would be conveyed by (2).[1]

2. Lisa began *reading* the book.

We briefly discuss some of these proposals below, but suffice it to say now that, for versions of the *interpolation* view (e.g., Pustejovsky, 1995, 2011; Lapata and Lascarides, 2003; Traxler et al., 2005) the meaning of (1) might be quasi-synonymous with (2). To a large extent, this assumption follows from the earliest studies of indeterminacy which proposed that sentences such as (1) appear almost as frequently in corpora as fully-determinate variants such as (2) (e.g., Briscoe et al., 1990). Moreover, an early empirical study found that participants rate indeterminate sentences as being just as sensible as their fully determined controls (McElree et al., 2001). Such findings suggest that indeterminate sentences trigger a systematic enrichment process that is believed to resolve indeterminacy, thus allowing for frequent and felicitous use in natural speech contexts.

This enrichment process has since been the object of considerable experimental investigation involving numerous techniques such as self-paced reading (McElree et al., 2001; de Almeida, 2004), eye-tracking (Traxler et al., 2002, 2005; Pickering et al., 2005; McElree et al., 2006a; Frisson and McElree, 2008; Katsika et al., 2012; Zarcone et al., 2014; Antal and de Almeida, 2020), probe recognition (Zarcone et al., 2014), sensibility judgments (McElree et al., 2006b), event-related potentials (ERPs; Baggio et al., 2010; Kuperberg et al., 2010), fMRI (Husband et al., 2011; de Almeida et al., 2016), magnetoencephalography (MEG; Pylkkänen and McElree, 2007), and aphasia (Piñango and Zurif, 2001). The majority of these studies have shown that indeterminate sentences engender processing delays or different activation patterns relative to various types of control sentences. The greater processing costs of indeterminate sentences are thought to correspond to the mental operations associated with *semantic* enrichment of sentences such as (1), leading to the creation of a proposition that corresponds to the content of (2).

However, several observations from recent reports cast doubt on the assumption that indeterminate sentences are fully enriched during online processing. At least four studies have reported plausibility norms for which indeterminate sentences were rated statistically lower than controls (Pylkkänen and McElree, 2007; Frisson and McElree, 2008; Katsika et al., 2012; de Almeida et al., 2016). In addition, one study that measured acceptability ratings incrementally, as sentences were being presented, showed that participants rate indeterminate sentences as unacceptable, suggesting that they might not achieve a fully specified event interpretation (McElree et al., 2006b).[2]

The vast majority of psycholinguistic experiments have emphasized processing times, including response times (RTs) and eye-movement behaviors, over other measures of

[1]We use *indeterminate* because, as we have noted (de Almeida and Riven, 2012), the actual event to which these sentences refer is not overtly specified, making sentence (1) compatible with any event over which *begin* would have scope. We eschew the terms *type-shifting* or *coercion*, as employed in the literature, because they are committed to particular theories on how indeterminate sentences are enriched – namely, by *type-shifting* or *type-coercing* the complement nominal (e.g., Pustejovsky, 1995, 2011; Pylkkanen, 2008; Asher, 2015), as we briefly review below. The use of the term *indeterminate,* in this regard, is *theory-neutral* with respect to whether or not sentences such as (1) are enriched, and, thus, more in line with the null hypothesis. We also want to avoid confusion with other terms, such as *underspecification*, which have been used in the semantics literature to mean something different from what we claim is indeterminacy. In some circles (see, e.g., Frisson, 2009), underspecification is associated with a representation based on features, some of which are absent in certain uses of a word, on the assumption that words are polysemous and represented as sets of features. We are not committed to features as constituents of lexical meaning, which we see as an empirical and theoretical issue yet to be determined (see also Fodor, 1998).

[2]It is also noteworthy that much of the research reviewed here has focused on a class of verbs that may have two subtypes: aspectual verbs (e.g., *begin*, *finish, continue*, etc.) and psychological or "try" verbs (e.g., *enjoy*, *prefer, attempt,* etc.; see de Almeida and Dwivedi, 2008). Katsika et al. (2012) suggested that these subtypes are processed differently. Although most papers in our review focus on a superordinate class that includes both subtypes, our discussion and, more importantly, the present experiments focus exclusively on the aspectual subtype, which is thought to be more exemplary of the processing effects reported in the literature.

interpretation, and have thus left open the question as to what sort of interpretation indeterminate sentences actually get. While the processing delays associated with such sentences are thought to correspond with some form of enrichment-related process, it may be the case that individuals' interpretations of indeterminate sentences are left indeterminate and in fact look more like (1) than (2). In other words, differences in RTs do not *per se* point to enrichment (de Almeida and Dwivedi, 2008) and, if they do, it is far from clear how sentences might be enriched.

Studies involving more direct measures of brain activity such as blood flow (fMRI) and electrical signals (ERP and MEG) have also consistently showed differences between indeterminate sentences and their controls, and have contributed to an understanding of functional and neuroanatomical resources involved in interpreting these sentences. But these techniques are also limited with regards to what information is attained when sentences are understood. One advantage of neuronal recording techniques (MEG and ERPs) over RTs and fMRI is that neuronal recordings yield information about the time-course of events associated with processing of indeterminate sentences. The study by Kuperberg et al. (2010), for instance, found a small but significant N400 effect in the contrast between indeterminate and control sentences in two grand-averaged sites. But similar to RT and MEG studies, these commonly found differences still call for an explanatory framework as to what sort of *content* comes into play. In the case of fMRI studies, results have also been far from conclusive. For instance, one study (Husband et al., 2011) supports a mandatory *type-shifting* effect, based on greater left-frontal and left-temporal activation for indeterminate sentences. Another study (de Almeida et al., 2016) calls for a pragmatic, inferential resolution, based on greater whole-brain activation, and in particular greater right-hemisphere and anterior cingulate cortex (ACC) activation, for indeterminate sentences, signaling a greater search for an interpretation. While the details of all their analyses are beyond the focus of the present paper, it is important to note that, as with other techniques, fMRI also does not allow for a clear understanding of what sort of content indeterminate sentences call for and when they are enriched.

We thus sought to investigate, in the present study, the nature of the *propositional content* elicited by indeterminate sentences in context and over time. Contrary to the studies briefly mentioned above, we were primarily focused on understanding what is retained in memory, both immediately and after a delay. We were motivated, in particular, by the classic effect of "gist" first obtained by Sachs (1967, 1974). In those studies, Sachs showed that participants quickly forget the *verbatim* form of a sentence and falsely recognize foils if these preserve the gist – i.e., propositional content – of the original stimulus sentence. Our goal was to go beyond this effect and determine if subjects would falsely recognize *Lisa began reading the book* (2) as being the originally presented *Lisa began the book* (1), within a biasing "book reading" context. While one would hardly dispute that discourse context exerts an influence on how one might ultimately interpret a sentence (see, e.g., O'Brien and Cook, 2015, for a review), our goal was to focus on the propositional content that is obtained as an indeterminate sentence and is heard and what becomes of this proposition in memory over time. Thus, beyond response times and neuronal recordings – which have provided us important insights into the processing of indeterminate sentences in isolation (but see de Almeida, 2004 and Traxler et al., 2005) – our goal was to focus on the nature of the proposition encoded in memory, in context, and over time.

(How) Are Indeterminate Sentences Enriched?

In order to further elaborate on the questions that motivate our study, it is important to examine how we might interpret indeterminate sentences in context. It seems clear that, if a sentence such as (1) is uttered in a discourse context related to reading, it may convey information that is compatible with what (2) says. Clearly, however, the two sentences – (1) and (2) – are not synonymous, even if the intention of the speaker of (1) is to communicate that Lisa began *reading* a book. The truth conditions of the two sentences also differ radically, because (2) but not (1) is true only if Lisa began *reading* a book, whereas (1) is true no matter what Lisa began doing with the book. Thus, while (1) and (2), on the surface, convey two different propositions, interpolation proposals for enrichment would assume that the proposition conveyed by (1) is actually modified to convey something different from what its meaning communicates at face value – something akin to (2). The studies mentioned above, involving several reading-time measures (self-paced reading and eye-tracking) as well as those involving ERP and MEG, have in fact indicated that the process of enrichment occurs rapidly, at or right after the processing of the complement noun (*book*). This suggests that the proposition conveyed by (1) is possibly fully formed as something like (2) by *default* (Pustejovsky, 1995, 2011; Lapata et al., 2003). While this is a viable outcome of how enrichment might work, there is no direct experimental evidence for this process other than differences in RTs and activation patterns, as discussed above, which are compatible with several theoretical explanations.

One such explanation attributes the greater processing cost of indeterminate sentences to the process of *type-shifting* the complement noun from an *entity* complement (*book*) to an *event* performed with the referent of the noun complement. The standard version of this theory (e.g., Partee, 1986) proposes a basic set of semantic types for noun phrases (NPs; e.g., *quantificational*, *entity/referential*, and *predicative*) with these NPs changing (viz., *shifting* – such as *lowering* or *lifting*) their types according to the requirements of the verb in order to semantically compose. An aspectual verb such as *begin*, by hypothesis, requires an *event* complement. Given that the NP *the book* may have a default *entity* type, it shifts to an *event* to compose with the verb. A type-shifting rule is not, in principle, a form of content enrichment – that is, it does not provide content to the resulting proposition other than changing the reading (thus, the computation) of the *entity* nominal into an *event* to allow for semantic composition.

Perhaps the most influential view in this camp is that of *generative lexicon* theory (GL; Pustejovsky, 1995, 2011; see also

Pustejovsky and Batiukova, 2019, for a recent review). GL proposes a greater variety of semantic *types* together with information that lexical representations ("qualia structure") encode about the meanings of nouns. The mechanism for how indeterminate sentences are enriched, more specifically, involves at least two component processes: (1) the retrieval of an event such as *reading*, supposedly stored within the conceptual representation of the noun *book* (the "qualia structure"); and (2) interpolation of the retrieved event into the indeterminate verb phrase (VP) *began [reading] the book*. Although this view has undergone several changes over the years, the basic mechanisms stand: in its more current version, GL proposes at first a form of type-shifting that relies on the subtypes carried by nominals. These subtypes (e.g., *physical* and *information*) are part of the qualia structure of the lexical entry and are associated with particular roles, such as *telic* (the purpose; e.g., *read* for *book*). In such cases, one of the operations involves selecting one of the subtypes for the noun complement and recover the role from its qualia structure. This qualia information is then used to enrich the representation of the sentence, adding semantic material to its logical form, that is, providing "a potential (default) interpretation for the predicate associated with the event" (Pustejovsky, 2011, p. 1422). In the GL framework, then, the analysis of a sentence such as (1) involves (i) the detection of a mismatch between the verbs' restrictions (viz., requiring an *event* complement) and the semantic type of the complement, followed by (ii) the insertion or *interpolation* of a plausible *event* to yield an enriched semantic composition similar to what sentence (2) conveys.

Although initial evidence of processing delays was taken to support this form of coercion with interpolation (McElree et al., 2001; Traxler et al., 2005; but see de Almeida, 2004), later psycholinguistic experiments suggested that coercion does not necessarily entail the retrieval of a specified event. Using a speed-accuracy trade-off paradigm, McElree et al. (2006b) proposed that participants do not build fully specified event interpretations. In a paradigm that involved incremental acceptability ratings of unfolding sentences, they showed that, while participants were slower to respond to indeterminate sentences like (3a) compared to controls like (3b), they maintained low acceptability ratings for several seconds after the sentence was presented, suggesting that the extra effort associated with processing these sentences did not deliver a fully enriched interpretation.

3.
a. The carpenter began the table.
b. The carpenter built the table.

These findings support type coercion without interpolation, that is, without necessarily activating a conceptual representation of the activity *per se* (e.g., *reading* and *building*). This implies that enriched event conceptualizations – i.e., those denoting specific activities – are not achieved autonomously from the lexical entry of the noun. However, they do not preclude the possibility that readers may access event interpretations *pragmatically* (viz., by deploying inferential processes) when additional constraints are presented to ease interpretation, such as a supportive discourse context.

In an experiment exploring the role of pragmatic constraints on processing times, de Almeida (2004) presented participants with short discourse contexts such as (4a), which were designed to activate knowledge about the type of events that are likely to unfold. These passages were followed by either an indeterminate sentence such as (4b), a preferred control such as (4c), or a non-preferred control such as (4d).[3]

4.
a. The secretary would always be sure to work ahead of schedule. She was asked to work on a memo.
b. The secretary *began* the memo long before it was due.
c. The secretary *typed* the memo long before it was due.
d. The secretary *read* the memo long before it was due.

Previous research had shown that when these indeterminate sentences are presented without context, they are costly relative to both preferred and non-preferred controls (McElree et al., 2001). However, de Almeida found that when context was provided, RTs to indeterminate sentences were slower compared to preferred controls only, and were equivalent to non-preferred sentences. The finding suggests that processing delays might reflect pragmatic-inferential means of enrichment (see Fodor and Lepore, 2002; de Almeida and Lepore, 2018). That is, indeterminate sentences may be more costly in isolation than in context because enrichment requires readers to draw local inferences about possible event interpretations. And readers may fail to retrieve event meanings when no contextual support is provided. Context, however, serves to further constrain interpretations – yielding indeterminate representations on a par with non-preferred interpretations.

This pragmatic proposal, more explicitly, takes the view that an early linguistic analysis of the sentence yields an *unenriched* proposition, one that is left indeterminate and is compatible with the sentence input. The proposal takes an indeterminate sentence such as (1), above, to be grammatical and felicitous, not one that is semantically defective. It assumes that any further enrichment of the initial proposition comes as a function of inferences triggered by the proposition, taking into account all possible sources of information but most importantly the context of the utterance. Crucially, the pragmatic hypothesis makes two proposals that set it apart from the interpolation view: first, the initial proposition is attained as a translation of the input, with no enrichment *by necessity*; and, second, any enrichment is a natural consequence of causal inferential processes, not by appealing to internal analyses of word meanings or what has been called semantic decomposition (see Fodor and Lepore, 2002; de Almeida and Dwivedi, 2008; de Almeida and Riven, 2012; de Almeida and Lepore, 2018; de Almeida and Antal, 2020). This hypothesis, therefore,

[3]The preferred/non-preferred classification was based on the most frequently given verb in frames like *The secretary was _____ the memo* from McElree et al. (2001). Verbs used in the preferred condition were provided more frequently (66%) than verbs used in the non-preferred condition (7%).

places the burden of enrichment – if anywhere – on inferential pragmatics, not on local default semantic processes.

Note that there are different views within the "pragmatic" camp on how the composition of (1) is achieved. For some (Fodor and Lepore, 2012), (1) means that Lisa began doing something with the book. This view relies on the notion that lexical concepts are *atomic* (i.e., nondecompositional) but also carry information about how they combine with other lexical concepts when they compose into propositions. For others (de Almeida and Dwivedi, 2008; de Almeida and Riven, 2012; de Almeida and Lepore, 2018), the *something* that Lisa began doing with the book is syntactically determined, that is, having a more complex VP – a syntactic representation that introduces a phonologically and morphologically empty *verb* node, one that is licensed by the aspectual verb. This node, although not filled semantically, operates as a processing trigger to generate pragmatic inferences about possible events [see de Almeida and Dwivedi, 2008 and de Almeida and Riven, 2012, for details on the syntactic analysis of sentences such as (1)].

It is also important to note that these views are in general agreement with the type-shifting (without interpolation) idea that a sentence such as *Lisa began the book* yields a different computation than a fully determined sentence such as *Lisa read the book*. Both camps assume that the former requires an analysis of the VP – and in particular the relation between the main verb and the complement NP – that relies on either the specification of a syntactic node, more lexical structure, or a semantic type-shifting operation on the complement NP. They thus agree that what Lisa began was an event with the book. In addition, both pragmatic and type-shifting views agree that however the sentence is analyzed, there is no *interpolation* of specified events by default.

A different but compatible proposal, "words-as-cues" (Zarcone et al., 2014) also assumes that inferences about events are generated online, not *via* mandatory semantic operations, with discourse or co-textual lexical units providing clues on what the event is most likely to be about. This proposal – billed as in-between the type-shifting and pragmatic hypotheses – takes the resolution of indeterminate sentences to be a function of the activation of multiple constraints, giving context the role of what the pragmatic theory assumes to be, roughly, that of inferences. It is rather difficult to fully contrast the "words-as-clues" with the pragmatic proposal on grounds that they are committed to different cognitive architectures. That is, while the pragmatic proposal makes a distinction between linguistic (viz., syntactic/semantic) level and pragmatic inferences, the word-as-cues hypothesis is committed to a connectionist/interactive-activation framework, which makes no clear distinctions between lexical units, what they mean (concepts), the propositions they partake (namely, how they compose meaning), and the inferences they trigger. Nonetheless, the view of Zarcone et al. is that information obtained to interpret the complement noun is general, encyclopaedic information, not linguistic *per se*, a view that is compatible with the pragmatic proposal for enrichment.

Those who support the interpolation hypothesis agree in principle that pragmatic inferences contribute to the enrichment process (e.g., Traxler et al., 2005) and, in fact, assume that co-text and context play a role in suggesting interpretations for the indeterminate VP. For instance, in an offline sentence completion task (Lapata et al., 2003), participants more frequently produced *writing* for a sentence fragment such as *The author began ___ the book*, containing an *agentive* subject – i.e., one that specifies the creation of the entity denoted by the noun. But when the agent was *The student*, participants more frequently produced *reading*, thus more in agreement with a *telic* subject – i.e., one that specifies one possible purpose of the entity denoted by the noun. In a neutral context, when the agent was a proper name, however, *reading* was produced more frequently than *writing*, suggesting that the *telic* information (or *role*) is the default event meaning associated with the entity. Lapata et al. suggested that the verb also provides information that might determine the interpretation of the whole verb phrase, that is, in combination with its complement noun, with cases in which a telic interpretation might be preferred (*endure the speech*), others in which an agentive interpretation might be preferred (*regret the speech*) and others in which there is no default (*enjoy the speech*). What these cases seem to show, in fact, is that the hypothesis of default interpretation is highly specific to contexts or situations beyond the information that the agent and complement provide.

Besides, the very idea that the subject NP suggests that what the author began doing is *writing* the book, just begs the question as to how this operation takes place in consonant with the interpolation process. Traxler et al. (2005) suggest that two key processes are triggered by the alleged mismatch between the selectional restrictions of the verb (*begin*) and the complement noun (*book*). According to them,

> "Comprehenders use salient properties associated with the complement noun and other relevant discourse elements (including but not necessarily limited to the agent phrase) to infer a plausible action that could be performed on the noun (Traxler et al., 2005, p. 4)."

And, further,

> "Comprehenders incorporate the event sense into their semantic interpretation of the VP by reconfiguring the semantic representation of the complement, converting [$_{\beta}$ began ($_{\alpha}$ the book)] into [$_{\beta}$ began ($_{\alpha}$ reading the book)]. (Conceivably, this could also require reconfiguration of an associated syntactic representation)" (p. 4).

If we understand this proposal well, these processes call for both, (a) inferences on potential actions performed by the subject over the object, but also (b) an actual semantic interpolation by a chosen activity. Crucial to the present discussion is the role attributed to pragmatics. As with the proposal of Zarcone et al. (2014), where there appears to be no distinct levels of representation for activated units, the proposal of Traxler et al. puts all processes bearing on content (lexical-semantic and pragmatic) in the service of an enriched semantic

composition – and with operations that appear to be over *sentences* (viz., syntactic reconfiguration), not propositions.

While the hypothesis that contextual information beyond sentence constituents may play a key role in the interpretation of indeterminate sentences, thus far few studies since de Almeida (2004) have investigated contextual influence on interpretation directly, and evidence for pragmatic enrichment remains mixed. Traxler et al. (2005), for instance, also manipulated discourse context and showed that including the to-be-inferred event in the immediate discourse did not consistently reduce processing times for indeterminate sentences. For instance, in their Experiment 1, participants were shown passages that either suggested an event such as in (5a) or were neutral, as in (5b), with both followed by an indeterminate sentence (5c) or a control (5d).

5.

a. The contractor had been building in the suburbs.
b. The contractor had been looking for new jobs.
c. That spring, he began a condominium …
d. The spring, he built a condominium …

Supporting the interpolation view, Traxler et al. report several analyses in which reading times for the indeterminate sentence are longer than the control in contexts such as (5a), suggesting that coercion is mandatory, despite contextual information providing an event interpretation (*building*) for the indeterminate sentence. But their results are difficult to interpret as unequivocal support for interpolation, for several reasons. First, their effects were not consistent across experiments manipulating similar, but slightly modified materials. Second, most effects appeared in the verb region rather than in the crucial post-verbal regions. And third, many of the effects were either statistical tendencies or marginal, even when measures of reading time were relatively late, such as in regressions and re-reading times (see, e.g., Traxler, et al., 2005, Experiment 3). Also important is an effect in their Experiment 2 showing greater total fixation times for indeterminate sentences [as in (5c)] at the noun complement region when these sentences followed neutral contexts [as in (5b)] than when indeterminate sentences followed event contexts such as (5a). This is what the pragmatic theory would predict: when an event is suggested by the context, it provides the indeterminate sentence with a possible interpretation; in the absence of such a suggestive context, indeterminate sentences may trigger pragmatic inferences, thus yielding greater costs (see de Almeida and Dwivedi, 2008).

Most RT experiments suggest that indeterminate sentences are difficult to process, but that this difficulty is not necessarily associated with an event retrieval process, be it *via* semantic interpolation or pragmatic inferences. Although the prevailing view is that sentence-enriching inferences occur autonomously and in a cost-free manner (Traxler et al., 2005; Frisson and McElree, 2008), there is no evidence that participants' interpretations actually include such interpolated representations. In fact, there seems to be growing evidence to the contrary in the form of plausibility, acceptability, and cloze norms. As we noted above, a growing number of studies reporting offline norming tasks corroborate McElree et al. (2006b) result of lower acceptability ratings. Although plausibility norms for these indeterminate sentences are consistently higher than anomalous sentences, they are often statistically lower than control sentences (Pylkkänen and McElree, 2007; Frisson and McElree, 2008; Katsika et al., 2012; de Almeida et al., 2016). And, occasionally, comprehension checks suggest that indeterminate sentences are interpreted less accurately (Husband et al., 2011). Even cloze tasks yield differences (Kuperberg et al., 2010) – with, for instance, subjects producing the complement noun ("*article*") significantly more for frames such as *The journalist wrote the…* than for *The journalist began the….*

All together, these data are difficult to reconcile with the view that an event interpretation is automatic, deploying some default interpolation operation. Poor comprehension metrics suggest that there is uncertainty about the events that indeterminate sentences are intended to convey and that they are likely not enriched by a mechanism that involves augmenting the proposition. If event information comes into play at all, it is presumably contingent on the availability of inferential cues within the broader utterance context (de Almeida and Lepore, 2018).

The Present Study

In order to disentangle these issues, we investigated the representation attained in memory for indeterminate sentences and the role that the contextual information plays in possibly suggesting interpretations for these sentences. Experiment 1 addressed the question of whether lower cloze probabilities and acceptability ratings of indeterminate sentences, as obtained in several empirical studies (e.g., McElree et al., 2006b; Kuperberg et al., 2010; Katsika et al., 2012; de Almeida et al., 2016), can be attributed to the absence of a supportive discourse context. Moreover, we were interested in obtaining a "default" meaning as in Lapata et al.'s (2003) neutral condition. In our Experiment 2, then, we further investigated whether contextualized indeterminate sentences trigger event interpretations, using a long-term memory (LTM) recognition paradigm that relies on recovering the propositional content of sentences (Sachs, 1967, 1974). Crucially, Experiment 2 aimed at tracing the "gist" or proposition obtained at the moment the original indeterminate sentence was presented and later, when it had been consolidated in LTM. We reasoned that probing at different points during the presentation of the discourse, as first done by the classic Sachs experiments, could give us information on how context might influence interpretation and thus potentially create false memories, beyond the immediate encoding of the indeterminate sentence.

It is important to stress that it is not under question in the present investigation whether or not context influences how we ultimately attain a particular meaning for a sentence. There are numerous sources of evidence for the influence of context on the interpretation of words and sentences. It is clear – to mention a classic example – that in cases of lexical ambiguity (Swinney, 1979; Tabossi and Zardon, 1993) sentential and even wider context help determine which meaning is attained, even when initially all possibilities are entertained.

And it has been amply demonstrated that information in a text is constantly generating inferences during reading – both local and global – that continually aid in the comprehension of sentences and of these in relation to global discourse (McKoon and Ratcliff, 1988, 1992; see also Kintsch, 1998 and O'Brien and Cook, 2015, for a recent review). Moreover, it has also been demonstrated that attention to sentences and discourse is not always accurate, leading to many interpretation errors that are taken to be at odds with the idea that we are constantly composing meaning that is faithful with the input stimuli (Sachs, 1967; Christianson et al., 2001; Sanford, 2002). But what is not clear is *how* context exerts its effects, that is, how the *proposition* that a sentence conveys is affected by context – whether enriched or impoverished – nor what is ultimately attained when an indeterminate sentence is processed in isolation and in a biasing context. And while we expect discourse to fully propose a resolution for a sentence deemed indeterminate, it is also not clear what happens with the memory trace of the original proposition – whether it is discarded and replaced by one that is supported by context.

EXPERIMENT 1

In this experiment, we examined the role of context in generating event interpretations and acceptability of indeterminate sentences. Typically, online processing studies of indeterminate sentences aim to match experimental and control sentences on measures of plausibility and/or comprehension to validate materials. In several studies, indeterminate sentences have been rated or interpreted on par with controls (McElree et al., 2001; Traxler et al., 2002; Pickering et al., 2005), but they are frequently considered less felicitous (Pylkkänen and McElree, 2007; Frisson and McElree, 2008; Husband et al., 2011; Katsika et al., 2012; de Almeida et al., 2016). Could their relatively low acceptability result from uncertainty about the events that indeterminate VPs refer to? And would participants provide a "default" filler, as obtained by Lapata et al. (2003)? To address these questions, we presented participants with indeterminate sentences in isolation and within discourse contexts in order to (a) evaluate the convergence of event interpretations that participants generate in a fill-in-the-blank task, and (b) measure their acceptability ratings. We hypothesized that embedding sentences within discourse contexts would, first, increase the proportion of participants providing the same event verb in a fill-in-the-blank task; and, second, context would improve acceptability ratings of indeterminate sentences to the level of fully specified control sentences.

Method

Participants

A total of 120 Concordia University undergraduate students divided into four groups participated in two between-subject tasks. Sixty students participated in the fill-in-the-blank task (42 in the no-context condition and 18 in the context condition) and 70 students participated in the ratings task (20 in the no-context condition and 50 in the context condition). All participants were native speakers of English and were compensated with course credit. They all gave informed written consent. The experiment was approved by the Concordia University Human Research Ethics Committee.

Materials and Procedure

Fill-in-the-Blank Task

A set of 19 indeterminate sentence frames such as *Lisa began _____ the book* were presented interspersed with 45 filler sentences lacking a verb such as *The cow _____ the field*. We presented participants with these frames on an Excel spreadsheet, leaving blank the column corresponding to the to-be-filled portion of the sentence. Participants were required to type a word that best completed the sentence, by filling the slot (see **Supplementary Material S1** for full instructions given to participants). All experimental sentences had proper names as subjects to prevent agent-patient semantic associations that could confound the source of the information – i.e., as to whether events were generated from the sentential context or from the broader discourse context. This manipulation was, thus, equivalent to the neutral condition in Lapata et al. (2003). These sentences were presented in isolation (no-context condition) to 42 participants. Another group of 19 participants were presented with the same sentences when these were preceded by a four-sentence paragraph (see Preceding Context in **Table 1**).

The contexts were designed to generate an event schema, specifically by establishing the agent's intention to perform a target activity – in this case *reading* – without mentioning the verb, neither before nor after the presentation of the target sentence. In both the no-context and context conditions, participants were instructed to provide a plausible verb to complete each sentence. See **Supplementary Material S2**, items 1–19, for the materials used in this task.

TABLE 1 | Sample materials employed in Experiments 1 and 2.

Condition/sentence type	Example
Preceding context	Lisa had been looking forward to the new Grisham novel ever since it came out. She had finally managed to set aside some time this weekend and made sure to make her home library nice and cozy. First thing Saturday morning, Lisa curled up on the sofa in her library with a blanket and a fresh cup of coffee. With everything in place, …
Indeterminate	Lisa began the book.
Biased foil/full-VP	Lisa began reading the book.
Non-biased foil	Lisa began writing the book.
Following neutral discourse	▲Suddenly, the doorbell rang. Lisa grunted, put down her coffee, and sluggishly made her way to the door. It was her neighbor John and he was out of peanut butter again. Looking through the cupboard, Lisa realized she was no better off. She told John that he was out of luck and suggested he try calling Mary, their mutual neighbor.▲

▲: *Locations of visual probe presentation in Experiment 2.*

Rating Task

Five additional sentences were added to the set of materials described above (see **Supplementary Material S2**, items 20–24) and all 24 items were used for ratings. The ratings task also had two parts, one with sentences in isolation (no-context condition) and one with each sentence preceded by a context. In the no-context condition, 24 indeterminate sentences were presented interspersed with 39 filler sentences for a total of 63 sentences. All 20 participants in the no-context condition saw the same set of 63 sentence materials.

In the contextual condition, indeterminate sentences were presented as the concluding clause of passages. For this condition, 50 participants were divided into five orthogonal lists that included an equal number of indeterminate sentences and full-VP sentences. The full-VP sentences included the event interpretation that was constrained by the context (see **Table 1**). These contextualized sentences were interspersed with 33 filler paragraphs.

In both the no-context and contextual conditions, participants were given a printed booklet and were instructed to rate the plausibility of each sentence on a 5-point scale, with higher ratings indicating greater plausibility (see **Supplementary Material S1**, for full instructions).

Results and Discussion

All context/no-context comparisons were analyzed by items using Welche's two-sample *t*-test in the "stats" package in R (R Core Team, 2014). In the fill-in-the-blank task, the mean proportion (*p*) of participants that generated the dominant verb increased by 22% when indeterminate sentences were embedded in a biasing context, $t(18) = 3.43$, $p < 0.001$, $d = 0.53$ (see **Table 2** for means and SDs). Thus, context constrained the range of event interpretations that participants assigned to indeterminate sentences. Context also improved plausibility ratings. Specifically, ratings were higher for indeterminate sentences with context compared to those without, $M_d = 0.66$, $t(23) = 4.20$, $p < 0.001$, $d = 0.80$. Finally, a dependent samples *t*-test showed that contextualized indeterminate sentences were rated similarly to full-VP sentences, $M_d = 0.08$, $t(23) = 0.34$, $p = 0.741$, $d = 0.12$.

Collectively, these results show that contexts contribute to the interpretation of indeterminate sentences by (a) constraining the range of event interpretations and (b) enhancing plausibility. It should be noted that our sentences, which used proper names, were less constraining than standard experimental sentences, which typically use semantically rich agentive nouns such as *author* or *student*. Similar to neutral condition of Lapata et al. (2003), we also aimed to determine the effect of verb-noun relations without the added semantic constrain of the agent. The data seem to suggest a preference for a particular meaning for the filing event – which in our case was more often associated with the *agentive* role.

TABLE 2 | Means (SD) for the fill-in-the-blank and plausibility ratings tasks in Experiment 1.

Task	No context	Context
Fill-in-the-blank[a]	0.59 (0.24)	0.81 (0.22)
Plausibility ratings		
Indeterminate	3.60 (1.00)	4.26 (0.76)
Full VP control	-	4.35 (0.89)

[a]*Proportion of participants that generated the dominant verb.*

The fill-in-the-blank convergence rates and plausibility ratings in the no-context condition were likely lower here than what has been observed in other studies (McElree et al., 2001; Traxler et al., 2002; Pickering et al., 2005). But our results nevertheless support the principle that extra sentential context is frequently required to enhance acceptability of indeterminate VPs, particularly when event information is insufficiently constrained by the intra-sentential semantics. We suggest, moreover, that this is often the case when indeterminate sentences are presented in isolation, even when the sentence has a semantically rich agent, as in *The carpenter began the table*. Although *carpenter* and *table* alone could yield a range of event representations (e.g., *building*, *demolishing*, *sanding*, *varnishing*, *repairing*, *renovating*, *restoring*, *measuring*, *moving*, etc.), such sentences likely breed uncertainty about the specific events that they convey when no information is available in the utterance context to constrain interpretations. The possibility that these sentences remain indeterminate can be further demonstrated by data of Lapata et al.: even with a typically *agentive* subject noun (e.g., *The author began…the book*) subjects complete sentences with verbs that are *not* in line with the agentive role (supposedly *writing*) in 47% of the cases. A similar pattern is obtained when the subject noun "favors" a *telic* interpretation (e.g., *The student began…the book*): although half of the sentence completions are in line with the telic role (supposedly *reading*), in 51% of the cases they are not.

The results of Experiment 1 – in conjunction with other studies reporting poor comprehension metrics of isolated sentences – underscore a crucial gap in the indeterminate sentence processing literature: event-enriching inferences have never been empirically demonstrated. To date, psycholinguistic experiments on indeterminate sentences have focused almost exclusively on time-course-of-processing paradigms using decontextualized sentences (but see de Almeida, 2004 and Traxler et al., 2005). These experiments typically show that indeterminate sentences produce online processing delays relative to a variety of control sentences. But while there is ample evidence that decontextualized indeterminate sentences are costly to process, we know little about the interpretations that these sentences generate. In the absence of direct empirical evidence that indeterminate sentences trigger event-enriching inferences, it is prudent to assume the null hypothesis – that the initial representations that individuals assign to these sentences are as indeterminate as the input. Experiment 1 suggested that, without context, it is unlikely that individuals consistently assign a specific event meaning to indeterminate sentences. The goal of Experiment 2 is to investigate (a) whether or not the proposition that listeners build of indeterminate sentences is indeed enriched over time, given sufficient contextual support, and (b) what is held in memory about the original indeterminate sentence.

EXPERIMENT 2

Numerous experiments investigating long-term retention of linguistic stimuli have demonstrated that delayed recognition reflects the meaning that individuals assign to utterances. For instance, a variety of now classical studies on false recognition for sentences (Sachs, 1967, 1974; Bransford and Franks, 1971; Bransford et al., 1972; Johnson et al., 1973; Brewer, 1977) have shown that while individuals quickly forget the *verbatim* form of linguistic expressions, the underlying propositional content or "gist" is retained.

This phenomenon is observed during recognition tasks when individuals erroneously recognize an expression that is synonymous with a presented sentence, albeit structurally distinct. For instance, Sachs (1967, 1974) demonstrated that, when they hear or read sentences like (6a) embedded in long contexts, participants frequently falsely recognize a semantically unchanged foil sentence such as (6b) upon delayed testing (80 syllables or up to 27 s after original presentation). Critically, this error does not occur for foils like (6c) that convey a fundamental change in meaning. Sachs' findings thus illustrated that sentences are transferred to LTM primarily in their semantic or propositional codes, which produces false recognition for altered sentences that convey the same proposition.

6.

a. *Target*: the founding fathers considered owning slaves to be immoral.
b. *Semantically unchanged foil*: owning slaves was considered to be immoral by the founding fathers.
c. *Semantically changed foil*: the founding fathers did not consider owning slaves to be immoral.

Later studies illustrated that the encoded "semantics" of a sentence is not restricted to its denotational representation. Information generated from pragmatic inferences are also encoded in LTM (Fillenbaum, 1971, 1974; Johnson et al., 1973; Brewer, 1977; Chan and McDermott, 2006). Moreover, sentences compatible with inferences that participants draw are often recognized more frequently than originally presented sentences (Johnson et al., 1973; Brewer, 1977), suggesting that information generated inferentially might ultimately supplant the denotational meaning of a sentence in LTM. For instance, participants who listened to a story about a boy who "was pounding a nail" later misrecognized a sentence that described the boy "using the hammer" more frequently than they recognized the original sentence, which made no mention of the instrument the boy was using (Johnson et al., 1973).

False recognition of inferred meaning has been extensively studied. And, typically, misrecognized inferences are either strongly implied (Johnson et al., 1973; Brewer, 1977; Chan and McDermott, 2006) or entailed (Bransford and Franks, 1971; Bransford et al., 1972; Rinck et al., 2001; Jahn, 2004) by the presented material. To our knowledge, this phenomenon has not been investigated in the context of sentence indeterminacy. We reasoned that if individuals indeed enrich *Lisa began the book* by ascribing the sentence a specific event meaning inferred from the utterance context, then they should misrecognize a fully determinate foil sentence such as *Lisa began reading the book* following a delay.

We used the contextual passages employed in Experiment 1 to investigate pragmatic event enrichment. Consider again the context presented in **Table 1** when it includes the indeterminate sentence *Lisa began the book*. As illustrated in Experiment 1, the passage biases the interpretation that Lisa began reading the book, although there is no mention of the *reading* event. During a delayed testing period, we presented one of three probe sentences: the indeterminate sentence that was presented during acquisition, a biased foil sentence, or a non-biased foil sentence.

Whereas Experiment 1 showed that event-enriching inferences are not likely achieved without context, the present experiment investigates whether the broader context indeed generates inferences for the biased event leading to enriched interpretations. Specifically, following Sachs (1967, 1974), we hypothesized that upon delayed recognition testing – in the present case, 25 s downstream – participants would falsely recognize biased foils but correctly reject non-biased foils.

Our experiment also addressed a secondary question concerning the time-course of the enrichment process and the nature of the proposition held in memory – whether original or "gist." Do enriched interpretations supplant the denotational representation of a sentence during acquisition or only during retrieval? We propose that the decision difficulty associated with long-term recognition responses might shed light on this question. Specifically, the longer it takes to misrecognize the foil, the more likely it is that the denotational representation of the sentence lingers, competing with the inferred event propositional content computed from context. In contrast, relatively rapid misrecognition would suggest that the event inference merged with or even replaced the denotational meaning of the sentence during acquisition, and only a single, fully enriched representation of the sentence was encoded.

Method

Participants

Seventy-two Concordia University students participated in this study – none of them participated in Experiment 1. They were all native speakers of English, and were compensated with course credit for their participation in a 40-min experimental session. They all gave informed written consent. The experiment was approved by the Concordia University Human Research Ethics Committee.

Materials

The materials consisted of the same 24 passages of continuous discourse used in Experiment 1. Each passage was comprised of three sentences of biasing context followed by an indeterminate sentence, which in turn was followed by several sentences of neutral discourse (see **Table 1**). The indeterminate sentence always appeared as the second clause of the fourth sentence in the discourse.

For each of the 24 paragraphs, three recognition probe sentences were generated for testing, as seen in **Table 1**: the original indeterminate sentence, a foil sentence that was biased by the discourse context, and a foil sentence that was not biased by the discourse context. The verb for the biased foil represented the response generated by the greatest proportion of participants during the fill-in-the-blank task from Experiment 1.

A set of 24 filler passages, which did not conform to any of the experimental features described above, were also prepared and included in the set of materials. The filler passages were of a similar length to the experimental materials and were written in the same general style, also describing mundane events, but without including indeterminate sentences. All 24 experimental passages as well as the 24 filler passages were recorded by the same female, native speaker of English, using natural prosody.

Procedure

Participants were seated in front of an iMac computer running PsyScope X (Cohen et al., 1993) and were provided with noise-canceling headphones. The experimenter then read the following instructions, which were also displayed on the screen for the participants to read:

> In the following task, you will be listening to a series of short stories through the headphones. At some point during playback, the story will be stopped and you will be presented with a few words on the screen. Your job is to indicate whether these words were present (exactly as they appear on screen) in the passage that you just heard. The words may form a full sentence, or just part of a sentence, but as long as they match word-for-word with a part of the passage you just heard, then you are to indicate a YES response. Otherwise, indicate a NO response. Use the keyboard in front of you to register your responses. Press the GREEN button if your answer is YES, and the RED button if your answer is NO.

Responses were registered on a keyboard with the "/" key marked by a green sticker for "yes" responses, and the "Z" key marked by a red sticker for "no" responses. RTs were recorded for each response. Participants were also instructed to rate how confident they were that their responses were correct on a 7-point scale, with 1 representing a guessed response and 7 representing total certainty. In addition to the "/" and "Z" keys, only the 1–7 keys and the space bar (for initiating trials) were visible on the keyboard.

During each trial, one sentence was presented for recognition at one of two probe points: immediate and delayed. The immediate condition occurred 0 s after the oral presentation of the indeterminate sentence, and the delayed condition occurred following an additional 25 s of neutral discourse (see **Table 1**). Each passage ran roughly 40 s long with the two probe periods occurring roughly 15 and 40 s after trial onset. To mask these probe points, testing of filler sentences occurred roughly 5 and 30 s after trial onset. The filler items were also used to balance the ratio of novel to repeated probe sentences throughout the experiment. While two-thirds of our experimental probes were novel (i.e., the biased and non-biased foils) and one-third was repeated (i.e., the indeterminate probe), the inverse was the case for the filler trials.

The session consisted of three practice trials, followed by the 24 experimental and 48 filler trials presented in random order. The item frames were distributed within six orthogonal lists, each containing four unique items for each condition – 3 (probe) × 2 (delay). Thus, each participant heard all of the 24 passages only once, and provided an equal number of responses in each of the six conditions. There were 12 participants per list.

Data Analyses

Recognition accuracy was used as the criterion in a mixed-effects logistic regression model, which tested for main effects of delay and probe type and all first order interaction terms. We conducted a secondary analysis on RTs to contrast the decision difficulty associated with the three probe sentences at the delayed test point. Given that the three sentences have different lengths, reading demands differ from probe to probe. Thus, we computed a variable to isolate RTs associated with decision difficulty alone. Specifically, we subtracted from each observation in the delayed condition the mean RT of the corresponding sentence in the immediate condition, the latter of which included only correct responses [e.g., $RT_{biased/delay}$ – mean ($RT_{biased/immediate/correct}$)]. These RTs were used to assess objectively the degree of difficulty associated with long-term recognition for each sentence type.

Results and Discussion

One item was removed from the analysis due to a typographical error in the probe sentences that were presented for testing. Thus, all subsequent analyses were conducted with 23 items. **Figure 1** presents recognition accuracy and confidence ratings. Descriptive statistics for these data are presented in **Supplementary Material S3**.

A binomial logit mixed model was fitted to the data using the "lme4" package (Bates et al., 2013) in R (R Development Core Team, 2012), with participants and items included as random effects, and delay period and probe type included as fixed effects. The overall model was evaluated relative to a null model consisting of only random predictors, and was found to provide a better fit to the data, $\chi^2(5) = 483.75$, $p < 0.001$.[4] A summary of fixed effects is presented in **Table 3**.

Our analysis revealed that recognition accuracy diminished with delay. In particular, the odds of correct recognition in the immediate testing period were 20 times that of the delayed testing period. An effect of probe type was also observed. The odds of correct recognition for the non-biased foils were 7.91 times that of the indeterminate probes, but the odds were approximately equal ($OR = 1.15$) for biased and indeterminate probes.

[4]Value of p less than 0.000 were reported as $p < 0.001$. Otherwise, the exact value of p is reported.

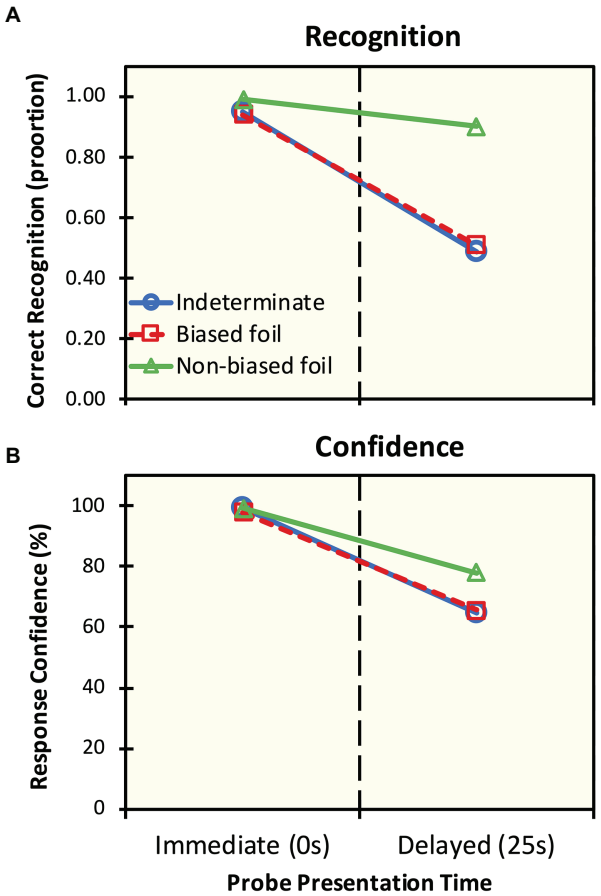

FIGURE 1 | Results from Experiment 2. Recognition accuracy **(A)** and response confidence ratings **(B)** for visual probe sentences shown at the offset of the original auditory presentation (immediate) and after 25 s (delayed) of intervening neutral discourse for indeterminate sentences (e.g., *Lisa began the book*), biased foils (*Lisa began reading the book*), and non-biased foils (*Lisa began writing the book*).

Interactions were not statistically significant, as identical and biased foils patterned together in both delay conditions – i.e., at least 94% accuracy upon immediate testing and chance performance at delayed testing – and the non-biased foils differed from the others in both testing periods.

These results show that participants are able to differentiate the indeterminate sentences from the foils immediately following the presentation of the indeterminate sentence in the discourse with at least 94% accuracy and 98% confidence for all probe types. However, during delayed testing, participants incorrectly recognized the biased foil, but not the non-biased alternative, and response confidence was as high as 4.99/7 (67%) for trials in which recognition of biased foils was false.

Our results extend Sachs' (1967, 1974) classical findings in significant ways. Sachs had shown that that "gist" representations serve as the primary source of sentence recognition in LTM once the verbatim trace has decayed. Recall that in Sachs' studies, with the exception of the semantically anomalous condition, all probe sentences – identical and foils – conveyed virtually the same proposition. Beyond Sachs' results, we show that the participants falsely recognize a probe that *does not* convey the proposition expressed by the original sentence presented in discourse: rather they accept sentences that are false, albeit contextually plausible. And participants do so with the same accuracy and confidence with which they accept the original sentence. Several other studies in recognition memory have pointed to the high acceptance and confidence associated with foils that are synonymous with the original stimulus sentence (e.g., Brewer and Sampaio, 2006). But, in the present case, *begin the book* and *begin reading the book* are not synonymous, for the former but not the latter is compatible with numerous events and both have different truth conditions.

In order to further investigate the processes underlying these false memories, we analyzed RTs associated with delayed recognition. Specifically, we measured the increase in RT from baseline by subtracting mean RTs for correct responses in the immediate condition from the RTs of the corresponding delayed condition. RTs that exceeded ±2.5 SDs from the mean – calculated separately for each condition – were replaced with the condition mean. This amounted to 2.6% of observations [24 indeterminate responses (16 at delay), 12 biased-foil responses (6 at delay), and 9 non-biased foil responses (5 at delay)].

A linear mixed-effects model was fitted to the RT data with probe type entered as a fixed effect and participants and items entered as random effects. The probe model was compared to a null model consisting of only random predictors and was found to provide a better fit to the data $\chi^2(2) = 47.74$, $p < 0.001$. **Table 4** presents the results of the fitted model, in which we observed a statistically significant estimate for both the biased and non-biased foils compared to indeterminate sentences. Specifically, participants were faster to respond to non-biased foils, $d = -0.54$, ~95% CI [−0.79, −0.28] and slower to respond to biased foils, $d = 0.36$, ~95% CI [0.13, 0.59].[5] Mean RT change is presented in **Figure 2**.

Biased and non-biased foils both differed from indeterminate sentences in terms of being novel, and thus both required a negative response. They also both differed from indeterminate probes in terms of having one extra word – e.g., *reading/writing*. But neither of these differences can explain our results, because we removed such extraneous variance by computing a delayed-minus-immediate difference score within each probe condition. Any processing difficulty associated with reading or rejecting the foils was thus internally controlled. The residual decision time, which differed in each condition, suggests that different processes governed responses for each of the three sentences.

Semantic coherence with the content of the discourse appears to be the primary source of decision difficulty. While the non-biased foils, which contradicted the event schemas described in the passages, were easiest for participants to classify,

[5]Effect size estimates (d) and approximate confidence intervals for d reported in text, as well as the condition means and SEs presented in **Figure 2** were all calculated with data averaged by participants, unlike the mixed-effects model in which all unique observations were included.

TABLE 3 | Logistic regression of recognition accuracy predicted by delay period and probe type in Experiment 2.

Predictor	Estimate	SE	*Z*	*p*	OR (e^E)	95% CI
Constant	2.94	0.27	10.84	<0.001	18.87	[11.10, 32.09]
Delay (delayed)	−2.96	0.29	−10.14	<0.001	0.05	[0.03, 0.09]
Probe (biased)	0.07	0.37	−0.18	0.857	0.94	[0.45, 1.93]
Probe (non-biased)	2.07	0.76	2.73	0.006	7.91	[1.79, 34.87]
Interaction terms						
delayed × biased	0.14	0.41	0.35	0.730	1.15	[0.52, 2.55]
delayed × non-biased	0.03	0.79	0.04	0.967	1.03	[0.22, 4.85]

TABLE 4 | Linear mixed-effects model of response time (RT) change by probe type in Experiment 2.

Predictor	Estimate	SE	*t*	95% CI (β)
Constant	1971	167.88	11.74	[1642, 2300]
Probe (biased)	458	168.49	2.72	[128, 788]
Probe (non-biased)	−712	168.36	−4.23	[−1042, −382]

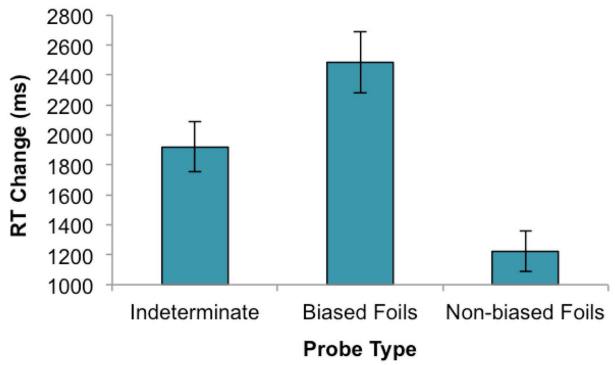

FIGURE 2 | Mean change in response times from immediate to delayed recognition in Experiment 2. Error bars represent the SE, calculated with data averaged by participants.

indeterminate, and biased-foils both generated additional decision delays. Presumably, participants who inferred that the biased event – say, *reading* – occurred in the discourse had to decide whether or not they acquired this information from the originally presented sentence. But why should this decision take longer for biased foils compared to indeterminate probes, especially since the two engendered equal levels of accuracy and confidence? We propose that there are additional interferences associated with biased foils due to denotational remnants of the acquired sentence. In other words: the denotational representation of the indeterminate sentence presented in discourse during acquisition interferes with the contextually-favored biased foil when this is evaluated in the delayed recognition condition.

Few studies have explored RTs to understand the processes that govern false memories (Jou et al., 2004; Coane et al., 2007; Jou, 2008), and this literature – which is restricted to the Deese/Roediger-McDermott (DRM) paradigm – is consistent with our findings. Participants are slower to respond to strongly biased foils compared to both non-biased foils and true items (Coane, et al., 2007; see Lopes and Garcia, 2014, for review). This pattern could be understood in terms of an activation/monitoring process (Roediger et al., 2001), whereby the false information is initially activated by the presented material during acquisition and is later erroneously reconstructed during retrieval due to a source monitoring failure. This dual-process account suggests that true and false information are encoded differently and produce phenomenologically distinct memory traces. However, the effects we obtained are more in line with Fuzzy-Trace Theory (FTT; Reyna et al., 2016, for a recent review). This theory also postulates a dual-representation of semantic information – verbatim (i.e., identical) and gist (which is meaning-preserving but not identical to the original stimulus). We assume that responding to a verbatim probe (*Lisa began the book*) requires the retrieval of the original proposition, a decision that takes longer in the delayed probe point due to (a) the intervening neutral context, but primarily (b) the inferences computed from contextual information. Yet, longer RTs for the biased foil (*Lisa began reading the book*) suggest that an enriched proposition is also available. Together with data on recognition and confidence, our RT data suggest that dual-representations – original and enriched propositions – are formed in the course of memory encoding over time. Beyond FTT, results compatible with the idea that sentences can yield true and false propositions were obtained by studies investigating "garden-path" sentences such as *While Susan wrote the letter fell off the table* (Christianson et al., 2001). When presented with these sentences, participants respond "yes" about half the time to questions that suggest a misparsing analysis, such as *Did Susan write the letter?* While it is possible that the subject of the matrix clause (*the letter*) is a plausible (although implicit) object of the subordinate clause, these results suggest that both propositions, [*Susan wrote the letter*] and [*The letter fell off the table*], linger in memory.

Our study was not designed to investigate these theoretical alternatives but rather the specific phenomenon of indeterminate sentence enrichment. Nonetheless, if the dual-representation account applies to the present study, the pattern of RTs we observed would suggest that the event inferences drawn from context and from the denotational representations of indeterminate sentences were encoded as distinct memory traces in LTM: a true proposition and a false, contextually-enriched one. It is theoretically possible that an event inference and a denotational interpretation would have been encoded as a single enriched proposition. But had this occurred, delayed recognition responses for biased foils should have been on par with the

indeterminate sentences. In the General Discussion, we elaborate on the implications of this finding for our understanding of the enrichment process.

GENERAL DISCUSSION

We conducted two experiments to investigate whether indeterminate sentences like *Lisa began the book* are enriched by assigning an interpretation that includes a specified event, such as *reading* in isolation and in context. Moreover, we were interested in the nature of the proposition that is obtained in memory for these sentences, both as participants first listen to the sentence and over time, as a function of context. In Experiment 1, participants were more likely to provide a dominant event interpretation when inferences were constrained by a broader discourse context. Similarly, participants provided higher plausibility ratings for indeterminate sentences in context, at levels comparable to event-specified sentences like *Lisa began reading the book*. Thus, Experiment 1 showed that, to the extent that indeterminate sentences generate event-enriching inferences, they are more likely to occur when interpretations are constrained by a broader discourse context. Lapata et al. (2003) have suggested that, in isolation, indeterminate sentences with neutral subjects (such as *Lisa*, in the present case) are filled primarily with a verb that is taken to be default for its object – such as *reading* for *began___the book*. In Experiment 1, we have shown that these "defaults" – if true – are ruled out by contextual demands. Compatible with classical studies demonstrating contextually-specific activation of properties in memory (e.g., Barsalou, 1983; McKoon and Ratcliff, 1988), our study shows that there are no defaults, but contextually appropriate enrichment.

Although contextual influence on sentence interpretation might be the norm, it should be noted that interpretations relying on some form of local semantic enrichment have long been assumed to occur, even for sentences in isolation. The GL framework (e.g., Pustejovsky, 1995, 2011), for instance, proposed a theory of coercion by which event meanings, such as *reading* and *writing*, are retrieved from the internal semantic representation of the nominal *book* and interpolated within the semantic representation (/logical form) of the sentence, as a kind of default. Thus, according to this view, a broader discourse context is not *necessary* for event-enriching interpretations to occur, for local semantic computations ought to determine the nature of the semantic filler, thus resolving the alleged mismatch between verb and complement.

To date, the vast majority of psycholinguistic and cognitive neuroscience experiments on indeterminate sentences have advanced theories of processing on the basis of sentences presented without context, de-emphasizing the role of context and even co-text in the event enrichment process, on the assumption that some sort of default meaning by necessity would ensue. Traxler et al. (2005, p. 5), for instance, proposed that "knowledge needed to enrich a complement is activated in an automatic and cost-free manner"… and that "the costs are due to additional operations needed to construct the appropriate event sense for the complement." The building of "the appropriate event sense" according to these authors is in line with Pustejovsky's interpolation proposal, whereby, say, *reading* is retrieved from *book*. This view, as we have argued elsewhere (de Almeida and Riven, 2012; de Almeida and Lepore, 2018) begs the question as to how the information that is "appropriate" is judged to be so. In order to enrich the propositional content that a sentence conveys (rather than to enrich a *sentence* qua linguistic object), there appear to be two alternatives: one is to rely on meaning decomposition, which in turn requires a criterion for determining what sort of content a concept carries (what is "analytic") from the content that is contingent on one's experience ("synthetic"). Meaning decomposition proposals cannot escape from the analytic-synthetic distinction and, thus far, a criterion for such distinction has not been set (see Fodor and Lepore, 2002; de Almeida and Riven, 2012; de Almeida, and Lepore, 2018; de Almeida and Antal, 2020; see also Quine, 1953). An alternative includes the semantic type-shifting of the complement NP, as discussed above. This view is committed to an ontology of semantic *types* for NPs, relying moreover on principles that adjust these types to fit the verbs' requirements. The assumption is that NPs carry information about their possible types, with semantic principles being informed about their modes of combination with their host verbs. Yet another alternative – one that we favor – is to assume that the content that enriches a proposition comes from contextual clues among other sources (expectations, beliefs, and conventions).

Our Experiment 1 suggests that event-enriching inferences are unlikely to occur reliably – that is, by *default* – in the absence of a strongly constraining discourse. Moreover, Experiment 1 showed that the plausibility of these sentences is tied to the availability of inferential constraints outside the indeterminate VP. We elaborate on the implications of these findings in conjunction with the results of Experiment 2 below.

Context and Enrichment

The results of Experiment 1 are particularly relevant because, to date, no studies have demonstrated that participants in fact generate event-enriching interpretations with or without context. The goal of Experiment 2 was to provide direct evidence of event inferences with contextualized sentences. After listening to short stories, which included indeterminate sentences like *Lisa began the book*, participants falsely recognized fully enriched foil sentences such as *Lisa began reading the book*. These foils included verbs that were implied by the discourse but never overtly mentioned. Although these event verbs were absent during acquisition, participants expressed 66% confidence that they indeed heard sentences like *Lisa began reading the book*, the same confidence they expressed for the original *Lisa began the book*. Thus, event inferences computed during acquisition left traces concerning the activity that was *began*, *finished*, or *continued* by the agent of the sentence. This is the first experiment to provide direct evidence that individuals build specified event representations, such as *reading*, *writing*, and *baking*, which they ultimately ascribe to such phrases as *started the book*, *continued the letter*, and *finished the cake*.

These results diverge from a previous experiment, which suggested that participants probably fail to generate event-specific interpretations for indeterminate sentences when presented in isolation (McElree et al., 2006b). Although our results are largely compatible with McElree et al. (as we discuss below), two critical differences in our methodology have enabled us to find evidence for the representations that have previously been undetected. Firstly, we employed a recognition paradigm (Sachs, 1967, 1974) that elicits the interpretations that individuals encode – the "gist" or, more technically, the *proposition* expressed by the sentence. Numerous studies of sentence and discourse memory (Sachs, 1967, 1974; Bransford and Franks, 1971; Bransford et al., 1972; Johnson et al., 1973; Brewer, 1977) show that what participants ultimately recognize reflects their understanding of the acquired information. Using a similar paradigm thus allowed us to assess directly participants' interpretations of our sentences. In contrast, McElree et al. (2006b) used acceptability ratings and processing times, which are good subjective and objective measures of processing fluency, but are not sufficiently sensitive to reveal the nature of the proposition that was encoded. More critically, our experiment embedded indeterminate sentences within strongly constraining discourse contexts. The participants of McElree et al. (2006b) saw sentences like *The carpenter began the table* decontextualized, and rated such sentences as less acceptable than controls. However, as our Experiment 1 showed, indeterminate sentences alone are likely insufficient to generate systematic event inferences. Although such sentences are likely to yield representations compatible with different events, individuals reading the sentence without a broader discourse context possibly fail to compute one specific interpretation, which is consistent with results of Lapata et al. (2003).

The results of Experiments 1 and 2 suggest that a detailed event schema must be established in the utterance context to produce sentence-enriching inferences. And in the absence of such contexts, indeterminate sentences breed uncertainty about the events that they are intended to describe. The idea that indeterminate sentences in isolation create uncertainty is supported by a recent fMRI study (de Almeida et al., 2016) showing the engagement of diverse brain areas – beyond those involved in the interpretation of fully determinate sentences – in particular the temporal and inferior frontal lobes bilaterally, the anterior cingulate cortex, and the thalamus. Although functional-neuroanatomical data should be seen with caution – for they alone cannot be taken to choose between theoretical or processing alternatives – they provide yet more evidence for how indeterminate sentences are (attempted) to be resolved: rather than resorting to local, default semantic operations, the resolution processes might involve generating inferences compatible with the original propositional content.

The Time Course of Enrichment

Having suggested that sentence-enriching inferences occur in the process of interpreting indeterminate sentences in context, a further issue for explication concerns the locus of enrichment. At what level of representation – or processing interface – does an inference enrich one's understanding of an indeterminate sentence? In the present study, delayed recognition RTs suggest the occurrence of interference between the denotational representation of the sentence and an enriched form, based on inferences computed from the discourse. Research on recognition for critical foils in DRM lists (Coane et al., 2007) suggests that RTs follow an activation/monitoring pattern (Roediger et al., 2001): participants are slower to respond to strongly biased words than they are to respond to non-biased words and true words (Coane et al., 2007). The additional interference associated with foils marks the presence of competition between originally presented and inferentially generated information. Compatible with FTT (Reyna et al., 2016), however, we take the false recognition of biased foils to be determined not by underlying associations but by the computation of a semantic alternative, a contextually-driven enriched proposition ("gist"), which is built over time and comes to compete with the original proposition.

Our participants' RTs followed this pattern. Recognition was slower for biased-foils than for indeterminate sentences and non-biased foils. Thus, it is likely that a proposition based on inferences about the event is ultimately encoded apart from the denotational representation of the sentence. An implication of this multiple trace account for indeterminate sentence processing is that enriched representations may not be built into the original propositional representation *per se*, but instead might occur beyond its composition. And interpretations of what a sentence means vs. what it implies are computed at distinct levels of representation – the former at the syntactic-semantic interface, and the latter at the level of thought, or pragmatics.

At first, our data can be seen as compatible with different accounts of indeterminate sentence processing, including those for which *interpolation* is a requisite for composing a semantic representation of the sentence. It may be that interpolation is automatically triggered by the input, and that this semantic operation works in tandem with context-driven inferential processes to produce a fully enriched sentence meaning. Specifically, it is possible that the context constraints the nature of events, providing information about a plausible predicate (e.g., *reading*) that serves to further enrich incoming sentences. When the sentence is parsed – and a mismatch is detected – the event suggested by the context becomes a predicate within the proposition encoded in memory. This process is compatible, then, with a full interpolating account of coercion, but one in which discourse information (viz., inferences based on implied events) provides the predicate for the local structural computations and yielding an enriched proposition. It is difficult, however, to determine how the content of a given context provides these directives – although it is possible to conceive of a general mechanism such as a "scoreboard" (Lewis, 1979) filled with common-ground information and presuppositions. And it is also difficult to constrain the boundaries of the context (see Cappelen and Lepore, 2005). However, while context *suggests*, it does not *determine* sentence enrichment. We note this because contextually supported foils

are rejected immediately and only later – over time, at the second probe point – they are accepted with relatively high levels of confidence. This, to us, suggests that the process of enrichment is primarily – if not uniquely – contextually-driven. Moreover, our data also suggest that the original proposition lingers, for sentences compatible with the original content of encoding are also accepted with the same level of confidence as the foils. If there is default coercion, the original, unenriched propositions should not linger in memory. Also, RTs to original and contextually-supported foils differ significantly, in the contrast between probe points, with greater costs for the enriched sentence at the late probe point. This suggests that, although seemingly confident, participants are reluctant to accept the enriched sentence, even when it is consistent with – and perhaps highly suggested by – the context. Therefore, while in principle compatible with the interpolation view, we see our data suggesting a process of enrichment that is primarily – if not uniquely – determined by contextual information over time.

Yet, another reason for casting doubt on the interpolation alternative relies on the theoretical morass that a commitment to analyticity entails (see, e.g., Quine, 1953; Fodor and Lepore, 2002; de Almeida and Antal, 2020). Simply put, there are no firm criteria for distinguishing between properties that are *constituents* of a concept (e.g., what goes into the "qualia structure") from those that are not. While this argument is not central to the interpretation of our results, it is a challenge to a proposal that relies on definitional or contingent properties of objects and events as contributing information to semantic computations.

What our experiments do not rule out is that sentences such as *Lisa began the book* might be, at first, subject to semantic algorithms that compute *semantic types*, thus triggering type-shifting operations akin to Partee (1986) and further extensions of type theory (e.g., Asher, 2015), without postulating lexical-semantic interpolation. These formal operations can very well precede interpretations (i.e., logical forms) that later become further enriched by context. Alternatively, it may be that pragmatic inferences are built on denotational representations of sentences derived from classical compositional mechanisms built out of unenriched syntactic analyses (de Almeida and Dwivedi, 2008; de Almeida and Riven, 2012). Either way, it is clear from our results that pragmatic inferences indeed occur in the service of enriching indeterminate sentences, playing a crucial role in building enriched propositions, thus in part accounting for effects obtained with behavioral and neuroimaging techniques.

CONCLUSION

Indeterminate sentences presented within strongly biasing discourse contexts trigger event inferences, which are encoded in LTM and later falsely recognized. Both the recognition of contextually biased sentences in the delayed probe point together with their longer RTs most likely suggest a competition between the original, *unenriched* sentence proposition and the proposition enriched with inferences computed from context. Our results are compatible with studies showing high rates of acceptance of false probes that are synonymous with original sentences (e.g., Sachs, 1967, 1974; Brewer and Sampaio, 2006) or that are their logical and pragmatic implications (Brewer, 1977). In the present study, however, the false-memory effects we obtained were even more surprising because they did not involve synonymous or entailed sentences. Taken together, the results from both experiments suggest that enrichment and consequent false recognition of indeterminate sentences can be attributed primarily to information generated by the context rather than to a default semantic interpolation.

More broadly, our study contributes to understanding the investigation of the division of labor between semantics and pragmatics, and their computations in the course of language comprehension. The cases we mentioned in the opening paragraph of the present article – viz., of "unarticulated constituents" (Perry, 1986; Recanati, 2004) – are examples of a pervasive approach to sentence meaning, namely one that takes a linguistically unmotivated form of silent meaning to contribute content to the representation of a sentence, beyond what it explicitly says. We suggest that sentences might hold their compositional meanings – without default interpolation – in isolation, with context being the source of enrichment in the form of pragmatic inferences computed over time.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Concordia University Human Research Ethics Committee. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

RdA devised the study. LR and RdA designed the experiments and wrote the manuscript. LR prepared materials, conducted the experiments, and analyzed the data. RdA is the main responsible for the final manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This study was supported by a grant from the Natural Sciences and Engineering Research Council of Canada (NSERC) to RdA, and fellowships from the Social Sciences and Humanities Research Council of Canada (SSHRC), and the Fonds de recherche société et culture (FQRSC) to LR.

ACKNOWLEDGMENTS

We thank Geoff O'Brien and Suzanne Purcell for their assistance in producing the materials, and Cailyn Antal for comments.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.616065/full#supplementary-material

REFERENCES

Antal, C., and de Almeida, R. G. (2020). "Indeterminate and enriched propositions linger in context: evidence from an eye-tracking false memory paradigm" Manuscript submitted for publication.

Asher, N. (2015). Types, meanings and coercions in lexical semantics. *Lingua* 157, 66–82. doi: 10.1016/j.lingua.2015.01.001

Baggio, G., Choma, T., van Lambalgen, M., and Hagoort, P. (2010). Coercion and compositionality. *J. Cogn. Neurosci.* 22, 2131–2140. doi: 10.1162/jocn.2009.21303

Barsalou, L. W. (1983). Context-independent and context-dependent information in concepts. *Mem. Cogn.* 10, 82–93. doi: 10.3758/bf03197629

Bates, D., Maechler, M., and Bolker, B. (2013). Lme4: linear mixed effects models using S4 classes. R package version 0.9999099–2. Available at: http://CRAN.R-project.org/package=lme4

Bransford, J. D., Barclay, J. R., and Franks, J. J. (1972). Sentence memory: a constructive versus interpretive approach. *Cogn. Psychol.* 3, 193–209. doi: 10.1016/0010-0285(72)90003-5

Bransford, J. D., and Franks, J. J. (1971). The abstraction of linguistic ideas. *Cogn. Psychol.* 2, 331–350. doi: 10.1016/0010-0285(71)90019-3

Brewer, W. F. (1977). Memory for the pragmatic implication of sentences. *Mem. Cogn.* 5, 673–678. doi: 10.3758/BF03197414

Brewer, W. F., and Sampaio, C. (2006). Processes leading to confidence and accuracy in sentence recognition: a metamemory approach. *Memory* 14, 540–552. doi: 10.1080/09658210600590302

Briscoe, T., Copestake, A., and Boguraev, B. (1990). "Enjoy the paper: lexical semantics via lexicology" in *Proceedings of 13th International Conference on Computational Linguistics*. Helsinki: Association of Computational Linguistics, 42–47.

Cappelen, H., and Lepore, E. (2005). *Insensitive semantics: A defense of semantic minimalism and speech act pluralism*. Oxford: Blackwell.

Chan, J. C. K., and McDermott, K. B. (2006). Remembering pragmatic inferences. *Appl. Cogn. Psychol.* 20, 633–639. doi: 10.1002/acp.1215

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

Coane, J. H., McBride, D. M., Raulerson, B. A., and Jordan, J. S. (2007). False memory in a short-term memory task. *Exp. Psychol.* 54, 62–70. doi: 10.1027/1618-3169.54.1.62

Cohen, J., MacWhinney, B., Flatt, M., and Provost, J. (1993). Psyscope: an interactive graphic system for designing and controlling experiments in the psychology laboratory using Macintosh computers. *Behav. Res. Methods Instrum. Comput.* 25, 257–271. doi: 10.3758/BF03204507

de Almeida, R. G. (2004). The effect of context on the processing of type-shifting verbs. *Brain Lang.* 90, 249–261. doi: 10.1016/S0093-934X(03)00438-3

de Almeida, R. G., and Antal, C. (2020). "How can semantics avoid the troubles with the analytic/synthetic distinction?" in *Concepts, frames and cascades in semantics, cognition and ontology*. eds. S. Löbner, T. Gamerschlag, T. Kalenscher, M. Schrenk and H. Zeevat (New York, NY: Springer), 1–19.

de Almeida, R. G., and Dwivedi, V. D. (2008). Coercion without lexical decomposition: type-shifting effects revisited. *Can. J. Linguist.* 53, 301–326. doi: 10.1017/S0008413100004503

de Almeida, R. G., and Lepore, E. (2018). "Semantics for a module" in *On concepts, modules, and language: Cognitive science at its core*. eds. R. G. de Almeida and R. Gleitman Lila (Oxford: Oxford University Press).

de Almeida, R. G., and Riven, L. (2012). "Indeterminacy and the source of coercion effects: minimal representations with pragmatic enrichment" in *Towards a biolinguistic understanding of grammar: Essays on interfaces*. ed. A. M. Di Sciullo (Philadelphia: John Benjamins), 277–303.

de Almeida, R. G., Riven, L., Manouilidou, C., Lungu, O., Dwivedi, V., Jarema, G., et al. (2016). The neuronal correlates of indeterminate sentence comprehension: an fMRI study. *Front. Hum. Neurosci.* 10:614. doi: 10.3389/fnhum.2016.00614

de Swart, H. E. (2011). "Mismatches and coercion" in *Semantics: An international handbook of natural language meaning*. eds. C. Maienborn, K. von Leusinger and P. Portner (Berlin: Mouton de Gruyter), 535–574.

Fillenbaum, S. (1971). On coping with ordered and unordered conjunctive sentences. *J. Exp. Psychol.* 87, 93–98.

Fillenbaum, S. (1974). Pragmatic normalization: further results for some conjunctive and disjunctive sentences. *J. Exp. Psychol.* 102, 574–578. doi: 10.1037/h0036092

Fodor, J. A. (1998). *Concepts: Where cognitive science went wrong*. Oxford: Oxford University Press.

Fodor, J. A., and Lepore, E. (eds.) (2002). "The emptiness of the lexicon: reflections on Pustejovsky" in *The compositionality papers*. Oxford: Oxford University Press, 89–119.

Frisson, S. (2009). Semantic underspecification in language processing. *Lang. Linguist. Compass* 3, 111–127. doi: 10.1111/j.1749-818X.2008.00104.x

Frisson, S., and McElree, B. (2008). Complement coercion is not modulated by competition: evidence from eye-movements. *J. Exp. Psychol. Learn. Mem. Cogn.* 34, 1–11. doi: 10.1037/0278-7393.34.1.1.

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

Jahn, G. (2004). Three turtles in danger: spontaneous construction of causally relevant spatial situation models. *J. Exp. Psychol. Learn. Mem. Cogn.* 30, 969–987. doi: 10.1037/0278-7393.30.5.969

Johnson, M. K., Bransford, J. D., and Solomon, S. K. (1973). Memory for tacit implications of sentences. *J. Exp. Psychol.* 98, 203–205. doi: 10.1037/h0034290

Jou, J. (2008). Recall latencies, confidence, and output positions of true and false memories: implications for recall and metamemory theories. *J. Mem. Lang.* 58, 1049–1064. doi: 10.1016/j.jml.2007.12.003

Jou, J., Matus, Y. E., Aldridge, J. W., Rogers, D. M., and Zimmerman, R. (2004). How similar is false recognition and veridical recognition objectively and subjectively? *Mem. Cogn.* 32, 824–840. doi: 10.3758/BF03195872

Katsika, A., Braze, D., Deo, A., and Piñango, M. M. (2012). Complement coercion: distinguishing between type-shifting and pragmatic inferencing. *Ment. Lex.* 7, 58–76. doi: 10.1075/ml.7.1.03kat

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

Kuperberg, G. R., Choi, A., Cohn, N., Paczynski, M., and Jackendoff, R. (2010). Electrophysiological correlates of complement coercion. *J. Cogn. Neurosci.* 22, 2685–2701. doi: 10.1162/jocn.2009.21333

Lapata, M., Keller, F., and Scheepers, C. (2003). Intra-sentential context effects on the interpretation of logical metonymy. *Cogn. Sci.* 27, 649–668. doi: 10.1207/s15516709cog2704_4

Lapata, M., and Lascarides, A. (2003). A probabilistic account of logical metonymy. *Comput. Linguist.* 29, 261–315. doi: 10.1162/089120103322145324

Lewis, D. (1979). Scorekeeping in a language game. *J. Philos. Log.* 8, 339–359.

Lopes, E. J., and Garcia, R. B. (2014). On the possibility of using reaction time to study false memories. *Psychol. Neurosci.* 7, 393–397. doi: 10.3922/j.psns.2014.047

McElree, B., Frisson, S., and Pickering, M. J. (2006a). Deferred interpretations: why starting Dickens is taxing but reading Dickens isn't. *Cogn. Sci.* 30, 181–192. doi: 10.1207/s15516709cog0000_49

McElree, B., Pylkkänen, L., Pickering, M. J., and Traxler, M. J. (2006b). A time course analysis of enriched composition. *Psychon. Bull. Rev.* 13, 53–59. doi: 10.3758/bf03193812

McElree, B., Traxler, M. J., Pickering, M. J., Seely, R. E., and Jackendoff, R. (2001). Reading time evidence for enriched composition. *Cognition* 78, B17–B25. doi: 10.1016/S0010-0277(00)00113-X

McKoon, G., and Ratcliff, R. (1992). Inference during reading. *Psychol. Rev.* 99, 440–466. doi: 10.1037/0033-295X.99.3.440

McKoon, G., and Ratcliff, R. (1988). Contextually relevant aspects of meaning. *J. Exp. Psychol. Learn. Mem. Cogn.* 14, 331–343. doi: 10.1037/0278-7393.14.2.331

O'Brien, E. J., and Cook, A. E. (2015). "Models of discourse comprehension" in *The Oxford handbook of reading*. eds. A. Pollatsek and R. Treiman (Oxford: Oxford University Press), 217–231.

Partee, B. (ed.) (1986). "Noun phrase interpretation and type-shifting principles" in *Compositionality in formal semantics*. Oxford: Blackwell, 203–230.

Perry, J. (1986). Thought without representation. *Proc. Aristot. Soc. Suppl. Vol.* 60, 137–151.

Pickering, M. J., McElree, B., and Traxler, M. J. (2005). The difficulty of coercion: a response to de Almeida. *Brain Lang.* 93, 1–9. doi: 10.1016/j.bandl.2004.07.005

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

Pustejovsky, J. (1995). *The generative lexicon*. Cambridge, MA: The MIT Press.

Pustejovsky, J. (2011). Coercion in a general theory of argument selection. *Linguistics* 39, 1401–1431. doi: 10.1515/LING.2011.039

Pustejovsky, J., and Batiukova, O. (2019). *The lexicon*. Cambridge: Cambridge University Press.

Pylkkänen, L. (2008). Mismatching meanings in brain and behavior. *Lang. Linguist. Compass* 2, 712–738. doi: 10.1111/j.1749-818X.2008.00073.x

Pylkkänen, L., and McElree, B. (2007). An MEG study of silent meaning. *J. Cogn. Neurosci.* 19, 1905–1921. doi: 10.1162/jocn.2007.19.11.1905

Quine, W. V. (ed.) (1953). "Two dogmas of empiricism" in *From a logical point of view. 2nd Edn*. (New York: Harper & Row), 1961.

R Core Team (2014). R: a language and environment for statistical computing. R Foundation for statistical computing, Vienna, Austria. Available at: http://www.R-project.org/

R Development Core Team (2012). R: a language and environment for statistical computing. R-Foundation for statistical computing, Vienna, Austria. Available at: http://www.R-project.org/

Recanati, F. (2004). *Literal meaning*. Cambridge: Cambridge University Press.

Reyna, V. F., Corbin, J. C., Weldon, R. B., and Brainerd, C. J. (2016). How fuzzy trace theory predicts true and false memories for words, sentences, and narratives. *J. Appl. Res. Mem. Cogn.* 5, 1–9. doi: 10.1016/j.jarmac.2015.12.003

Rinck, M., Hannel, A., and Becker, G. (2001). Using temporal information to construct, update, and retrieve situation models of narratives. *J. Exp. Psychol. Learn. Mem. Cogn.* 27, 67–80. doi: 10.1037/0278-7393.27.1.67

Roediger, H. L., Balota, D. A., and Watson, J. M. (2001). "Spreading activation and arousal of false memories" in *The nature of remembering: Essays in honor of Robert G. Crowder*. eds. R. G. Crowder and H. L. Roediger (Washington, DC: American Psychological Association), 95–115.

Sachs, J. (1967). Recognition memory for syntactic and semantic aspects of connected discourse. *Percept. Psychophys.* 2, 437–442. doi: 10.3758/BF03208784

Sachs, J. (1974). Memory in reading and listening to discourse. *Mem. Cogn.* 2, 95–100. doi: 10.3758/BF03197498

Sanford, A. J. (2002). Context, attention and depth of processing during interpretation. *Mind Lang.* 17, 188–206. doi: 10.1111/1468-0017.00195

Swinney, D. A. (1979). Lexical access during sentence comprehension: (Re) consideration of context effects. *J. Verb. Learn. Verb. Behav.* 18, 645–659. doi: 10.1016/S0022-5371(79)90355-4

Tabossi, P., and Zardon, F. (1993). Processing ambiguous words in context. *J. Mem. Lang.* 32, 359–372.

Traxler, M. J., McElree, B., Williams, R. S., and Pickering, M. J. (2005). Context effects in coercion: evidence from eye movements. *J. Mem. Lang.* 53, 1–25. doi: 10.1016/j.jml.2005.02.002

Traxler, M. J., Pickering, M. H., 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

Zarcone, A., Padó, S., and Lenci, A. (2014). Logical metonymy resolution in a words-as-cues framework: evidence from self-paced reading and probe recognition. *Cogn. Sci.* 38, 973–996. doi: 10.1111/cogs.12108

Conflict of Interest: 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 © 2021 Riven and de 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) and the copyright owner(s) 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: 06 April 2021
doi: 10.3389/fpsyg.2021.629724

Information Structure and Word Order Canonicity in the Comprehension of Spanish Texts: An Eye-Tracking Study

Carolina A. Gattei[1,2,3]*, *Luis A. París*[4] and *Diego E. Shalom*[1]

[1] Instituto de Física de Buenos Aires (CONICET), Universidad de Buenos Aires, Buenos Aires, Argentina, [2] Laboratorio de Neurociencia, Universidad Torcuato Di Tella, Buenos Aires, Argentina, [3] Facultad de Filosofía y Letras, Pontificia Universidad Católica Argentina, Buenos Aires, Argentina, [4] Grupo de Lingüística y Neurobiología del Lenguaje, Instituto de Ciencias Humanas Sociales y Ambientales (CONICET), Mendoza, Argentina

Word order alternation has been described as one of the most productive information structure markers and discourse organizers across languages. Psycholinguistic evidence has shown that word order is a crucial cue for argument interpretation. Previous studies about Spanish sentence comprehension have shown greater difficulty to parse sentences that present a word order that does not respect the order of participants of the verb's lexico-semantic structure, irrespective to whether the sentences follow the canonical word order of the language or not. This difficulty has been accounted as the cognitive cost related to the miscomputation of prominence status of the argument that precedes the verb. Nonetheless, the authors only analyzed the use of alternative word orders in isolated sentences, leaving aside the pragmatic motivation of word order alternation. By means of an eye-tracking task, the current study provides further evidence about the role of information structure for the comprehension of sentences with alternative word order and verb type, and sheds light on the interaction between syntax, semantics and pragmatics. We analyzed both "early" and "late" eye-movement measures as well as accuracy and response times to comprehension questions. Results showed an overall influence of information structure reflected in a modulation of late eye-movement measures as well as offline measures like total reading time and questions response time. However, effects related to the miscomputation of prominence status did not fade away when sentences were preceded by a context that led to non-canonical word order of constituents, showing that prominence computation is a core mechanism for argument interpretation, even in sentences preceded by context.

Keywords: information structure, word order, eye-tracking, text comprehension, prominence, psych verbs

OPEN ACCESS

Edited by:
Valentina Cuccio,
University of Messina, Italy

Reviewed by:
Masatoshi Koizumi,
Tohoku University, Japan
Tao Gong,
Educational Testing Service, United States

***Correspondence:**
Carolina A. Gattei
carolina.gattei@conicet.gov.ar

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

Received: *15 November 2020*
Accepted: *22 February 2021*
Published: *06 April 2021*

Citation:
Gattei CA, París LA and Shalom DE (2021) Information Structure and Word Order Canonicity in the Comprehension of Spanish Texts: An Eye-Tracking Study. Front. Psychol. 12:629724. doi: 10.3389/fpsyg.2021.629724

1. INTRODUCTION

Word order alternation is a frequent feature in many languages across the world. Several works have tried to explain the psycholinguistic principles that govern comprehension of alternative word orders. Based on theoretical accounts of word order alternation or "scrambling," many of these studies assume the existence of a particular canonical word order for each language (e.g., SVO for English, SOV for German, etc.), and alternative orders derived from it (Comrie, 1989).

Experimental evidence suggests that alternative word orders are more difficult to process than canonical ones, as reflected by longer reading times, response times and lower accuracy rates (Hyönä and Hujanen, 1997; Bader and Meng, 1999; Kamide and Mitchell, 1999). Studies about the role of word order for incremental comprehension have also shown that word order alternation is a relevant cue for lexico-semantic argument interpretation and posterior realization of syntax-to-semantics linking. In other words, incremental processing of word order features are useful to predict *"who does what to whom"* in a given event (see Bader and Bayer, 2006; Bornkessel and Schlesewsky, 2006, for two different reviews on this issue). For instance, a Spanish cloze task has shown that while the appearance of a nominative-marked argument in first position leads readers to expect an activity verb, the appearance of a dative-marked argument in first position leads them to expect an Object Experiencer psychological verb (heareafter ObjExp psych verb, Gattei et al., 2015b, Experiment 2). The violation of these expectations generates higher error rates and response times, longer reading times and amount of regressions to previous regions (Gattei et al., 2015a, 2017), and differential neural correlates (Gattei et al., 2015b), even in the canonical word order of the language. The interpretive function of word order has also been evaluated in a spectrum of languages with different degrees of complexity regarding morphological case marking, such as German (Bornkessel et al., 2003, 2005), Italian (Dröge et al., 2014), and Chinese (Wang et al., 2012) with very similar and robust results. Bornkessel and Schlesewsky (2006) suggest that word order—as well as case marking, animacy, and definiteness—are key linguistic features for the computation of argument prominence, which comprises the hierarchical relation among arguments in a sentence (Lamers and De Swart, 2012). The evidence suggests that the human sentence parser tends to interpret the first argument of a sentence as the most "Actor-like" possible according to the prominence status provided by those features. This proposal suggests that speakers tend to compute arguments prominence status by following a more-to-less prominent order, this is, following the stipulated order of arguments in the lexico-semantic structure of verbs. Hence, a Spanish animate, nominative-marked, definite argument in first position will most likely be the Actor of an activity event, and an animate, dative-marked, definite argument in first position will most likely be the Experiencer of a psych state.

However, an aspect that has not been taken into account by most studies that address word order alternation in sentence comprehension is that the appearance of non-canonical word order is not arbitrary but rather motivated by discursive factors like, for instance, if a referent has been previously introduced or if it is part of a referent mentioned before (Givón, 1984; Lambrecht, 1994; Birner and Ward, 1998). Along with prosody, word order is considered one of the key information structure markers across languages. The way in which given and new information is conveyed can modulate pragmatic interpretation by stipulating the status of constituents as discourse topic and focus. For instance, when unmarked, Spanish favors given information in the left-most position of the sentence, even when that means to change from canonical SVO word order to a non-canonical one (Zubizarreta, 1998) as it may be seen in (1):

(1) *¿Quién le gritó a Ana? [A Ana]*
Who CL-DAT yelled to Ana-DAT ? [To Ana-DAT]
le gritó María.
CL-DAT yelled María-NOM

"Who yelled at Ana? María yelled (at Ana=."

Typologically speaking, Spanish is considered to be a flexible language regarding both the possibility of alternating word order and the lack of constraints about the syntactic positions in which focus can potentially be assigned (Van Valin, 1999; Belloro, 2012). Hence, the same question posited in (1) may present a response in which new (focused) information takes place in first position, as shown in (2).

(2) *¿Quién le gritó a Ana? María le*
Who CL-DAT yelled to Ana-DAT ? María-NOM CL-DAT
gritó a Ana.
yelled to Ana-DAT

"Who yelled at Ana? María yelled at Ana."

In this example, the response to the question may be interpreted as narrow focus, in the sense that it was María, and not Juan, for instance, who yelled at Ana. Thus, the appearance of new information in first position may modify the way speakers interpret the response. From a psycholinguistic point of view, not many studies have addressed the role of word order for discourse on-going interpretation. In a study about processing of declarative sentences in Finnish with non-canonical information structure Kaiser and Trueswell (2004) argue that readers may need additional presuppositions in order to understand isolated sentences with a non-canonical word order. Hence, showing the right discourse setting for this type of sentences should facilitate comprehension. The authors showed that the presentation of a referent providing new information in first position entailed longer reading times irrespective of word order (SVO vs. OVS), and that overall, sentences with non-canonical word order (OVS) were more difficult to understand that sentences that followed the canonical word order of the language. This means that the presentation of a supportive discourse context partially alleviates the usual difficulty associated to a non-canonical construction.

In a series of Event-Related Potentials (ERPs) experiments in German that manipulated position of the referents and givenness, Schumacher and Hung (2012) showed that new inferred information in sentence-medial positions engender a Late Positivity when compared to new given information. This difference does not take place when the constituents are in sentence-initial position. The authors claim that "information presented in sentence-initial position is treated differently than information in other positions during both language processing," and the construction of discourse representation structure.

Burmester et al. (2014) also showed that topic-first order eases OVS sentence processing in German-speaking adults, as evidenced by an offline comprehensibility judgment task and a

late positivity effect at the ERP mean voltage when comparing sentences preceded by a neutral context and those preceded by a topicalized context.

The present study seeks to go a step forward and to evaluate how the pragmatic use of word order alternation interacts with its use as a cue for arguments prominence computation. In other words, if prominence is considered a hierarchy composed by other independent hierarchies (e.g., animacy features are independent from case marking and word order), and in a particular sentence these hierarchies may conflict with each other (e.g., the innanimate argument bares nominative case, Chow and Phillips, 2013), it is worth exploring when the language word order (SVO) is incongruent with the canonical word order stipulated by the lexico-semantic structure of the verb (SVO for activity verbs and OVS for ObjExp psych verbs) and that of the rhematic hierarchy ("given" referents precede "new" ones).

The paper is organized as follows: We first present a brief description of Spanish word order alternation, with special emphasis on Object Experiencer Psych Verbs (hereafter ObjExp psych verbs) and stipulate the hypotheses and predictions related to the processing of these sentences when embedded in context. We then present an eye-tracking study addressing these issues. Finally, we discuss the results of the current experiment under the light of previous findings.

1.1. Word Order Alternation in Spanish

Spanish is rather flexible in terms of word order, although it is argued to be a Subject-Verb-Object (SVO) language (Contreras, 1976; Suñer, 1982; Ocampo, 1995). Take for instance sentences in (3) and (4)

(3) *María le responde a Ana.*
MaríaNOM CL-DAT responds to Ana-DAT

"María responds to Ana"

(4) *A Ana le responde María.*
To Ana CL-DAT responds María-NOM

"María responds to Ana"

The verb in these examples expresses the same type of event, in which an Actor ("María") carries out an activity ("to respond") that affects another participant ("Ana"). The main difference between both sentences is that, apart from showing the canonical word order of the language, sentence (3) shows a canonical order of its arguments, with an Actor preceding the affected participant or "Undergoer" (Foley and Van Valin, 1984). Sentence (4), on the contrary, exhibit both a non-canonical word order and a non-canonical arguments order, with the Undergoer preceding the Actor.

The same morphological case marking is applied to arguments in sentences with ObjExp psych verbs, as shown in (5) and (6)

(5) *María le encanta a Ana.*
MaríaNOM CL-DAT loves to Ana-DAT

"Ana loves María"

(6) *A Ana le encanta María.*
To Ana CL-DAT loves María-NOM

"Ana loves María"

Altough sentence (6) carries a non-canonical word order, it reflects the canonical order of arguments established by its lexico-semantic structure, as exemplified in (7), in which the verb's left-most argument ("x") is associated to an Experiencer of a state predicate, and "y" is associated to the Theme that generates this state (Van Valin, 2005, p. 45).

(7) encantar'(x,y)

This subclass of psych verbs has become relevant for understanding how the sentence processor uses both syntactic and lexico-semantic information in order to predict the thematic structure of a particular event. In a series of studies run in this language, when presented with sentences like (4) and (6), readers found higher difficulty to integrate the verb and the second argument of the sentence in trials with activity verbs than in sentences with ObjExp psych verbs. The opposite pattern was found for subject-initial sentences, showing longer reading times (Gattei et al., 2015a) and higher amount of regressions (Gattei et al., 2017) to previous regions when the sentence included an ObjExp psych verb than an activity verb.

This pattern of results suggests that readers are not only guided by word order canonicity in order to interpret sentences, but that they use word order together with a semantic principle that stipulates that the first argument will take the most prominent status possible to form predictions about the type of thematic structure that the event will carry and assign a thematic role to the preverbal argument accordingly (Bornkessel et al., 2005; Wolff et al., 2007; Haupt et al., 2008). The appearance of a verb that required a correction of this assumption resulted in longer reading times in the regions that comprised the second argument of the sentence. Furthermore, when asked "who did/felt what for whom" after reading each sentence, accuracy rates were lower and response time longer when the sentences arguments did not reflect the canonical order of arguments of their lexico-semantic structure, showing that the effects of not respecting the order established by the lexico-semantic structure of the verb are so robust that can persist even once all the processes of linguistic integration have been completed.

1.2. Hypotheses and Predictions

By means of an eye-tracking reading task we aim at weighing the relative processing load imposed by the violation of two types of linguistic hierarchies related to word order alternation: the rhematic hierarchy—given referents precede new referents (Contreras, 1976)—and that related to arguments' prominence—"the Actor precedes the Undergoer" (Van Valin and LaPolla, 1997).

We propose to replicate the findings from the study of Kaiser and Trueswell (2004), but using two different verb types (i.e.,

activity verbs and ObjExp psych verbs), as in Gattei et al. (2015a, 2017).

Following the results of Kaiser and Trueswell (2004) and basing our hypotheses on the assumption that word order alternation is motivated by discursive factors (Givón, 1984; Lambrecht, 1994; Birner and Ward, 1998) we expect that overall, the appearance of an adequate context facilitates sentence comprehension. In the current study, context adequacy is provided by the pragmatic status of referents. This means that an adequate context will lead to a sentence with a "given" referent in first position and a "new" referent as second argument, giving rise to a canonical rhematic hierarchy. Conversely, an inadequate context will give rise to a "new" referent in first position and a non-canonical information structure.

We also expect that all effects related to the interaction between syntactic, semantic and pragmatic factors are reflected in late eye-movement measures, since they are assumed to reflect later parsing stages (see Clifton et al., 2007; Vasishth et al., 2013, for a review on this discussion).

In relation to the interaction between both prominence and rhematic hierarchies, we predict two possible outcomes:

1. Context adequacy causes possible effects of prominence miscomputation fade away. The rationale of this prediction is that thematic reanalysis effects found in previous studies could have been the result of making additional presuppositions related to the use of a non-canonical word order without any previous context. This hypothesis predicts a main effect of information structure, and a triple interaction between word order, verb type and information structure, with sentences with non-canonical word order showing higher processing demand when an unsupportive context is used than when preceded with a supportive context. When a supportive context is used, prominence miscomputation effects should disappear. This interaction should take place once the verbs are read and in subsequent regions.
2. Context adequacy plays a role at initial stages of sentence processing but does not make the effect of prominence miscomputation fade away. The rationale of this prediction is that the relation between syntax-to-semantics linking and word order involves a mechanism -semantic roles and syntactic functions- that belongs to the grammatical nucleus of any given language. Thus, the violation of the prominence hierarchy comprises the alteration of a core relationship in a sentence. On the contrary, the relationship between a non-canonical rhematic structure and non-canonical word order involves the manipulation of a more flexible system (i.e., Pragmatics), designed to adapt linguistic form to the dynamics of context. In other words, this hypothesis predicts greater difficulty for sentences with unsupportive context than for those with supportive one. This difficulty should be reflected at the initial regions of sentences (i.e., where the new referent takes place). The hypothesis also predicts higher processing demands for those sentences that do not respect the prominence hierarchy than for those that respect it (i.e., for SVO sentences with ObjExp psych verbs and for OVS sentences with Activity verbs) irrespective of whether they are preceded by a supportive or unsupportive context. Following Gattei et al. (2017), effects of prominence miscomputation should take place at late eye-movement measures at later regions of the sentence (i.e., verb region onward for reading measures, and at initial regions of the sentence and verb for regression measures).

2. MATERIALS AND METHODS

We designed a text reading task using the eye-tracking method in order to study the interaction between word order, verb type and information structure. This technique allows to register with great temporal precision what eyes do during naturalistic reading, and what strategies readers use in order to overcome cognitive difficulties that could arise from linguistic complexity (also see Just and Carpenter, 1980; Just et al., 1982, for a discussion on the advantages of this paradigm).

2.1. Participants

Seventy-two native Spanish speakers (47 female, age range 18–54 years old; $M = 22.6$, $SE = 0.74$) participated in this study. All participants had normal or corrected-to-normal vision and had no history of prior neurological disease, drug or alcohol abuse, psychiatric disorders, developmental speech/language disorders, or learning disabilities. All of them provided written consent prior to the study. Sixty-nine of the participants entered the final data analysis, the remaining three having been excluded on the basis of equipment-related artifacts and/or insufficient accuracy in the comprehension task (an error rate higher than 40% in the critical conditions). All participants were compensated with 150 Argentinian Pesos (approximately US$ 9 at that time) after finishing the experiment session.

2.2. Materials

A total of 384 texts were built following the studies of Gattei et al. (2017) and Kaiser and Trueswell (2004). The texts consisted of three sentences [hereafter S_1 refers to the first sentence, S_2 refers to the second sentence and S_3, to the third sentence of the text, see example (8)]. S_1 introduced the first referent (R_1: Richard) and the situation in which s/he was. S_2 introduced the second referent (R_2: Mary/Ana) and stated that this person was performing an action with a person whose name has not been mentioned (R_3: Ana/Mary). S_3 comprised the target sentence, which described that R_1 (Richard) saw or heard that one of the two referents introduced in S_2 did or felt something for the other person. Sentences were built in such way that R_1 always had a different gender than R_2 and R_3. This was done in order to avoid possible ambiguity in the use of pronouns ("he" or "she") in the text, so that it was always clear for the reader that it referred to R_1. In other words, if R_1 was feminine, then R_2 and R_3 were masculine proper names and vice versa.

(8) *Ricardo estaba en la sala corrigiendo unos exámenes.*
Richard was in the room grading some exams.
Él vio que ***María/Ana*** *estaba en el pasillo*
He saw that Mary/Ana was in the corridor
charlando por teléfono. Poco después, él oyó que
chatting by telephone. Little later, he heard that
María *le respondía a Ana...*
Mary CL-DAT responded to Ana...

"Richard was in the room grading exams. He saw that Mary/Ana was in the corridor chatting on the phone. Little later, he heard that Mary responded to Ana..."

In S_3, 24 ObjExp psych verbs and 24 activity verbs with dative-marked objects were used. Both verb groups were matched according to length (ObjExp: $M = 6.8$, $SE = 0.31$; Act: $M = 6.3$, $SE = 0.25$) and log-transformed frequency (ObjExp: $M = 4.32$, $SE = 0.17$; Act: $M = 4.47$, $SE = 0.11$) according to the LEXESP database (Davis and Perea, 2005). An independent samples *t*-test revealed that there were no significant differences between groups [Length: $t_{(46)} = -1.35$, $p > 0.05$; log Frequency: $t_{(46)} = 0.71$, $p > 0.05$].

Verbs from S_3 were framed between a Noun Phrase (NP) and a Prepositional Phrase (PP) that consisted of 48 pairs of proper names matched in length and counterbalanced in gender (half masculine and half feminine). Target sentences could also follow the Subject-Verb-Object (SVO) order or the Object-Verb-Subject (OVS) word order. In this way, we tested the role of constituents order for these types of sentences. Finally, information structure of S_3 was also manipulated. In four of the target sentences, the referent that appeared in first position in S_3 had already been mentioned in S_2, while in the other four, the referent that appeared in first position in S_3 had not been previously mentioned by its proper name. This means that the configuration of half of the sentences' information structure comprised a given referent in first position while in the other half, a new referent was provided at sentence-initial position.

The 384 total sentences were divided into eight lists of 48 sentences each (six per condition) so that participants would see each verb twice, each time in a sentence with different word order and framed by two different pairs of proper names and a different context.

In order to avoid *wrap-up* effects (Just and Carpenter, 1980), additional phrases were added at the end of S_3 so that regions of interest did not coincide with the last word of the text. These phrases were semantically neutral so that they would not facilitate the interpretation of S_3's argument structure. In order to facilitate the posterior statistical analysis, both the syntactic structure and length of the first two sentences of the text were kept constant among the 48 sets, with a length range between 35 and 52 characters in S_1 ($M = 42.5$ characters), and 29–40 characters in S_2 ($M = 35.42$ characters). Length of S_3 would only vary according to the length of the additional phrase used in order to avoid "wrap-up" effects, with a length range of 75–88 characters ($M = 82.33$ characters).

In addition, a set of three practice items and 72 filler texts that were unrelated to the purposes of the study were used. The latter texts included sentences with different syntactic complexity and length to the target texts, and referred to diverse semantic topics, so that participants could not realize what the main purpose of the study was.

Finally, 123 questions were designed in order to test the comprehension of each practice, critical and filler item. Questions for the critical items were formulated in two ways: In order to respond 32 of the questions, participants had to retrieve the argument structure of the target sentence (S_3) and participants had to respond whether one of the referents did / felt what for the other referent, while in the remaining 16 texts the question tested the comprehension of one of the two previous sentences (S_1 and S_2). The rationale of doing this was to assure that participants would read the context previous to the target sentence. Half of the questions were responded affirmatively and half of them were responded negatively. Half of the questions that referred to S_3 asked about the subject constituent and half of them referred to the object constituent. Participants had to choose the correct answer by clicking on it with the mouse. Position of the correct answer was half of the times on the right side of the screen and was randomly assigned between trials for each participant. **Table 1** shows an example of one of the 48 sets of 8 texts used in the current experiment. A complete list of the experiment materials may be found at Appendix A of the Supplemental Material, available at https://osf.io/kp4dn/.

2.3. Equipment

Similarly to (Gattei et al., 2017), participants were seated in front of a 19-inch screen (Samsung SyncMaster 997 MB, 1024 × 768 pixels resolution, 100 Hz refresh rate) at a viewing distance of 65 cm. Head movements were prevented with a chinrest aligned with the center of the screen. Gaze locations of both eyes during reading was recorded with an EyeLink 1000 eye-tracker (SR Research Ltd.) at a sampling rate of 1 kHz. As given by the manufacturer, nominal average accuracy was 0.5 ° and space resolution was 0.01° root mean square (RMS). A standard 13-point grid for both eyes was used to calibrate participant's gaze. All recordings and calibration were binocular but only left eye data were used for the analysis.

All eye movements were labeled as fixations, saccades and blinks by the eye-tracker software using the default thresholds for Cognitive experiments (30°/s for velocity, 8,000°/s for acceleration, and 0.1° for motion, Cornelissen et al., 2002). Stimuli presentation was developed using Matlab (http://www.mathworks.com/, Massachusetts, United States) and Psychophysics Toolbox Version 3.

2.4. Procedure

All texts were displayed on five lines, the critical sentence being displayed on the fourth line. Neither the first nor the last word of each line displayed any of the main regions of interest from the critical sentence nor any of the referents from S_2.

In sum, the design of the text was such that: (i) the critical sentence did not exceed one line of the text; (ii) the line of the critical sentence was always the same across trials (line four); and (iii) the line of the critical sentence never started or ended with a critical word.

TABLE 1 | Critical sentences used in the current eye-tracking study.

Condition	Context	Critical sentence	Question
(a) ObjExp SVO G-N	Ricardo estaba en la sala corrigiendo unos exámenes. Él vio que María estaba en el pasillo charlando por teléfono. Poco después él oyó que...	María le encantaba a Ana aunque no estuvieran de acuerdo. *Ana loves Mary although they wouldn't agree.*	¿Es María quien le encantaba a alguien?
(b) ObjExp OVS G-N		A María le encantaba Ana aunque no estuvieran de acuerdo. *Ana loved Mary although they wouldn't agree.*	Is it Mary who is loved by someone?
(c) Act SVO G-N	*Richard was at the room grading exams. He saw María was at the corridor talking on the phone. Later he heard that...*	María le respondía a Ana aunque no estuvieran de acuerdo. *Mary responded to Ana although they wouldn't agree.*	¿Es María quien le respondía a alguien?
(d) Act OVS N-G		A María le respondía Ana aunque no estuvieran de acuerdo. *Ana responded to Mary although they wouldn't agree*	Is it Mary who responded to someone?
(d) ObjExp SVO N-G	Ricardo estaba en la sala corrigiendo unos exámenes. Él vio que Ana estaba en el pasillo charlando por teléfono. Poco después él oyó que...	María le encantaba a Ana aunque no estuvieran de acuerdo. Ana loves Mary although they wouldn't agree.	¿Es María quien le encantaba a alguien?
(e) ObjExp OVS N-G		A María le encantaba Ana aunque no estuvieran de acuerdo. *Ana loved Mary although they wouldn't agree.*	Is it Mary who is loved by someone?
(f) Act SVO N-G	*Richard was at the room grading exams. He saw Ana was at the corridor talking on the phone. Later he heard that...*	María le respondía a Ana aunque no estuvieran de acuerdo. *Mary responded to Ana although they wouldn't agree.*	¿Es María quien le respondía a alguien?
(g) Act OVS N-G		A María le respondía Ana aunque no estuvieran de acuerdo. *Ana responded to Mary although they wouldn't agree*	Is it Mary who responded to someone?

ObjExp, Object Experiencer Psych Verbs; Act, Activity Verbs; SVO, Subject-Verb-Object, OVS, Object-Verb-Subject; G-N, Given-New; N-G, New-Given.

Sentences were presented in Courier New Bold font. At a distance of 65 cm, each letter subtended 0.44° of visual angle laterally. Subjects were instructed to read the texts at their own rate. No instructions were given to suppress eye blinks.

Before the eye-tracking experiment began, they had a practice session of three texts. At the beginning of each trial, a dot appeared at the top left edge of the screen and after participants fixated on this dot, the text appeared. The first letter of the text was located at the position of the dot. Participants were instructed to look at a second dot at the bottom right corner of the screen to indicate they had finished reading. The total reading time of each trial was measured starting from when participants triggered the appearance of the text by fixating on the left dot until they fixated on the bottom right dot and the text disappeared. Comprehension questions appeared after every text. Participants responded by mouse-clicking on one of two possible answers ("Yes" or "No") displayed horizontally. Response time was measured starting from the appearance of the question until participants clicked on one of the possible responses. A calibration procedure was performed at the beginning of the eye-tracking experiment. Experimental sessions lasted approximately 45 min.

2.5. Data Analysis

Eye movement data from the 69 participants was screened for blinks and track losses. Fixations shorter than 50 ms and longer than 1,000 ms were removed from the analysis. After this screening process, fixations were assigned to their respective word and line. Boundaries between words (x axis) were set by splitting the space between two words in half. Boundaries between lines (y axis) were set by splitting the space between two lines in half. Upper and lower boundaries of the first and last lines were calculated so that they were symmetrical with the lower and upper boundaries of these lines, respectively. Fixations that fell outside the boundaries of the text were eliminated whenever participants continued reading after fixating outside the text area.

Visual inspection was carried out for each trial by providing a number to each fixation and a line that linked consecutive fixations. With this representation it could be easily established whether participants were reading the whole text. Trials in which participants skipped sentences from the context or the critical sentence were erased. Whenever there was a vertical misalignment between the fixations and the lines they belonged to, manual correction of fixations was performed by taking into account the trajectory of the reading path and realigning the

fixation to the correct line. Visual inspection and subsequent correction resulted in the removal of 5,293 fixations (0.79% of the data) and realignment of 17537 fixations (2.62%). Besides, 44 trials were removed due to track loss, the appearance of a random reading pattern, or incomplete text reading. This comprises 0.5% of the total data.

Eye-tracking measures were computed using em2 package for R language for statistical computing (Logacev and Vasishth, 2013, version 3.0.2).

For the purpose of analysis, we divided the sentences into ten regions that consisted of the first ten words of each sentence, as shown in **Table 2**. Note that in order to facilitate statistical analysis and visual presentation of the results, we aligned the critical regions that comprised the proper names (regions 2 and 6), the clitic (region 3) and the verb (region 4). The region of the preposition has been labeled as (5) in subject-initial sentences, and (1) in object-initial sentences. The regions "PP1," "PP2," "PP3," and "PP4" correspond to the first to the fourth word of the prepositional phrase following the second noun phrase.

For each fixated word, we computed the following measures: (1) *First Fixation Duration* (FFD; the duration of the first fixation on the word); (2) *First Pass Reading Time* (FPRT; the sum of all fixation durations on the word before any other word was fixated); (3) *Regression Path Duration* (RPD; also known as *go-past time*, it is the sum of all first-pass fixation durations on the word and all preceding words in the time period between the first fixation on the word up to the point where the reader leaves the critical region with a progressive saccade; (4) *Right-Bounded Regression Count* (RBRC; the number of regressions from the word before any word further to the right has been fixated); (5) *Total Fixation Time* (TFT; the sum of all fixations durations on a word); and (6) *Total Incoming Regressions* (TIR; the number of regressions to a specific word). Measures 1–2 are typically considered early measures, whereas measures 3–6 are considered late measures (Clifton et al., 2007; Vasishth et al., 2013).

Data analysis was conducted in the R programming environment (R Core Team, 2013). For measures comprising reading or response time (i.e., Comprehension Task Response Time, FFD, FPRT, RPD, and TFT) a linear mixed-effects model was fit to the data using the package lme4 (Pinheiro and Bates, 2000; Bates et al., 2014). For the accuracy measure, the data was fit to a generalized linear mixed-effects model with a binomial function, which is adequate for analyzing data measured on a dichotomous scale, namely "Correct" and "Incorrect" response. Count data (RBRC and TIR), on the other hand, was analyzed with a generalized mixed-effects model with Poisson link function, which is appropriate for counts of events in a fixed time window (Baayen, 2008, p. 322).

For the regression models, Verb Type, Word Order and Information Structure were considered fixed effects and Subject, and Item were fit as random effects. Log Frequency and inverse length of each word were included as control factors in every region except for regions 1 and 5 (preposition "a") and region 3 (clitic). These two variables may explain a significant part of the variability in reading times and amount of fixations on these regions (Just and Carpenter, 1980; Rayner and Well, 1996; Kliegl et al., 2004). As for collinearity between both factors, model comparison among models that included one, the other or both were significantly different. AIC values indicated that models where both factors were included were significantly better than the other two. In consequence, the two of them were included.

A maximal random-effects structure was included in both LMMs and GLMMs whenever it was possible, as linear mixed-effects models that do not consider random intercepts and slopes involve the risk of Type I error inflation (Barr et al., 2013). When models either did not converge or the correlation between variance components could not be estimated, the random effects structure was simplified by removing the correlations. For large samples like the ones collected in this study, 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 that involved reading time were scaled and centered.

Finally, we used an orthogonal contrast coding to test the interactions among verb type, word order and information structure at the pertinent regions. For the verb type contrast, sentences with activity verbs were coded as −1 and sentences with ObjExp psych verbs were coded as 1. For the word order contrast, SVO sentences were coded as −1 and OVS sentences were coded as 1. Finally, for the information structure contrast, sentences with a new referent in first position were coded as −1 and sentences with a given referent in first position were coded as 1.

3. RESULTS

3.1. Comprehension Task

3.1.1. Total Reading Time

Figure 1A shows the average total reading time for the critical texts used in the current eye-tracking experiment. The statistical

TABLE 2 | Regions of interest used for the statistical analysis of the current eye-tracking experiment according to Word Order.

	1	2	3	4	5	6	7	8	9	10
SVO		María	le	respondía|encantaba	a	Ana	PP1	PP2	PP3	PP4
		Mary	*clitic*$_{[DAT]}$	*responded|loved*	*to*	*Ana*$_{[DAT]}$	*PP1*	*PP2*	*PP3*	*PP4*
OVS	A	María	le	respondía|encantaba		Ana	PP1	PP2	*PP3*	*PP4*
	To	*Mary*$_{[DAT]}$	*clitic*$_{[DAT]}$	*responded|loved*		*Ana*	*PP1*	*PP2*	*PP3*	*PP4*

SVO, Subject-Verb-Object; OVS, Object-Verb-Subject; PP1, First Word of the Prepositional Phrase; PP2, Second Word of the Prepositional Phrase; PP3, Third Word of the Prepositional Phrase; PP4, Fourth Word of the Prepositional Phrase.

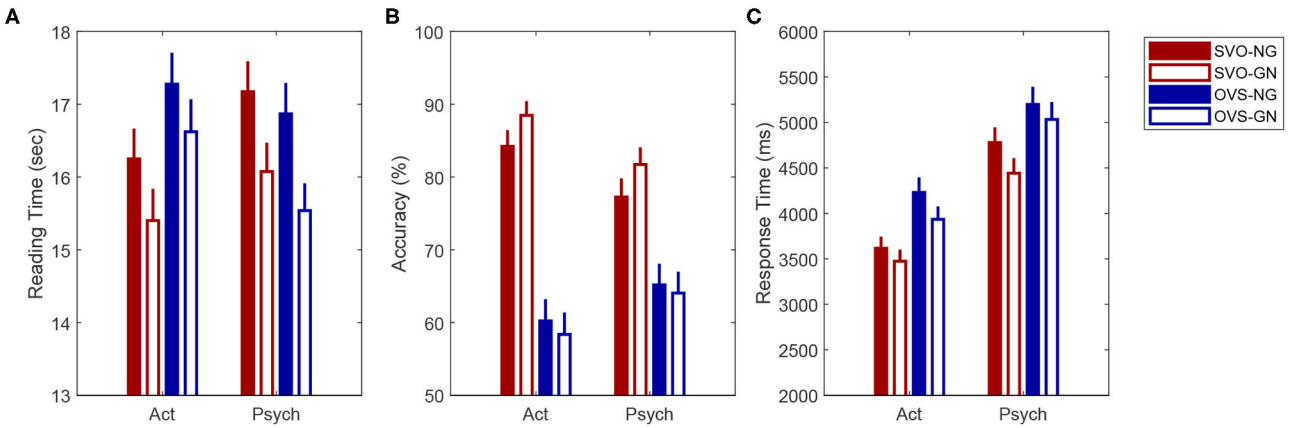

FIGURE 1 | Mean total reading time for the critical texts **(A)**, percentage of accurate answers **(B)**, mean response times for the comprehension question **(C)**, in the current eye-tracking study according to verb type (ObjExp psych verb vs. Act), word order (SVO vs. OVS) and information structure (given-new vs. new-given). Error bars correspond to Standard Error of the Mean. ObjExp psych verb, Object Experiencer Psych Verb; Act, Activity Verb; SVO, Subject-Verb-Object; OVS, Object-Verb-Subject; GN, given-new; NG, new-given.

analysis revealed an interaction between verb type and word order; $\beta = 0.022$, $SE = 0.007$, $t = 3.193$, $p < 0.01$. Resolving this interaction showed that participants spent significantly longer time reading sentences with SVO word order when they included an ObjExp psych verb than when they included an activity verb; $\beta = 0.071$, $SE = 0.020$, $z = 3.624$, $p < 0.01$. Although this difference was not significant among OVS conditions, sentences with activity verbs were read slower when they followed the OVS word order than when they followed the SVO order; $\beta = 0.077$, $SE = 0.020$, $z = 3.955$, $p < 0.001$. Information structure also affected texts' reading times significantly. Participants took longer time to read texts in which the Information Structure of the critical sentence included a new referent in first position ($M = 16{,}955$ ms, $SE = 210$ ms) than when it included a given referent in first position ($M = 16{,}015$ ms, $SE = 207$ ms; $\beta = 0.032$, $SE = 0.007$, $z = 4.552$, $p < 0.001$).

3.1.2. Question Accuracy

Mean accuracy for all comprehension questions was 86.27% ($SE = 0.38\%$). This indicates that participants were paying attention to the content of the texts. Mean accuracy of critical questions was 77.45% ($SE = 0.74\%$). **Figure 1B** shows mean accuracy according to condition. Differences in accuracy according to verb type and word order were analyzed with a generalized linear mixed-effects model. The analysis revealed a significant interaction between verb type and word order; $\beta = -0.198$, $SE = 0.055$, $z = -3.614$, $p < 0.001$. Resolving this interaction revealed that accuracy was significantly higher for questions about sentences with Activity verbs and SVO word order than for the other three conditions (ActSVO - ActOVS; $\beta = 1.640$, $SE = 0.164$, $z = 9.981$, $p < 0.001$; ActSVO - ObjExpSVO; $\beta = 0.532$, $SE = 0.172$, $z = 3.089$, $p = 0.011$; ActSVO - ObjExpOVS; $\beta = 1.385$, $SE = 0.165$, $z = 8.397$, $p < 0.001$). A significant effect of word order was also found. On average, participants responded more accurately after reading texts with sentences in SVO order ($M = 84.51\%$, $SE = 0.9$) than texts with sentences in OVS order ($M = 70.64\%$, $SE = 1.13$; $\beta = -0.626$, $SE = 0.056$, $z = -11.27$, $p < 0.001$).

3.1.3. Response Time

Figure 1C shows mean response time (RT) according to condition. Analysis of differences in RT between verb type, word order and information structure revealed main effects of the three factors. On average, response time was significantly longer for questions about texts that included ObjExp psych verbs ($M = 4{,}629$ ms; $SE = 70$) than for questions about texts with activity verbs ($M = 3{,}889$ ms, $SE = 58$; $\beta = -0.118$, $SE = 0.013$, $t = -9.066$, $p < 0.001$). Participants also took longer time to respond to questions about texts that included sentences in OVS order ($M = 4{,}396$ ms; $SE = 67$) than when they included sentences in SVO order ($M = 4{,}121$ ms, $SE = 61$; $\beta = 0.054$, $SE = 0.011$, $t = 5.031$, $p < 0.001$). Finally, questions about texts that included critical sentences with non-canonical information structure were responded significantly slower ($M = 4{,}383$ ms; $SE = 67$) than questions about texts that included sentences with a canonical rhematic hierarchy ($M = 4135$ ms, $SE = 62$; $\beta = 0.025$, $SE = 0.011$, $t = 2.270$, $p < 0.05$). Interactions among the three factors were not significant.

3.2. Eyetracking Measures

Figure 2 summarizes the contrast between sentences with activity verbs and sentences with ObjExp psych verbs according to both word orders (SVO in red and OVS in blue) and information structure (GN in dashed lines; NG in solid lines). Positive values mean that reading time is longer and regression counts are higher for sentences with activity verbs than for sentences with ObjExp psych verbs. A positive blue line and a negative red line correspond to an interaction between Verb Type and Word Order. Absolute values higher for solid lines than for dashed lines show an effect of Information Structure as expected, with non-canonical information structure showing

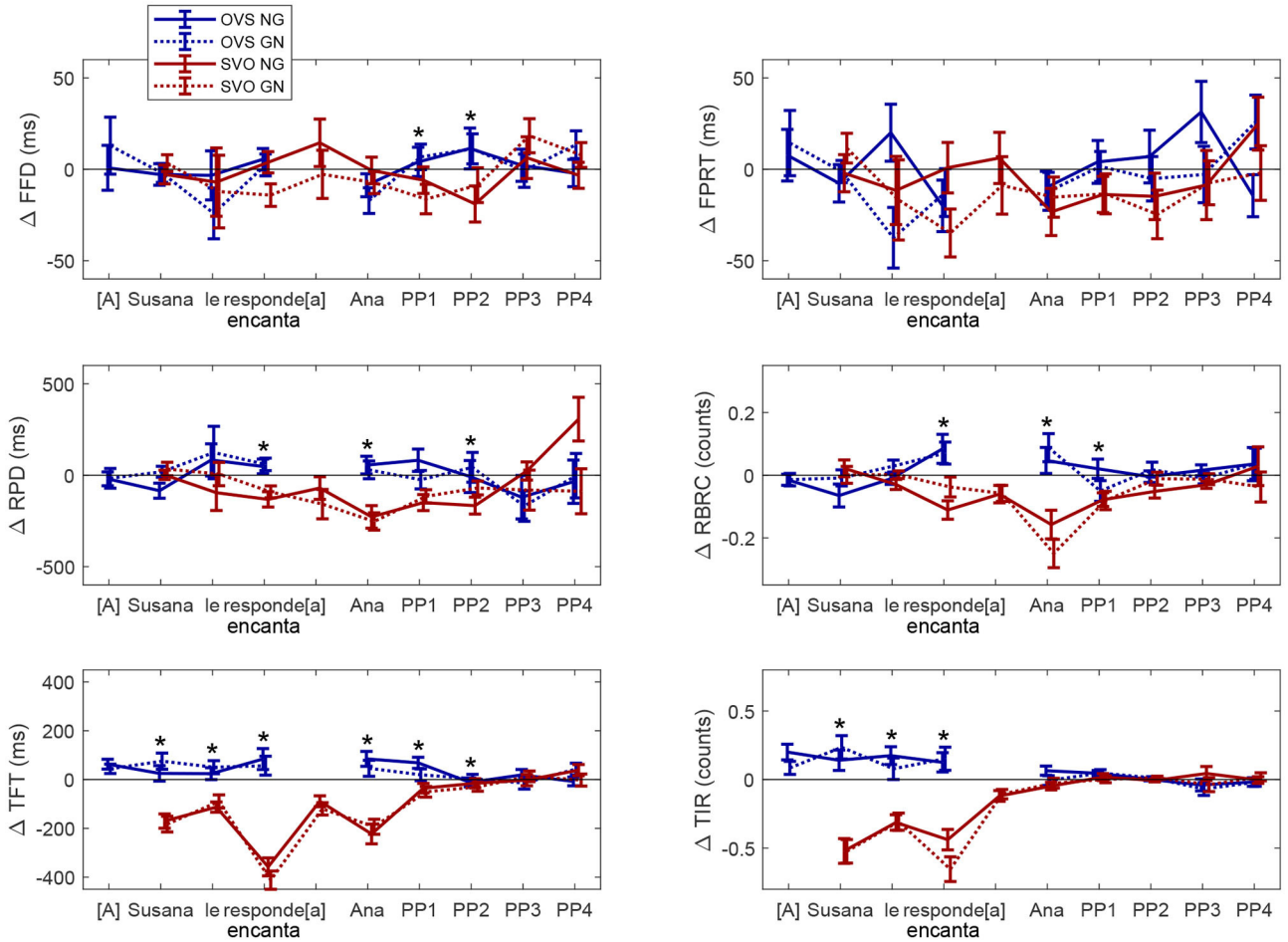

FIGURE 2 | The Figure shows the difference (Δ) in mean fixation times (ms) and the amount of regressive saccades (counts) between conditions with Activity Verbs and conditions with Psych verbs according to the sentence word order (SVO vs. OVS) and information structure (Given-New vs. New-Given) Error bars correspond to Standard Error of the Mean. Eye-tracking measures: FFD, First Fixation Duration; FPRT, First Pass Reading Time; TFT, Total Fixation Time; RPD, Regression Path Duration; RBRC, Right-Bounded Regression Count; TIR, Total Incoming Regressions. Word Order: SVO, Subject-Verb-Object; OVS, Object-Verb-Subject. The asterisk shows that the interaction between Word Order and Verb Type was significant.

higher cognitive demand than canonical information structure. This representation makes the interaction and Information Structure effect visually clear. The asterisks show the regions where the interaction was significant.

We now provide the analysis of regions of interest for both the early and late measures mentioned in the section 2.5. For each region, we first present the analysis of the interactions among factors since they comprise the contrasts of interest of the current study. We then provide the relevant results of the multiple comparisons test whenever was needed. Finally, we report main effects of Verb Type, Word Order or Information Structure. Appendix B shows the final converging models for each measure at each region. A table with all statistical coefficients, standard errors and *t* values may be found at Appendix C.

Region 1 (Case marking preposition "a")

Analysis of late eye-movement measures revealed no interactions among factors. A main effect of Verb Type was found for Total Fixation Time (TFT) showing longer fixation time for sentences with activity verbs than for sentences with ObjExp psych verbs; $\beta = 0.636$, $SE = 0.019$, $t = 3.311$, $p < 0.001$.

Region 2 (First proper name)

Late eye-movement measures showed a significant interaction between Verb Type and Word Order for Total Fixation Time (TFT) and Total Incoming Regressions (TIR); TFT: $\beta = 0.086$, $SE = 0.011$, $t = 7.923$, $p < 0.001$; TIR: $\beta = 0.192$, $SE = 0.018$, $z = 10.619$, $p < 0.001$). Resolving these interactions revealed that for subject-initial sentences, the probability of regressing to this region and the total fixation time were significantly longer when the sentence included an ObjExp psych verb than when it included an activity verb (TFT: $\beta = 0.259$, $SE = 0.031$, $z = 8.384$, $p < 0.001$; TIR: $\beta = 0.580$, $SE = 0.060$, $z = 9.590$, $p < 0.001$). Conversely, for object-initial sentences, the probability of regressing into this region and the total fixation time were significantly longer when the sentence included an

activity verb than when it included an ObjExp psych verb (TFT: $\beta = 0.087$, $SE = 0.031$, $z = 2.811$, $p = 0.025$; TIR: $\beta = 0.181$, $SE = 0.056$, $z = 3.258$, $p = 0.005$). Analysis of Regression Path Duration (RPD) also showed a main interaction between word order and information structure $\beta = -0.027$, $SE = 0.0127$, $t = -2.158$, $p < 0.05$. Tukey *post-hoc* test revealed that non-canonical word order (OVS) led to significantly longer regression path duration before continuing reading when the sentence presented a non-canonical information structure, $\beta = -0.107$, $SE = 0.035$, $z = -3.022$, $p < 0.05$. This difference was not significant between sentences with canonical information structure.

Analysis of early eye-movement measures revealed a main effect of Information Structure for First Fixation Duration (FFD) and First Pass Reading Time (FPRT), showing longer reading time for sentences with non-canonical Information Structure than for sentences with canonical Information Structure (FFD: $\beta = -0.032$, $SE = 0.006$, $t = -5.018$, $p < 0.001$; FPRT: $\beta = -0.052$, $SE = 0.008$, $t = -6.714$,, $p < 0.001$.) A similar effect was found for Right-Bounded Regression Count (RBRC), showing higher amount of regressions from this region for sentences with non-canonical Information Structure: $\beta = -0.016$, $SE = 0.039$, $t = -4.161$, $p < 0.01$.

Analysis of the probability of regressions into this region (TIR), RBRC , RPD, and Total Fixation Time also revealed a main effect of Word Order. On average, participants regressed to this word significantly more, fixated on this word for longer time and regressed for significantly longer time and higher amount of times from this region in object-initial sentences than in subject-initial sentences; (TIR: $\beta = 0.155$, $SE = 0.033$, $z = 4.710$, $p < 0.001$; TFT: $\beta = 0.063$, $SE = 0.011$, $t = 5.744$, $p < 0.001$; RBRC: $\beta = 0.010$, $SE = 0.040$, $t = 2.502$, $p < 0.05$; RPD: $\beta = 0.026$, $SE = 0.013$, $t = 2.065$, $p < 0.05$. Finally, effects of Verb Type and Information Structure were present for TFT and effects of Information Structure were found at RPD(TFT: Verb Type: $\beta = -0.043$, $SE = 0.011$, $t = -3.963$, $p < 0.001$; Information Structure:$\beta = -0.110$, $SE = 0.013$, $z = -8.203$, $p < 0.001$, RPD - Information Structure: $beta = -0.083$, $SE = 0.013$, $z = -6.543$, $p < 0.001$. The sign of these effects reveal that participants fixated for longer time when the sentence included an ObjExp Psych verb and fixated and regressed to previous regions for longer time when the noun corresponded to a new referent.

Region 3 (Clitic)

Analysis of early eye-movement measures revealed no interactions among fixed factors nor main effects. Analysis of late eye-movement measures showed a significant interaction between Verb Type and Word Order for total fixation time, and for the probability of regressions into this region (TFT: $\beta = 0.074$, $SE = 0.014$; $t = 5.331$; TIR: $\beta = -0.218$, $SE = 0.024$ $z = 8.977$, $p < 0.001$). This interaction follows the same direction as the interaction found on Region 2. For subject-initial sentences, participants regressed and fixated on this region significantly more when the sentence contained an ObjExp psych verb than when it included an activity verb (TFT: $\beta = 0.226$, $SE = 0.040$, $z = 5.685$, $p < 0.001$; TIR: $\beta = 0.624$, $SE = 0.072$, $z = 8.630$, $p < 0.001$). In object-initial sentences; participants regressed to this region significantly more when the sentence contained an activity verb; TIR: $\beta = 0.246$, $SE = 0.067$, $z = 3.748$, $p < 0.001$. This difference was not significant for TFT; $\beta = 0.071$, $SE = 0.039$, $z = 1.817$, $p < 0.265$. The analysis of these measures also revealed main effects of Verb Type (TFT: $\beta = -0.039$, $SE = 0.014$; $t = -2.791$; TIR: $\beta = -0.094$, $SE = 0.024$ $z = -3.833$, $p < 0.01$), Word Order (TIR: $\beta = -0.075$, $SE = 0.024$ $z = 3.095$, $p < 0.01$; RPD: $\beta = 0.076$ $SE = 0.029$, $t = 2.639$, $p < 0.05$), and Information Structure (TFT: $\beta = -0.042$, $SE = 0.016$; $t = -2.597$, $p < 0.01$). Participants showed higher processing load whenever the sentences included an ObjExp psych verb than when the included an activity verb. They also regressed to this region significantly more when the sentence followed the OVS order than when it followed the SVO word order, and fixated for longer time on this region when the first NP belonged to a new referent than when it belonged to an already given one.

Region 4 (Disambiguating verb)

Analysis of early eye-movement measures revealed no interactions among factors. However, a main effect of Information Structure was found for FPRT, showing significantly longer reading time for this region whenever the sentence presented a new referent in first position; $\beta = -0.042$, $SE = 0.009$, $t = -4.706$, $p < 0.001$. A significant interaction between Verb type and Word Order was found for all late eye-movement measures (RPD: $\beta = 0.053$, $SE = 0.012$, $t = 4.449$, $p < 0.001$; RBRC: $\beta = -0.135$, $SE = 0.033$, $z = 4.081$, $p < 0.001$; TFT: $\beta = 0.119$, $SE = 0.010$, $t = 11.852$tcr, $p < 0.001$; TIR: $\beta = 0.202$, $SE = 0.019$, $z = 10.531$, $p < 0.001$) Tukey *post-hoc* test showed that this interaction follows the same direction as in the previous region (SVO: RBRC: $\beta = 0.276$, $SE = 0.102$, $z = 2.711$, $p < 0.05$; TFT: $\beta = 0.358$, $SE = 0.047$, $z = 7.701$, $p < 0.001$; TIR: $\beta = 0.631$, $SE = 0.081$, $z = 7.751$, $p < 0.001$; OVS: RPD: $\beta = 0.122$, $SE = 0.038$, $z = 3.222$, $p = 0.007$; TFT: $\beta = 0.118$, $SE = 0.046$, $z = 2.544$, $p < 0.05$; RBRC: $\beta = 0.266$, $SE = 0.091$, $z = 2.930$, $p < 0.05$). TIR also showed a significant interaction among the three main factors, $\beta = 0.041$, $SE = 0.019$, $z = 2.112$, $p < 0.05$. Tukey HSD multiple comparisons showed that this triple interaction depended on the interaction between Verb and Word Order: when new information was in both first and second position, and sentences included an ObjExp psych verb, participants regressed significantly more to this region in SVO conditions than in OVS sentences (New-Given:$\beta = 0.033$, $SE = 0.073$, $z = 4.546$, $p < 0.001$; Given-New: $\beta = 0.040$, $SE = 0.071$, $z = 5.641$, $p < 0.001$). The opposite pattern took place for sentences with activity verbs: participants regressed significantly more to this region when the sentence followed the OVS word order than when it followed the SVO order (New-Given: $\beta = 0.033$, $SE = 0.079$, $z = 4.160$, $p < 0.001$; Given-New: $\beta = 0.057$, $SE = 0.084$, $z = 6.762$, $p < 0.001$). A significant effect of Word Order was found for RPD, RBRC and TFT in the same direction as in the previous region: participants found higher processing cost at this region for OVS sentences than for SVO sentences (RPD: $\beta = 0.044$, $SE = 0.012$, $t = 3.739$, $p < 0.001$; RBRC: $\beta = 0.110$, $SE = 0.033$, $z = 3.280$; TFT: $\beta = 0.419$, $SE = 0.010$, $t = 4.173$, $p < 0.001$) Furthermore, participants fixated for significantly longer time

on this region and regressed significantly more to it whenever it included an ObjExp psych verb than when it included an activity verb (TFT: $\beta = -0.060$, $SE = 0.021$, $t = -2.865$, $p < 0.01$; TIR: $\beta = -0.115$, $SE = 0.035$, $z = -3.251$, $p < 0.01$). Finally, Information Structure modulated both RPD and TFT. Participants fixated for longer time on that region and previous regions before continuing reading and fixated for longer time on that word whenever the sentence included a new referent in first position (RPD: $\beta = -0.062$, $SE = 0.012$, $t = -5.175$, $p < 0.001$; TFT: $\beta = -0.073$, $SE = 0.011$, $t = -6.827$, $p < 0.001$).

Region 5 (Case marking preposition "a")

Analysis of this region showed a significant main effect of Verb Type for most late eye-movement measures, with participants experiencing greater cognitive load and regressing significantly more to this region whenever the sentence included an ObjExp psych verb (RPD: $\beta = -0.126$, $SE = 0.041$; $t = -3.074$, $p < 0.01$; RBRC: $\beta = -0.28617$, $SE = 0.089$; $z = -3.225$, $p < 0.001$;TFT: $\beta = -0.132$, $SE = 0.020$; $t = -6.544$, $p < 0.001$; TIR: $\beta = -0.348$, $SE = 0.059$; $z = -5.921$, $p < 0.010$).

Region 6 (Second proper name)

Analysis of this region showed an interaction between Verb Type and Word Order for all late eye-movement measures (RPD: $\beta = 0.116$, $SE = 0.015$, $t = 7.598$, $p < 0.001$; RBRC: $\beta = 0.177$, $SE = 0.028$, $z = 6.253$, $p < 0.01$; TFT: $\beta = 0.010$; $SE = 0.011$; $t = 9.150$, $p < 0.001$; TIR: $\beta = 0.084$; $SE = 0.040$; $t = 2.125$, $p < 0.05$). Resolving these interactions revealed that in SVO sentences, participants fixated for longer time at this and previous regions and regressed significantly more times from and to this region when the sentence contained an ObjExp psych verb than when it included an activity verb (RPD: $\beta = 0.398$, $SE = 0.043$, $z = 9.235$, $p < 0.001$; RBRC: $\beta = 0.534$, $SE = 0.081$, $z = 6.573$, $p < 0.001$; TFT: $\beta = 0.332$, $SE = 0.033$, $z = 9.962$, $p < 0.001$; TIR: $\beta = 0.624$, $SE = 0.072$, $z = 8.693$, $p < 0.001$).

Difference among OVS conditions was only significant for Total Incoming Regressions, with conditions with activity verbs showing a higher amount of regressions to this region than conditions with ObjExp psych verbs ($\beta = 0.246$, $SE = 0.065$, $z = 3.777$, $p < 0.001$).

A main effect of Word Order was found for FFD and FPRT, with longer reading times for sentences with SVO word order than for sentences with OVS order (FFD: $\beta = -0.016$; $SE = 0.007$; $t = -2.292$, $p = 0.002$; FPRT: $\beta = -0.034$; $SE = 0.008$; $t = -3.989$, $p < 0.01$).

A significant effect of Verb Type was found for FPRT and all late eye-movement measures except for TIR. The sign of the effect shows longer reading time, regression duration and amount of regressions from this region when the sentences included an ObjExp psych verb than when they included an activity verb (FPRT: $\beta = -0.023$; $SE = 0.008$; $t = -2.724$, $p < 0.01$; RPD: $\beta = -0.083$; $SE = 0.015$; $t = -5.438$, $p < 0.001$; RBRC: $\beta = -0.090$; $SE = 0.028$; $z = -3.180$, $p < 0.01$; TFT: $\beta = -0.064$; $SE = 0.012$; $t = -5.172$). Finally, a significant effect of Information Structure was also for FPRT. Contrary to the effect of Information Structure found in previous regions, this region shows longer reading time for conditions with new information in second position than for conditions with new information in first position; $\beta = -0.026$; $SE = 0.008$; $t = -3.101$, $p < 0.01$.

Region 7 (First word of the Spill-over region)

Analysis of this region showed that the interaction between Verb Type and Word Order was significant for FFD and for three out of five late eye-movement measures (FFD: $\beta = 0.020$, $SE = 0.008$, $t = 2.649$, $p = 0.008$; RPD: $\beta = 0.066$, $SE = 0.016$, $t = 4.206$, $p < 0.001$; RBRC: $\beta = 0.151$; $SE = 0.051$, $z = 2.953$, $p < 0.05$; TFT: $\beta = 0.052$; $SE = 0.011$, $t = 4.579$, $p < 0.001$). The multiple comparisons test showed that in subject-initial sentences, participants fixated for longer time at this region when the sentence contained an ObjExp psych verb than when it included an activity verb (TFT: $\beta = 0.0332$, $SE = 0.032$, $t = 10.477$, $p < 0.001$). Participants also spent longer time reading and regressing to previous regions and regressed significantly more times from this region for sentences with ObjExp psych verbs than for sentences with activity verbs (RPD: $\beta = 0.021$, $SE = 0.045$, $z = 4.724$, $p < 0.001$; RBRC: $\beta = 0.665$; $SE = 0.146$, $t = 4.551$, $p < 0.001$).

Differences among object-initial sentences were marginally significant for Total Fixation Time, with participant fixating for longer time on this region when the sentence included an activity verb than when it included an ObjExp psych verb; TFT: $\beta = 0.077$, $SE = 0.032$, $z = 2.436$, $p = 0.070$. For the other four measures this difference was not significant.

Analysis of this region also showed main effects of word order and verb type for late measures RPD and RBRC. Participants regressed significantly more from this region and spent significantly longer time on previous regions before continuing reading when the sentences included ObjExp psych verbs than when they included activity verbs and when they followed the OVS word order than when they were subject-initial sentences (Word Order: RPD: $\beta = 0.049$, $SE = 0.016$, $t = 3.127$, $p < 0.01$; RBRC: $\beta = 0.021$, $SE = 0.048$, $z = 4.454$, $p < 0.001$; Verb: RPD: $\beta = -0.039$, $SE = 0.016$, $t = -2.517$, $p < 0.05$; RBRC: $\beta = -0.184$, $SE = 0.046$, $z = -3.983$, $p < 0.01$).

Region 8 (Second word of the Spill-over Region)

A significant interaction between Verb Type and Word Order was found for FFD and RPD (FFD: $\beta = 0.023$, $SE = 0.010$, $t = 2.555$, $p < 0.05$; RPD: $\beta = 0.042$, $SE = 0.019$, $t = 2.224$, $p < 0.05$). The multiple comparisons Tukey HSD test revealed significant differences among SVO conditions for RPD only, with conditions with ObjExp psych verbs engendering longer regression path duration than sentences with activity verbs. Differences among OVS conditions were not significant for any of the above-mentioned measures. Analysis of FFD and TFT also showed a significant interaction between Word Order and Information Structure (FFD: $\beta = 0.021$, $SE = 0.010$, $t = 2.087$, $p < 0.05$; TFT: $\beta = 0.028$, $SE = 0.013$, $t = 2.138$). However, the interaction was not confirmed by the multiple comparisons tests from both measures. Finally, a main effect of Verb Type was found for RPD, with sentences with ObjExp psych verbs engendering longer regression path duration than sentences with activity verbs; $\beta = -0.046$, $SE = 0.019$, $t = -2.413$, $p < 0.05$.

Region 9 (Third word of the Spill-over Region)
Analysis of early eye-movement measures revealed a significant interaction between Word Order and Information Structure for FPRT; $\beta = 0.021$, $SE = 0.010$, $t = 2.037$, $p < 0.05$. However, the multiple comparisons test showed no significant differences among conditions.

Analysis of late eye-movement measures showed a main effect of Verb Type for RPD and a main effect of Word Order for TIR (Verb: $\beta = -0.055$, $SE = 0.026$, $t = -2.113$, $p < 0.05$; Word Order: $\beta = 0.056$, $SE = 0.026$, $z = 2.198$, $p < 0.05$). The sign of the effects show that participants regressed significantly longer time to previous regions when the sentence included an ObjExp psych verb and that they regressed more times to this region when the sentence followed the OVS word order.

Region 10 (Fourth word of the Spill-over Region)
Analysis of this region showed a significant interaction among Word Order, Verb Type and Information Structure for FPRT and RPD (FPRT: $\beta = 0.024$, $SE = 0.010$, $t = 2.484$, $p < 0.05$; RPD: $\beta = 0.056$, $SE = 0.021$, $t = 2.680$, $p < 0.01$). Resolving these interactions revealed that for First Pass Reading Time, participants spent longer time reading this region in sentences with activity verbs and a new referent in first position when the sentence followed the SVO order than when it followed the OVS one; $\beta = 0.122$; $SE = 0.039$, $t = 3.164$, $p < 0.05$. Differences among the other conditions did not reach significance for this measure nor for RPD. A main effect of Word Order was also sound for FPRT. Participants spent longer time on this region when word order was SVO than when it was OVS, $\beta = -0.0239$, $SE = 0.010$, $t = -2.486$, $p < 0.05$ Finally, a main effect of Information Structure was found for RPD, $\beta = -0.064$, $SE = 0.021$, $t = -3.078$, $p < 0.01$. The sign of the effect shows longer reading time for conditions with new information in second position than for conditions with new information in first position.

4. DISCUSSION

Evidence about the comprehension of isolated Spanish sentences with alternative word orders has shown that readers manifest increasing difficulty to understand sentences with a word order that does not respect the order of arguments at the lexico-semantic structure of the verb, independently of whether the sentence follows the canonical word order of the language (SVO) or not (Gattei et al., 2015a,b, 2017). In these studies, the authors used sentences with activity verbs and object experiencer verbs in order to compare events that required alternative linking between syntax and semantics. When reading SVO sentences, participants required significantly more time to read and figure out "who did / felt what for whom" when the sentence included an ObjExp psych verb. Conversely, when reading OVS sentences, participants required more time to read when the sentence included an activity verb. This interaction was present when using both self-paced reading and eye-tracking techniques.

These studies support the hypothesis that during incremental parsing, readers use the morphosyntactic and semantic information provided by the first sentential argument to generate predictions about the verb type that will take place in the sentence according to the prominence status of that argument. This proposal assumes that the language processing system gives rise to predictions about arguments order by following a prominence hierarchy that canonically stipulates that more prominent arguments precede less prominent ones (Bornkessel et al., 2005; Wolff et al., 2007; Haupt et al., 2008). Nonetheless, it is relevant to ask whether the pragmatic use of constituents order has any influence on the mentioned results. In the current study, we focused on the distinction between "given" and "new" referents in a sentence in relation with a previous context. According to Givón (1984), the use of a non-canonical word order is expected when mentioning a referent that has already been introduced by the previous context so that the rhematic hierarchy (i.e., given-new) is respected. Experimental evidence about the role of rhematic hierarchy during incremental reading has shown that effects of word order non-canonicity are alleviated when an adequate context precedes the sentence (Kaiser and Trueswell, 2004; Burmester et al., 2014), suggesting that increasing reading times in isolated sentences with non-canonical word order could partly be due to higher cognitive demands related to making assumptions about possible contexts in which a non-canonical word order could take place. However, the role of information structure in relation to alternative word orders stipulated by different lexico-semantic configurations had not been explored yet. Thus, the question that motivated the present study was whether the prominence effects found for Spanish sentence comprehension were caused by the lack of a context that could motivate the election of a specific word order. We thus framed sentences used in Gattei et al. (2017) in texts that would favor the appearance of a specific referent in first position of the sentence and compared them with sentences in which the first argument comprised a new referent. By means of a comprehension offline task, we also evaluated the cognitive cost of understanding "who did/felt what for whom" correctly.

Regarding the hypotheses and predictions outlined at the section 1, the current work shows that context adequacy plays a role for processing of sentences with non-canonical word order but does not make effects of prominence miscomputation fade away.

On the one hand, results of the current study revealed that the use of an adequate context facilitated the comprehension of the target sentences. Participants took significantly less time to read the texts when, in first position, the final sentence introduced a referent that had explicitly been presented before. They also took less time to respond the comprehension questions when the target sentence included a canonical information structure. In other words, these results replicate the findings that Kaiser and Trueswell (2004) showed for Finnish sentences with activity verbs. Additionally, the current study revealed that a non-canonical information structure is detrimental to comprehension even in SVO sentences, as it is evidenced by sentences' response time of questions about sentences with ObjExp psych verbs.

On the other hand, effects related to incorrect syntax-to-semantic linking were present during reading for late eye-movement measures as predicted, showing a disruption of processes of higher-level text integration (Clifton et al., 2007).

When encountering a verb that did not match the predictions stipulated by the computation of prominence status of the first argument, participants took longer time to read the word and following content words, and regressed more times and for longer time to previous regions. These results replicate the findings by Gattei et al. (2017) for isolated sentences, and yield further evidence in favor of the hypothesis that one of the central mechanisms for argument interpretation is prominence computation (Bornkessel and Schlesewsky, 2006), and that prominence computation follow a principle that assumes that the first argument will be the most "Actor-like" possible.

The current study also provides interesting insight about the cognitive cost and strategies used by readers to process new information. Several proposals have tried to explain the cognitive effects derived from the use of a non-canonical rhematic hierarchy. Although there is an agreement regarding the type of effects caused by the unpredictable appearance of a new referent, there is not a unique view with respect to which mechanisms are involved in information structure processing. For instance, it has been proposed that speakers tend to choose the syntactic constructions that allow them to place the most "accessible" (already mentioned) information earlier in the utterance (Ferreira, 2003), possibly because this allows them to postpone the difficult part of the utterance, which requires more resources to plan. The assumption behind this hypothesis is that when information has a strong representation in memory it is easier to retrieve and to process. Evidence in favor of this view shows that speakers choose word order according to visual attention (Gleitman et al., 2007).

Kaiser (2012) argues that the pragmatic status of referent emerges naturally from memory and attention. Theories about memory distinguish between working memory, which stores information currently being used, and long-term memory, which stores the conceptual and procedural knowledge for posterior use. From this point of view, given referents can be defined as those accessed through working memory (and thus easier to retrieve) and new referents as those which have not been retrieved by long-term memory yet (Arnold et al., 2013).

From a neurobiological perspective of language and its processing, Bornkessel-Schlesewsky and Schumacher (2016) propose that, instead of postulating specific neural correlates for information structure, a more promising approach is to consider that information structure affects domain-general mechanisms when hierarchically guiding predictive processing or when providing cues for attentional shift. The authors claim that the status of discourse referents feeds the predictive processes during discourse, as it is shown by how the preference for a continuity of the same referent or for certain types of linearization (i.e., given referents precede new ones) facilitate language processing. Errors in the predictions at this level elicit negative potentials (for instance N400 for unpredictable information structure properties) and result in attentional reorienting and mental model updating required in topic shift scenarios.

A general idea that stems from these approaches is that the violation of rhematic structure involves the modification of a more flexible system, designed to adapt linguistic form to the dynamics of context -possibly through working memory capacity-, and prepared to deal with domain-general mechanisms like attentional shift and reorienting. Although the current work was not aimed at disentangle whether the effects produced by the use of a non-canonical rhematic hierarchy were related to factors associated to referents' accessibility, readers' working memory capacity or a failure of an expected structure and the subsequent need for attentional reorientation, the current findings are informative with regards to the mechanisms underlying incremental processing of new information.

In the current study, eye-movement measures showed information structure effects, evidenced by increasing reading times for the first and the second NP whenever readers found a new referent. However, it is generally argued that reading words that are repeated throughout a text entails a decrease of reading time (Rayner et al., 1995; O'Brien et al., 1997; Kamienkowski et al., 2018). Hence, it is fair to ask whether the effects found at new referents respond to the manipulation of rhematic hierarchy or if they should be interpreted as lexical repetition effects related to word recognition processes. For instance, Lowder et al. (2013) ran en eye-tracking study in which participants had to read sentences with two NPs composed by two and one proper names, respectively. The authors manipulated proper names' frequency and repetition (the second NP could mention one of the proper names from the first NP or not) and showed that when reading the second NP, repeated names were processed more quickly than new names in both early and late eye-movement measures. Following Gordon and Hendrick (1998), the authors argue that "while basic word recognition goes on, the effort to understand the meaning of a sentence or short discourse leads to the construction of a discourse model that represents patterns of reference and co-reference and which captures the predicate-argument relationships described in the text" (Lowder et al., 2013). The results of the current study showed a similar pattern of results, with a modulation of both early and late eye-movement measures at the first NP and following two regions, when the proper name comprised a new referent. Interestingly, when the sentence followed a canonical information structure, effects of a new referent were only present for First Pass Reading Time at the proper name region. We interpret this pattern of results as a difference in the control and time course of oculomotor processes for word recognition, with short-lived, early effects, and for information structure manipulation, which affected late eye-movements and caused a longer comprehension disruption.

As for offline measures collected in the current study, total reading time is informative of the time required by readers to guarantee that they have understood the text. Although these were the instructions provided, this measure did not reflect comprehension success, as shown by accuracy rates. In particular, participants responded questions significantly better when the final sentence followed Spanish canonical word order (SVO), independently of whether the initial constituent consisted of a new referent or not. Although this was expected for sentences with activity verbs, a preference for SVO word order was not expected for accuracy rates of sentences with ObjExp psych verbs, diverging from the results found in previous studies about this issue with isolated sentences. In Gattei et al. (2017) participants

showed overall higher accuracy rates (around 90% accuracy for critical questions), and higher accuracy rates for questions about sentences with activity verbs than for those about ObjExp psych verbs. While further investigation of this difference between experiments is needed, it is possible that the use of additional context and the requirement of keeping referents in working memory in order to reply the comprehension questions, had a negative effect on the comprehension of Spanish overall less frequent word order.

With regards to the interaction between verb type and word order found for accuracy rates, results replicate the findings of Gattei et al. (2017), with higher accuracy for questions about SVO sentences with activity verbs than for the other conditions. We argue that this pattern is expected as in this type of sentences both semantic and syntactic canonical orders coincide, while the other conditions present an alteration of either semantic order (as in SVO sentences with ObjExp psych verbs), constituents order (as in OVS sentences with ObjExp psych verbs) or both (as in OVS sentences with Activity verbs). In other words, results show that the alignment of both canonical linking and canonical word constituents order facilitates comprehension, whereas non canonical arrangement of either type of information makes it more difficult.

A final aspect that needs to be taken into account is the response time for comprehension questions, which show that readers needed extra time to respond questions about sentences with either OVS word order, non-canonical information structure or ObjExp psych verbs. We believe that the lack of information structure effects for overall accuracy shows that while the use of a non-canonical rhematic hierarchy require longer reading time and response time for comprehension questions, the consequences of not following a canonical order for information structure are not as strong as to show a modulation of comprehension success, as it occurs with OVS sentences. However, it is matter of future research to evaluate whether differences in the trade-off between response time and accuracy for word order and information structure non-canonicity respond to task-related factors (as structural complexity or types of questions used) or individual differences (like working memory or attentional capacities).

4.1. Possible Methodological Caveats and Future Directions

Although results of the current study support previous results on this issue, possible methodological caveats should be taken into account for future research and replication in other languages. Most importantly, while the study asks about the role of prominence computation in sentences embedded in context, the materials were designed in such way that they do not directly compare comprehension of isolated sentences with comprehension of sentences embedded in texts within the same group of subjects. The rationale of not doing so was that adding no-context trials would have implied to double the amount of conditions to sixteen conditions. Considering the short amount of ObjExp psych verbs available in Spanish, this would have implied that participants either read each verb four times (as opposed to two as it occurs in the current version of the experiment), enabling the possibility of introducing the confound of structure repetition effect and other possible confounds due to participants tiredness or boredom, or that the amount of subjects tested was doubled to approximately 150 to yield results comparable to the current ones. Considering that the sentences without context have been repeatedly tested (Gattei et al., 2015a,b, 2017) we considered that the design of the current study was a fair trade-off between running the ideal experiment and getting reliable results. Still, adding further isolated conditions in languages that have not been previously tested would be crucial for results' replication.

A second aspect that needs further investigation is how participants deal with referent's activation when encountering sentences with non-canonical information structure. In other words, can regressive saccades from regions comprising new referents to previous sentences be informative of participants' reading strategies related to referent updating? (Chafe, 1976, 1994). While, this question was out of the scope of our work, we are currently addressing this issue with the data currently collected.

5. CONCLUSIONS

We presented an exploratory study that evaluated the interaction between word order, lexico-semantic structure of the verb and information structure in the comprehension of Spanish texts. Previous studies about this language have only evaluated the role of the first two factors, leaving aside the pragmatic aspect involved in the election of constituents' word order. Understanding the role of information structure is crucial to explain sentence processing in this language, since previous evidence has shown that when sentences are presented in isolation, constituents order is a relevant cue for incremental argument interpretation. It was pertinent to ask whether word order is still a relevant cue for argument interpretation when the previous context justifies (or not) the appearance of a specific word order. By evaluating reading of texts that manipulated the relation between "given" and "new" referents we showed that while information structure canonicity enhances comprehension, the use of an adequate context for a specific word order does not alleviate comprehension effects caused by argument misinterpretation. This type of evidence is crucial for any model of language comprehension that attempts to explain sentence processing in languages that allow alternative word orders.

By conducting an eye-tracking experiment, we could also provide further information about the time course of on-going processing of new referents, which show a different gaze signature to lexically-driven word retrieval.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article are available at https://osf.io/kp4dn/.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Facultad de Filosofía y Letras - Universidad de Buenos Aires. The participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

CG designed the materials of the study, collected the data, ran the statistical analysis, and sketched the first draft of this manuscript. DS programmed the Matlab script for data collection, data preanalysis, and figures' design. LP supervised the adequacy of hypotheses and predictions to formal approaches of information structure canonicity. All authors equally contributed with the initial discussion of hypotheses, predictions and results of the current study, and writing the final version of this manuscript.

FUNDING

This work was part of the CG postdoctoral fellowship work funded by the National Council of Scientific and Technical Research (CONICET) for the project Prediction and integration during information structure parsing in Spanish and its interaction with the syntax-semantics interface. This work was also funded by the James McDonnell Foundation JSMF Grant #220020334, awarded to Laboratorio de Neurociencia at Universidad Torcuato Di Tella.

ACKNOWLEDGMENTS

We would like to thank the contribution of the attendees and organizers of the I Encuentro Latinoamericano de Interfaces Gramaticales (ELIG) for the thorough discussion of the current results previous to this article's submission.

REFERENCES

Arnold, J. E., Kaiser, E., Kahn, J. M., and Kim, L. K. (2013). Information structure: linguistic, cognitive, and processing approaches. *Wiley Interdiscipl. Rev. Cogn. Sci.* 4, 403–413. doi: 10.1002/wcs.1234

Baayen, R. H. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge, MA: Cambridge University Press. doi: 10.1017/CBO9780511801686

Bader, M., and Bayer, J. (2006). *Case and Linking in Language Comprehension: Evidence from German, Volume 34 of Studies in Theoretical Psycholinguistics*. Dordrecht: Springer Science & Business Media.

Bader, M., and Meng, M. (1999). Subject-object ambiguities in German embedded clauses: an across-the-board comparison. *J. Psycholinguist. Res.* 28, 121–143. doi: 10.1023/A:1023206208142

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., Machler, M., Bolker, B., and Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*. doi: 10.18637/jss.v067.i01

Belloro, V. (2012). "La estructura informativa," in *El funcionalismo en la teoría lingüística: la Gramática del Papel y la Referencia; Introducción, avances y aplicaciones*, eds R. Mairal, L. Guerrero, and C. González Vergara (Madrid: Akal), 225–244.

Birner, B. J., and Ward, G. L. (1998). *Information Status and Noncanonical Word Order in English*, Vol. 40. Amsterdam: John Benjamins Publishing. doi: 10.1075/slcs.40

Bornkessel, I., and Schlesewsky, M. (2006). The extended argument dependency model: a neurocognitive approach to sentence comprehension across languages. *Psychol. Rev.* 113:787. doi: 10.1037/0033-295X.113.4.787

Bornkessel, I., Schlesewsky, M., and Friederici, A. D. (2003). Eliciting thematic reanalysis effects: the role of syntax-independent information during parsing. *Lang. Cogn. Process.* 18, 269–298. doi: 10.1080/01690960244000018

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

Bornkessel-Schlesewsky, I., and Schumacher, P. B. (2016). "Towards a neurobiology of information structure," in *The Oxford Handbook of Information Structure*, eds C. Fery and S. Ishihara (Oxford: Oxford University Press), 581–598. doi: 10.1093/oxfordhb/9780199642670.013.22

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

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

Chafe, W. (1994). *Discourse, Consciousness, and Time: The Flow and Displacement of Conscious Experience in Speaking and Writing*. University of Chicago Press.

Chow, W.-Y., and Phillips, C. (2013). No semantic illusions in the 'semantic P600' phenomenon: ERP evidence from Mandarin Chinese. *Brain Res.* 1506, 76–93. doi: 10.1016/j.brainres.2013.02.016

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 (Elsevier Science), 341–372. doi: 10.1016/B978-008044980-7/50017-3

Comrie, B. (1989). *Language Universals and Linguistic Typology: Syntax and Morphology*. University of Chicago Press.

Contreras, H. (1976). *A Theory of Word Order with Special Reference to Spanish*. Amsterdam: North-Holland.

Cornelissen, F. W., Peters, E. M., and Palmer, J. (2002). The eyelink toolbox: eye tracking with MATLAB and the psychophysics toolbox. *Behav. Res. Methods Instrum. Comput.* 34, 613–617. doi: 10.3758/BF03195489

Davis, C. J., and Perea, M. (2005). Buscapalabras: a program for deriving orthographic and phonological neighborhood statistics and other psycholinguistic indices in Spanish. *Behav. Res. Methods* 37, 665–671. doi: 10.3758/BF03192738

Dröge, A., Maffongelli, L., and Bornkessel-Schlesewsky, I. (2014). "Luigi piace a Laura?" in *Structuring the Argument: Multidisciplinary Research on Verb Argument Structure*, eds A. Bachrach, I. Roy, and L. Stockall (Amsterdam: John Benjamins Publishing Company), 83–118. doi: 10.1075/lfab.10.05dro

Ferreira, V. S. (2003). The persistence of optional complementizer production: why saying "that" is not saying "that" at all. *J. Mem. Lang.* 48, 379–398. doi: 10.1016/S0749-596X(02)00523-5

Foley, W. A., and Van Valin, R. D. Jr. (1984). *Functional Syntax and Universal Grammar, Volume 38 of Cambridge Studies in Linguistics*. London: Cambridge University Press.

Gattei, C. A., Dickey, M. W., Wainselboim, A. J., and París, L. (2015a). The thematic hierarchy in sentence comprehension: a study on the interaction between verb class and word order in Spanish. *Q. J. Exp. Psychol.* 68, 1981–2007. doi: 10.1080/17470218.2014.1000345

Gattei, C. A., Sevilla, Y., Tabullo, A., Wainselboim, A., París, L., and Shalom, D. (2017). Prominence in spanish sentence comprehension: an eye-tracking study. *Lang. Cogn. Neurosci.* 33, 587–607. doi: 10.1080/23273798.2017.1397278

Gattei, C. A., Tabullo, Á., París, L., and Wainselboim, A. J. (2015b). The role of prominence in Spanish sentence comprehension: an ERP study. *Brain Lang.* 150, 22–35. doi: 10.1016/j.bandl.2015.08.001

Givón, T. (1984). *Syntax: A Functional-Typological Approach*. Amsterdam: J. Benjamins. doi: 10.1075/z.17

Gleitman, L. R., January, D., Nappa, R., and Trueswell, J. C. (2007). On the give and take between event apprehension and utterance formulation. *J. Mem. Lang.* 57, 544–569. doi: 10.1016/j.jml.2007.01.007

Gordon, P. C., and Hendrick, R. (1998). The representation and processing of coreference in discourse. *Cogn. Sci.* 22, 389–424. doi: 10.1207/s15516709cog2204_1

Haupt, F. S., Schlesewsky, M., Roehm, D., Friederici, A. D., and Bornkessel-Schlesewsky, I. (2008). The status of subject-object reanalyses in the language comprehension architecture. *J. Mem. Lang.* 59, 54–96. doi: 10.1016/j.jml.2008.02.003

Hyönä, H., and Hujanen, J. (1997). Effects of case marking and word order on sentence parsing in finnish: an eye fixation analysis. *Q. J. Exp. Psychol. A* 50, 841–858. doi: 10.1080/027249897391919

Just, M. A., and Carpenter, P. A. (1980). A theory of reading: from eye fixations to comprehension. *Psychol. Rev.* 87:329. doi: 10.1037/0033-295X.87.4.329

Just, M. A., Carpenter, P. A., and Woolley, J. D. (1982). Paradigms and processes in reading comprehension. *J. Exp. Psychol. Gen.* 111:228. doi: 10.1037/0096-3445.111.2.228

Kaiser, E. (2012). Taking action: a cross-modal investigation of discourse-level representations. *Front. Psychol.* 3:156. doi: 10.3389/fpsyg.2012.00156

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

Kamide, Y., and Mitchell, D. C. (1999). Incremental pre-head attachment in Japanese parsing. *Lang. Cogn. Process.* 14, 631–662. doi: 10.1080/016909699386211

Kamienkowski, J. E., Carbajal, M. J., Bianchi, B., Sigman, M., and Shalom, D. E. (2018). Cumulative repetition effects across multiple readings of a word: evidence from eye movements. *Discourse Process.* 55, 256–271. doi: 10.1080/0163853X.2016.1234872

Kliegl, R., Grabner, E., Rolfs, M., and Engbert, R. (2004). Length, frequency, and predictability effects of words on eye movements in reading. *Eur. J. Cogn. Psychol.* 16, 262–284. doi: 10.1080/09541440340000213

Lambrecht, K. (1994). *Information Structure and Sentence Form: Topic, Focus, and the Mental Representations of Discourse Referents*, Vol. 71. Cambridge, MA: Cambridge University Press. doi: 10.1017/CBO9780511620607

Lamers, M., and De Swart, P. (2012). *Case, Word order and Prominence: Interacting Cues in Language Production and Comprehension*, Vol. 40. Dordrecht: Springer Science & Business Media. doi: 10.1007/978-94-007-1463-2

Logacev, P., and Vasishth, S. (2013). *em2: A Package for Computing Reading Time Measures for Psycholinguistics*. R package version *0.9*. Available online at: http://CRAN.R-project.org/package=em2

Lowder, M. W., Choi, W., and Gordon, P. C. (2013). Word recognition during reading: the interaction between lexical repetition and frequency. *Mem. Cogn.* 41, 738–751. doi: 10.3758/s13421-012-0288-z

O'Brien, E. J., Raney, G. E., Albrecht, J. E., and Rayner, K. (1997). Processes involved in the resolution of explicit anaphors. *Discourse Process.* 23, 1–24. doi: 10.1080/01638539709544979

Ocampo, F. (1995). "The word order of two-constituent constructions in spoken Spanish," in *Word Order in Discourse*, eds P. A. Downing and M. Noonan (Amsterdam: John Benjamins), 425–447. doi: 10.1075/tsl.30.15oca

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*, eds J. C. Pinheiro and D. M. Bates (New York, NY: Springer), 3–56. doi: 10.1007/978-1-4419-0318-1_1

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rayner, K., Raney, G. E., and Pollatsek, A. (1995). "Eye movements and discourse processing," in *Sources of Coherence in Reading*, eds R. F. Lorch Jr., and E. J. O'Brien (Hillsdale), 9–35.

Rayner, K., and Well, A. D. (1996). Effects of contextual constraint on eye movements in reading: a further examination. *Psychon. Bull. Rev.* 3, 04–509. doi: 10.3758/BF03214555

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

Suñer, M. (1982). *Syntax and Semantics of Spanish Presentational Sentence-Types*. Georgetown University School of Language.

Van Valin, R. D., and LaPolla, R. J. (1997). *Syntax: Structure, Meaning, and Function*. Cambridge, MA: Cambridge University Press. doi: 10.1017/CBO9781139166799

Van Valin, R. D. Jr. (1999). "A typology of the interaction of focus structure and syntax," in *Typology and the Theory of Language: From Description to Explanation*, eds E. Raxilina and J. Testelec (Moscow: Languages of Russian Culture), 511–524.

Van Valin, R. D. Jr. (2005). *Exploring the Syntax-Semantics Interface*. Cambridge, MA: Cambridge University Press. doi: 10.1017/CBO9780511610578.001

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

Wang, L., Schlesewsky, M., Philipp, M., and Bornkessel-Schlesewsky, I. (2012). "The role of animacy in online argument interpretation in Mandarin Chinese," in *Case, Word Order and Prominence*, eds M. Lamers and P. de Swart (Dordrecht: Springer), 91–119. doi: 10.1007/978-94-007-1463-2_5

Wolff, S., Schlesewsky, M., and Bornkessel-Schlesewsky, I. (2007). The interaction of universal and language-specific properties in the neurocognition of language comprehension: evidence from the processing of word order permutations in Japanese. *J. Cogn. Neurosci* 288.

Zubizarreta, M. L. (1998). *Prosody, Focus, and Word Order*. Cambridge, MA: MIT Press.

Conflict of Interest: 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 © 2021 Gattei, París and Shalom. 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) and the copyright owner(s) 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

CONCEPTUAL ANALYSIS
published: 13 April 2021
doi: 10.3389/fpsyg.2021.662276

Conceptual Conflicts in Metaphors and Pragmatic Strategies for Their Translation

*Ilaria Rizzato**

Department of Modern Languages and Cultures, University of Genoa, Genoa, Italy

This article seeks to provide a theoretical exploration of Prandi's model of conceptual conflicts in metaphors (2017) and to highlight the advantages such model presents in its applications to translation and the text analysis preceding and preparing translation. Such advantages are mainly identified in the model aptness to meet the pragmatic requirements of translation, seen as a practice-based, goal-oriented and context-driven activity. These advantages also distinctly emerge from a comparison with the main tenets of the cognitive tradition. The theoretical basis for an understanding of conceptual conflict and its applications to translation are illustrated through the analysis of three brief excerpts from literary texts in English and their Italian translation.

Keywords: conceptual conflict, metaphors, pragmatics, English-to-Italian translation, translation studies

OPEN ACCESS

Edited by:
Valentina Cuccio,
University of Messina, Italy

Reviewed by:
Daniele Borgogni,
University of Turin, Italy
Elena Negrea-Busuioc,
National School of Political Studies and Public Administration, Romania

***Correspondence:**
Ilaria Rizzato
ilaria.rizzato@unige.it

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

Received: *31 January 2021*
Accepted: *15 March 2021*
Published: *13 April 2021*

Citation:
Rizzato I (2021) Conceptual Conflicts in Metaphors and Pragmatic Strategies for Their Translation. Front. Psychol. 12:662276. doi: 10.3389/fpsyg.2021.662276

INTRODUCTION

Professional translation is a goal-oriented activity based on strong practical objectives and aimed at concrete and effective outputs (Baker, 2011). Since it involves the interaction of two languages, linguistics is an obvious, major contributor to Translation Studies (Baker, 2011) and given the practical nature of translations as concrete products in contingent, specific communicative situations, pragmatics is at the forefront of the linguistic disciplines Translation Studies look to Snell-Hornby (1995, 2006), House (1997), and Baker (2011). Metaphors are traditionally considered a particularly challenging element in translation (Newmark, 1981; Schäffner, 2004). Thus, in the text analysis that precedes and prepares translation, especially in difficult cases, an approach to metaphor analysis which puts an emphasis on pragmatic aspects of metaphors should be adopted and preferred over approaches that do not foreground a pragmatic method.

This article will seek to illustrate how the notion of conceptual conflict in metaphors (Prandi, 2017) may account for pragmatic aspects in metaphors and hence be a reliable model for the metaphor analysis needed in preparation of translation. This theoretical exploration of the conceptual conflict model is conducted by comparing it to the dominant paradigm of metaphor theorization and analysis offered by Conceptual Metaphor Theory (CMT), based on the dual mapping of source and target domains (Lakoff and Johnson, 1980). Its central tenets will be exemplified through literary texts drawing on the author's experience in English to Italian literary translation practice and teaching.

CONFLICTUAL CONCEPTS IN METAPHORS: IMPLICATIONS FOR TRANSLATION

Although CMT has revealed the pervasiveness of metaphorical concepts in everyday language and thier importance in shaping human thought and communication, its emphasis on generalizable

features of metaphor may present detrimental aspects in its application to translation. Emphasis on universal elements in conceptual metaphors and the identification of similar patterns in conceptual metaphors across languages may play down on culture- or language-specific differences that may make a difference between a quality translation and an unacceptable one. The claim of universality for conceptual metaphor, moreover, may have diminished the role of context and situational circumstances in metaphor interpretation, with negative consequences on translation. Snell-Hornby; Snell-Hornby (1995; 2006) critique of universalist theories in favor of differentiation aims to highlight the risk of underestimating the rendering of significant differences in translation.

A consequence of this tendency toward generalization also lies in the idea that there is no difference between conventional and living (that is, unconventional) metaphors since they originate from the same metaphorical concepts (Lakoff and Turner, 1989). This view has significantly turned attention away from living metaphors and reinforced problematic tendencies in translation theory and practice[1]. Identifying conventional meanings across languages is in fact facilitated by dictionaries, linguistic repertories, databases, and similar tools, which rely on normativity and on recurrence in use. Conversely, translating unconventional elements not only requires knowledge of norms, but also sensitivity to deviation from such norms and to possible ways of recreating deviation in the target text with comparable effects. Hence, a model for metaphor identification and analysis encompassing unconventional elements as well as conventional ones may represent a precious tool for translators. This is precisely what Prandi's theory of conceptual conflict does and the reason why this article seeks to explain its main principles and their applicability to translation.

The central notion in Prandi's theoretical framework is that conceptual conflict arises in complex meanings presenting conceptual relations inconsistent with our shared system of conceptual presuppositions or natural ontology (Prandi, 2016: 73–81). A prototypical example of conflict in conceptual relations is Emily Brontë's poetic line "And winter pours its grief in snow" (2017: 23). The metaphor in this excerpt is inconsistent with our shared conceptual structures, since grief cannot be poured, nor can winter feel grief or dispose of it in the form of snow, nor can grief and snow be easily identified as the same thing. Thus, a conflictual meaning arises that cannot rely on shared uses of language to make sense, but that needs an act of interpretation based on co-text, context, and/or the communicative situation at hand. Under a cognitivist perspective, the conceptual metaphors EMOTIONS ARE LIQUIDS (WITHIN A PERSON) and SEASONS ARE PERSONS may be identified as lying at the origin of this example. These two mappings, however, are insufficient to fully unveil the figurative meaning of this line. EMOTIONS ARE LIQUIDS (WITHIN A PERSON) may explain quite well the import of an utterance such as "being overwhelmed by grief," through one of the many instantiations of the conceptual metaphor we share in our everyday use of language, representing grief as a mass of liquid submerging a person way beyond his or her capacity. But "pouring one's grief" is not a shared instantiation of the metaphor, since "pour" does not collocate with "grief" in the first place, which makes it conflictual with our shared representation of grief as a mass of liquid within the body or submerging the body. The personification of winter, moreover, may serve to present it as a human being performing actions or feeling emotions, but "pouring grief" is hardly a human action to be performed, nor does the phrasing "its grief" in this context unambiguously express an emotion felt by winter rather than an inner characteristic of this season. Grief is also said to be poured "in snow," which is not a consistent representation of snow under existing conceptualizations. These elements are inconsistent with our habitual conceptualizations of and presuppositions about them and from this inconsistency conflictual concept arises. Moreover, put together in this context, these elements provide a unique, complex expression that unleashes its figurative and creative potential through an additional interpretative effort on the part of the reader, which is something that with shared conceptualizations does not apply. Being aware of this is a valuable resource for the translator, both for the interpretative act triggered by the metaphor to make sense of it and for the creative act required to translate it in the target language: conflictual elements will have to be identified as such rather than referred back to pre-existing conventional elements, and solutions that render this conflict (rather than relying on conventional conceptualizations) will have to be found.

Conventional metaphors, on the other hand, do not feature conflict since they are consistent conceptual structures belonging to a shared heritage of everyday expressions, emerging from polysemy (Prandi, 2017: 23). An example of this is provided by "wasting time." The verb "waste" is polysemous and when used with "time" it appears in its metaphorical sense, consistent with the underlying metaphorical concept TIME IS MONEY. The latter concept is part of a shared way of representing time as a valuable resource, already existing in our vocabulary and system of communicative options. Hence, there is no conflict between the idea of "wasting" and the concept of "time." No particular interpretative effort is required to make sense of the phrase, since its meaning is already conventionalized in our shared linguistic background. In other words, to make sense of such metaphors, one only has to master shared conceptual structures and lexical systems. This has obvious consequences in translation: on the one hand, smaller effort is required to understand what is conventionalized in a shared lexical system, especially in the typical professional translation situation, with a non-native (although usually near-native) speaker of the source language translating into their native language. When concepts and lexis are already shared in a language, they will be immediately accessible to language users or retrievable through dictionaries, corpora, databases, and similar sources. The translator will be aware whether that concept is already available in their native language, and select the conceptual framework that is more apt to translate that metaphor in that context. As emphasized by

[1]Although many metaphor scholars, including, to mention but a few, Kovecses, Steen, Semino, and Gibbs have looked at living metaphors from a cognitive perspective, they are not analyzed here because the focus of this article is on the model provided by Lakoff and Johnson and on its unparalleled influence and persistence in approaches to analysis and translation.

the cognitivist tradition, conceptual metaphorical frameworks will often go beyond the boundaries of individual languages, making the translator's task theoretically simpler, especially when closely related languages are under analysis. Going back to Prandi's example TIME IS MONEY, it is true that IL TEMPO È DENARO provides a nearly identical conceptualization in Italian, so that its projections "wasting time" and "saving time," for example, will easily be translated, with the necessary adaptations to co-text, through "sprecare tempo" and "risparmiare tempo," respectively. "Spending time," on the other hand, is not effectively translated by "spendere tempo," which may be occasionally found in Italian usage, mainly as a calque from English. The shared Italian equivalent in use will be "passare/trascorrere del tempo," which does not pertain to the same conceptualization, and using "spendere tempo" in its place will sound marked and unconventional, if not simply awkward. This is a typical case of anisomorphism, which introduces an element of arbitrariness in the shared metaphorical motivation (Prandi, 2017: 186–188). A similar phenomenon warns us that little differences play an important role in a translation effectiveness and that attention to them should not decrease in view of metaphorical pattern similarities between source and target language.

Awareness of anisomorphism also points to the importance of the distinction between non-conflictual/conventional metaphors and conflictual/living metaphors in translation. With conventional metaphors, the metaphor lies in polysemy, that is, in lexical meaning, and the major difficulty lies in translating polysemy itself (as is the case with "spending time"). Conversely, with living metaphors, the metaphor is not translated—what is translated is conflict, and it is up to the reader to interpret it. Text examples in support of this claim are provided in the next section.

Prandi's model also explicitly highlights the importance of the pragmatic level of figurative interpretation in metaphors. In the presence of an extended sense of polysemous word—as in *"wasting"* time —metaphor is logically independent of interpretation, since the metaphorical meaning of "wasting" belongs to the shared lexis of English, and if it occurs in a sentence, it contributes to its complex meaning like any other lexical meaning. In the case of consistent metaphorical expression such as Dante's "In the middle of life's road/I found myself in a dark wood," the conceptual content of the metaphor is not the outcome of an act of interpretation, but part of our shared conceptualization of life as a journey. Hence, if only conventional metaphors are taken into consideration, there is no interpretative meaning specific to metaphor and its comprehension functions just like any other act of understanding, which is underlined by Sperber and Wilson (2008: 84). In the presence of conceptual conflict, on the other hand, the relationship between meaning, interpretation, figure, and message takes on a specific shape. This is due to two main reasons: firstly, a conflictual meaning lacks conceptual consistency, which imposes an act of figurative interpretation; secondly, with conflictual meanings the process of contingent interpretation not only connects a complex meaning and a message, but also gives shape and content to the figure. A living metaphor is not encapsulated within the meaning of the conflictual expression. A specific level of figurative interpretation becomes necessary, the relevance of which is explicitly excluded by Sperber and Wilson (2008) and (Prandi, 2017: 255–256). Conflictual metaphor may be open to many interpretative paths, not necessarily going in one, identifiable direction. A shared metaphorical concept may also be involved in its structure. The distinctive element, however, is that "the context plays the active role of urging the addressee to infer unexpected projections that go beyond conventional mappings. In such cases, the pressure of conventional coherence and relevance really turns into creative energy" (Prandi, 2017: 257). Thus, figurative interpretation is exclusive of conflictual metaphors and participates in their making, which distinguishes them from conventional metaphors and is extremely relevant to their translation, as the examples in the next section will hopefully clarify. These distinctions are major differences between Prandi's and CMT's views and shift the focus of attention from conventional metaphors to unconventional ones. Given the more complex nature of conflictual concepts as opposed to non-conflictual ones, further structural entities are required to explore conflict that go beyond the dual model of source and target domain as conceived by Lakoff and Johnson (1980). These notions provide an effective description of consistent concepts because they organize whole conceptual areas in conventional metaphors (Prandi, 2017: 29). For example, in the metaphorical concept LIFE IS A JOURNEY, "life" and "journey" are not confined to individual sentence level, but feature as labels for whole conceptual domains. The theory of conceptual conflict, on the other hand, extends and puts an emphasis on the analysis of living metaphors, that is, on individual complex expressions and complex meanings, which are characterized by further structure and thus require further notions for a thorough description. For these notions, Prandi draws on Black theory 1954—frame, focus, and subsidiary subject—and on Richards (1936)—tenor. *Frame* and *focus* are the immediate constituents of a conflictual complex meaning. The *frame* is characterized by its coherence with the ongoing text or discourse, whereas the *focus* introduces an incoherent and typically inconsistent concept: for example, in "the moon smiles," "the moon" is consistent with the communicative situation—the nocturnal setting in Blake's poem—whereas *smiles* is the strange, inconsistent element. Frame and focus are overt constituents of a more complex conceptual structure that also includes covert constituents. The latter are represented through the notions of *tenor* (Richards, 1936) and *subsidiary subject* (Black, 1954), which identify the possible covert counterparts of frame and focus, respectively. For example, in "the moon smiles," the focus "smiles" acts as a subsidiary subject for its covert tenor, say, "glittering," whereas the covert element emerging from the frame is the human being, which acts as a covert subsidiary subject on the tenor "the moon" in this specific metaphor (Prandi, 2017). Thus, implicit meanings are elicited, which should also be conveyed in translation.

Prandi's introduction of his conflict-based distinction between living and conventional metaphors is only a part of a plural consideration of metaphors which shows more pragmatic applications to translation than CMT. In the cognitive tradition, the singular term "metaphor" usually defines a wide range of figures and linguistic phenomena, including metaphors in

the strict sense of the word, obviously independent of their conventional or unconventional nature, as well as simile and even metonymy. This also emphasizes similarities among figures with different purposes, implications, effects, and translation outcomes. Prandi's model, on the other hand, detects differences among figures through an accurate typology (2004; 2010; 2012; 2015; 2017) and his monograph *Conceptual Conflicts in Metaphors and Figurative Language* (2017) specifically examines constitutive aspects of different figures and their different impact on meaning construction. For example, the different implications of the mechanism of analogy displayed by simile, as opposed to the workings of conceptual conflict in metaphor, are analyzed in detail, in open contrast with the cognitivist view of simile and metaphor as substantially the same phenomenon (Prandi, 2017: 166–170). This view is also reflected in the ingrained practice of translating metaphors as similes, which Newmark (1981) poses as one of the third best strategy out of seven to "solve the problem" of metaphor translation. Even if Newmark's decontextualized set of translation strategies may no longer be a strong point of reference in Translation Studies, the idea that simile may be a good substitute—and thus a good translation—for a metaphor is still widespread. A more thorough approach to metaphors and their distinction from simile and other figures makes it possible to investigate their linguistic features and communicative functions in text, with obvious benefits for translation, which is always a creative act of a specific text in a specific context.

A further effort in Prandi's taxonomy of conflictual metaphors is the notion of metaphorical swarm (2012: 157–166, 2017: 140–145). A metaphorical swarm is a network of interconnected metaphoric associations revolving around the same conflictual concept. The core conflictual concept generates a set of related conflictual expressions through the mechanism of projection. As Prandi suggests, projection "does not share the conflictual structure of the complex meaning that triggers it but can be completely accounted for from within the structure of consistent thought" (2017: 151). The main example of metaphorical swarm provided by Prandi is based on the conflictual concept LIGHT IS A LIQUID SUBSTANCE in Romantic literature: "if light is a liquid, it can *flow in rivers and streams*, *form waves*, *drops and waterfalls*, *ponds and lakes*, and so on" (2017: 143). This example illustrates how projection may apply to a number of interconnected inconsistent expressions, "each of which frames in words one node of the complex conceptual network projected by the seminal conflictual expression" (2017: 143). *Swarm* seems to be the perfect term to encapsulate such a constellation of expressions, since it suggests exactly the right inferences: "unpredictability of time, location and size; high mobility; and uneven density" (2012: 158, 2017: 144).

DISCUSSION OF TEXT EXAMPLES

In this section, the applicability of the conceptual conflict model to a pragmatic view of text analysis and translation will be explored through literary text excerpts. These notions, however, are not specific to literature, but apply to any text type, as Prandi's applications to the realm of science have demonstrated (Prandi, 2013). The first example is from Sidney's *The Countess of Pembroke's Arcadia*, where the character Dametas is said to be "muttering and **champing** as though his **cud** troubled him" (Sidney, 1973: 44). "Champing" indicates the act of noisy chewing performed by animals, and is the inconsistent element representing our focus. Together with "muttering" it reinforces the idea he is grumbling his complaints in an irate fashion, adding a beastly flavor to it. Such an animal element is confirmed by "cud," which points to the half-digested food in ruminants, and has no other conventional meaning that could be consistent with the ongoing text. Thus, thanks to co-textual elements, covert elements emerge: behind "champing," which is both focus and subsidiary subject, is a covert tenor- Dametas's beastly ruminating of thoughts and worries—and behind the tenor Dametas is the subsidiary subject of a ruminant. This interpretation is not only consistent with immediate co-text, but also with the wider context in which Dametas is repeatedly referred to as coarse and uncouth, and such a context reinforces it. It should be noted that an analysis based on source and target domain only would explain this as an instantiation of the THOUGHT IS FOOD conceptual metaphor, but that would provide no access to the covert meanings responsible for the effects pursued by text. Identifying the covert elements in text is very important to attempt a translation that preserves these complex meaning relations as well as the comic intention at the heart of this passage. A few major problems, however, arise due to anisomorphism in English and Italian: there is no Italian verb for "champing" that is specific to animals, and thus may act as a focus, apart from "ruminare" which has a slightly different meaning, and also the conventional figurative meaning of "ruminating" intended as "thinking"; neither "cud" has any equivalents that are specific to animals; and translating "troubled" requires a more specific solution in Italian. After long research, I would propose this version: "mugugnando e ruminando come se il fieno gli fosse rimasto sullo stomaco," which may be backtranslated as "muttering and ruminating as if his fodder lay heavy on his stomach," where "fieno/fodder" makes reference to his animal nature unambiguous, thus selecting the animal—therefore conflictual—meaning of "ruminare."

The accurate description of conflictual metaphorical structures provided by Prandi through the notions of frame, focus, tenor, and subsidiary subject takes into account aspects of fundamental importance for a pragmatic view of translation, such as implicit meaning, context-relatedness, the salience of world knowledge and inherent complexity. These factors all play a part in making a living metaphor an individual act of linguistic creation of something new, escaping the boundaries of linguistic conventions. In this sense, conceptual conflict is closely related to the mechanisms of conceptual creativity made possible by linguistic expressions, and represents a major resource for expressing innovative contents and projecting new interpretative frameworks onto real-world scenarios.

Innovative concepts, however, may find resistance on their way to translation, and sometimes the problem does not lie in anisomorphism or any other interlinguistic issue, at least not directly. An interesting example concerns the nocturnal setting depicted in William Trevor's short story "Bravado:"

> Not many people were about; it was after midnight, almost one o'clock, the widely spaced lampposts casting *pools of misty yellow illumination*. A man walked his dog in Blenning Road in the same blotchy lamplight, the first of autumn's leaves gathering there also. (Trevor, 2004: 73)

"Blotchy" is a strange, inconsistent element as it describes the lamplight as a liquid. The same conflictual concept also features in "pools of misty yellow illumination," which precedes and prepares "the same blotchy lamplight," which is a more marked choice than "pools" to project the characteristics of liquids onto light. This marked representation of light is strictly linked to this specific context, as the lamplight appears "blotchy" precisely because of the "widely spaced lampposts" which cast a patchy illumination rather than a uniform one.

> Non c'era molta gente in giro; era dopo mezzanotte, quasi l'una, e i radi lampioni proiettavano *pozze di fioca luce gialla*. Un uomo portava a passaggio il cane in Blenning Road sotto la stessa luce pallida, e anche lì si ammucchiavano le prime foglie autunnali. (Trevor, 2009: 64)

The Italian translation maintains the liquid light metaphor in the case of "pools," "pozze di fioca luce gialla," where it is interesting to note that misty is rendered as "fioca/dim," — which is one of the meanings expressed by "misty." Blotchy, on the other hand, is rendered as "pallida/pale," which appears to be context-driven as a synonym for "fioca/pale." This way, however, it does not translate the idea of liquid light forming blotches, which is the most marked—and creative—element in the description of this urban landscape; rather, it erases this outstanding element and replaces it with a substitute simply repeating the meaning expressed by "misty/fioca." Thus, no conflictual meaning is suggested and consistency is restored at the expense of the peculiar visual input suggested by the metaphorical focus in context. A reason for this choice might lie in the fact that solutions that are as marked in the target text will also stand out and sound "strange," which may cause a revisor to edit it, or a translator to erase it in the first place to avoid revision. A possible solution features *in quelle stesse chiazze di luce* (which, in back-translation, would read *in the same blotches of light*). This version would foreground the creative image of the "blotches of lamplight," motivated by the wide spacing of the lampposts mentioned in the previous sentence, producing a patchy light. This interpretation is reinforced by the mention of *leaves gathering there also*, which refers to leaves lying on the pavement within the perimeter of the "blotches" receiving illumination—whereas "luce pallida/pale light" does not seem to draw attention to the pavement, but to the whole space illuminated by the lampposts.

A similar phenomenon characterizes the translation of a passage from the novel *Abela* by a final year student on the MA in Translation and Interpreting at the University of Genoa where I teach:

> Suddenly the boys set up an excited shouting, waving their arms and skipping, as *the cloud of noisy red dust that was a bus came bumping* toward them. (Doherty, 2007).

> Improvvisamente i ragazzi iniziarono a urlare eccitati, agitando le braccia e saltando, mentre *una vistosa nuvola di polvere rossa, che proveniva da un autobus, si muoveva* verso di loro.

In this translation, "the cloud of noisy red dust that was a bus" is rendered, in back-translation, as "a flashy/huge cloud of red dust, which came from a bus." This translation offers a simplified interpretation of this figure, deliberately modifying the defining clause after the figure "that was a bus." Moreover, it turns "noisy" into "vistosa/flashy or huge," which cancels the synesthetic reference triggered by the sound adjective associated with visual input, by suggesting a consistent association of visual elements only. The conflict implied in the representation of a bus as a bumpy and noisy cloud of dust is leveled down, and this leveling is reinforced by the translation of "came bumping" with "si muoveva/moved." Again, conflict is underrepresented, although Italian is certainly equipped with the linguistic material to provide a suitable representation of it.

The phenomenon of diminishing marked and innovative elements in translation is well-documented in the literature, and is usually related to a position of inferiority of the translator, to the so-called translator's invisibility, and to publishers' determination to play down on politically conflictual issues (Venuti, 2008). The latter two examples, however, do not seem to justify this explanation, since they do not contain reference to political or sensitive issues. Nor does the linguistic material used present major problems such as wordplay or untranslatable culture-bound elements. The conceptual conflicts in these examples, however, lack conceptual consistency, which is a preliminary condition for a test of coherence and therefore calls for a figurative interpretation (Prandi, 2017: 256). Their translations refuse that lack of consistency and reach coherence by omitting the conflictual elements in favor of a literal, non-figurative solution. This reflects the common habit of considering conflictual concept as a mistake to be corrected rather than an instance of creativity, which has dominated most of twentieth-century—especially Chomsky's—linguistics and is also well-rooted in public opinion. A model of metaphor analysis that takes conflictual concept into account and explains how it works in terms of constructing goal-oriented effects such as Prandi's may well provide a tool to contrast this tendency and a solid basis for the text analysis needed to render conflictual metaphors effectively.

Identifying a metaphorical swarm also provides an asset in the linguistic analysis aimed at translation, since it makes an important textual feature evident, enabling the translator to recognize the network of figurative elements that may be scattered over longer stretches of text, thus paving the way to recreating text objectives and effects in a credible way. In the following dialogue from *The Two Gentlemen of Verona*[2], for example, a metaphorical swarm emerges in the dialogue

[2]The Italian translation provided here is the author's work (Shakespeare, 2015). This passage has already been analyzed in two previous writings by the author, dealing with the translation of the play (Rizzato, 2019) and with the specific functions associated with metaphorical swarms in the source text and in translation (Rizzato, 2021).

between Valentine who, on his departure for Milan, addresses his humorous reproaches to Proteus, and Proteus himself, who refuses to leave Verona to pursue his love for Julia:

> PROTEUS [. . .] For I will be thy beadsman,Valentine.
> VALENTINE And on a love-book pray for my success?
> PROTEUS Upon some book I love I'll pray for thee.
> VALENTINE That's on some shallow story of deep love-
> How young Leander crossed the Hellespont.
> PROTEUS That's a deep story of a deeper love,
> For he was more than over-shoes in love.
> VALENTINE Tis true,for you are over-boots in love,
> And yet you never swam the Hellespont.
> PROTEUS Over the boots? "Nay,give me not the boots".
> VALENTINE No, I will not; for it boots thee not.
> PROTEUS What?
> VALENTINE To be in love [. . .] (Shakespeare, 2005:s3)

In this scene, Valentine introduces the myth of Leander, who would cross the Hellespont every night in order to visit his lover Hero, by comparing his love to Proteus's. This myth offers an opportunity to represent love as a liquid, through the opposition, among others, of "deep" and "shallow," referring to "love" and "story," but also evocative of the Hellespont waters, which makes the metaphorical swarm based on the metaphor LOVE IS WATER emerge. The representation of love as a substance is by no means a novelty, like the locative expression "to be in love" presupposes. Here, however, context and reference to myth represent love in a conflictual way, as water, which is unique to this text. Moreover, a network of metaphorical foci consistent with the central metaphorical concept is constructed. An important part of it makes use of the idiomatic expression "over shoes, over boots," representing a person as immersed with one's feet into some substance (presumably water or mud), metaphorically meaning "expressing reckless continuance in a course already begun" (Anon, 1933: 996). This expression, no longer comprehensible to today's speakers of English, is related to the locative phrase "in love" twice, as it is not used in its canonical form, "over shoes, over boots," but is split into two, so that "over-shoes in love" describes Leander, whereas "over-boots in love" is associated with Proteus, which seems to condemn Proteus to an even more desperate condition than Leander's. This creative use of the "split idiom" also elicits the exploitation of the domain of footwear to construct projections of the LOVE IS WATER metaphorical swarm providing humorous connection for each character's line in the remaining part of the dialogue. "Boots" is in fact reprised in Proteus' line "Nay, give me not the boots"—an idiom meaning "don't make a laughing stock of me" (Shakespeare, 1969: 4)—and then in Valentine's reply, "it boots thee not," this time as a verb, an instance of the multiple uses of polysemy characterizing the texture of much of Shakespeare's plays.

Awareness of the swarm of metaphorical expressions and the functions these metaphors have, both in isolation and as an ensamble, may represent a first step toward recreating them in translation. A few difficulties arise in this attempt. First, in Italian "in love" is not translated by an equivalent prepositional phrase, but by the past participle "innamorato," the locative reference of which is much less transparent. Hence, in order to suggest the idea of love as a substance in which the lover may swim, the translation uses "immerso nell'amore" (immersed in love), which makes the idea of a liquid or a fluid explicit. Second, the idiomatic expression "over-shoes, over-boots" has no immediate equivalent in Italian. No doubt there are translations for the propositional meaning of the expression, but they may not be suitable for constructing a meaningful sentence, nor are they connected with the domain of shoes and boots evoked by the source text and also represented in the following lines. The Italian translation should ideally draw from the same domain and convey "give me not the boots" with an idiomatic expression of the same meaning containing reference to footwear, and "it boots thee not" with a verb semantically related to shoes or boots of the same meaning. Thus, the connection among each character's utterances would be equivalent to that of the source text, and the figurative pattern established in the source text reproduced in the target text. If you consider the single phrases in isolation, solutions may be available presenting near perfect isomorphism with the source text expressions. For example, "calzare" (to fit and, by extension, to be apt) as a translation of the verb "boot" evokes the domain of footwear and means "to suit perfectly," which could work in this context. Similarly, "esserci dentro con tutte le scarpe" (to be into something over the shoes) is an idiomatic form including reference to shoes (scarpe), which could serve well the purpose of translating "he was over-shoes in love" and, with some adaptations, "you are over-boots in love" and the following "Over the boots?". And here is where the third problem arises: there seems to be no equivalent for "give me not the boots" in Italian with a similar meaning and a translation of "boots" or some other item of footwear as a focus in the metaphor. Therefore, the penultimate ring in the chain of figures in the same swarm is broken, and the effect of the whole sequence is put at risk. Evidence of this problem emerges from Perosa's translation, where the LOVE IS WATER swarm ends with the translation of "Over the boots?", and the next two metaphorical foci ("boots" as a noun and then as a verb) are replaced with a pun based on the stem of the verb "dire[3]", which is unrelated to the preceding exchange, thus diminishing coherence in the dialogue.

Bompiani translation (Shakespeare, 2015) tries to prevent this by identifying an alternative domain to that of footwear that may translate the pair "over-shoes. . . over-boots" and also cover the utterances "over the boots," "give me not the boots" and "it boots thee not," while remaining consistent with the LOVE IS WATER swarm:

> PROTEO [. . .] perché io, Valentino, intercederò per te.
> VALENTINO Pregando per me su un libro d'amore?
> PROTEO Pregando per te su un libro che amo.
> VALENTINO Ossia sulla storia superficiale di un amore profondo, come quella del giovane Leandro che attraversò l'Ellesponto.
> PROTEO Quella è la storia profonda di un amore ancora più profondo, tant'è che Leandro si immerse nell'amore fino al collo.

[3]PROTEO: No, non mi *dire*./VALENTINO E io non lo *dirò*: non ti si *addice*. (Shakespeare, 1990, p. 419) PROTEUS: Don't even tell me that!/VALENTINE: I won't tell: it doesn't suit you.

VALENTINO Vero, e tu vi sei immerso fino al naso, anche se non ti sei mai bagnato nell'Ellesponto.
PROTEO Il naso? Sei tu a non dovermi prendere per il naso!
VALENTINO No, no. A fiuto, direi che non fa per te.
PROTEO Che cosa?
VALENTINO L'amore [. . .] (Shakespeare, 2015: 25–27)

A source very productive domain in terms of idioms and polysemy was required in Italian that could cover the five footwear-related items in the most similar way possible—translating, for example, an idiom for an idiom, a polysemous lexeme with an equally polysemous one, and so on. In this context, existing metaphor research was taken into consideration and applied to the search for appropriate solutions, so that the domain of the human body, considered the major source domain for conceptual metaphors (Kövecses, 2010: 18), was identified and actually used in the passage to construct a very similar figurative pattern in the target language[4].

Thus, "fino al collo" (up to one's neck) was selected for "over-shoes," "fino al naso" (up to one's nose) for "over-boots," and "prendere per il naso" (pull somebody by one's nose, metaphorically meaning "making a laughing stock fun of somebody") for "giving somebody the boots." The element of the nose was then reprised in the next utterance by "a fiuto" (according to one's sense of smell), which is also idiomatic. Here, it modalises the target text equivalent of "it boots thee not," "non fa per te," which translates its propositional meaning without including any metaphors of the body. Thus, a number of adjustments were made necessary by the unavailability of Italian immediate equivalents, but the figurative pattern suggested by the metaphorical swarm in the source text was preserved, creating a similar pace in the translation, and making sense of each line in this further sparring match in a lively manner.

This version does not solve all problems. For example, "fino al naso" is neither part of a lexicalized expression together with "fino al collo," nor is it as idiomatic, but it serves the purpose of constructing a comprehensible crescendo structure relying on the same domain. In addition, in the final part of the passage, a modaliser ("a fiuto") is added in the target text to justify the presence of a meaningful metaphorical element from the desired source domain. Moreover, the ridiculous, low-brow connotations of shoes and boots, especially when referred to the noble feeling of love, are not completely translated by the notions of nose and neck. The main objectives pursued by the metaphorical swarm in the source text, however, are at least partially achieved in the target dialogue, where making the figurative texture evident conveys extra meaning, sustains dramatic progression and strongly contributes to comic effects.

[4]Perosa had already used the source domain of the human body to translate the first three footwear-related items, but not the remaining ones, thus altering the structure of the swarm completely, whereas Bompiani translation attempts to preserve such structure in the target text as a fundamental item in meaning-making.

CONCLUSION

Translation is an inherently pragmatic activity, since it needs to tackle context- and situation-related issues and pay attention to minimal cross-linguistic differences which are unique and specific to the text under analysis. In metaphor translation, the model for metaphor analysis provided by CMT does not seem to offer the same range of tools for translation as Prandi's model, since it emphasizes universal and conventional aspects of metaphors, while translation difficulties are usually due to linguistic difference, anisomorphism and unconventional meanings, as the passages analyzed in this article were meant to demonstrate. Prandi's theory of conceptual conflict seems to provide a better analytic model, since it identifies significant distinctions to be realized in translation, and puts an emphasis on living metaphors, revolving around conceptual conflict and unconventional elements. The notion and structure of conflictual concepts offers more specific resources for text analysis and translation than the mapping of source and target domains. In addition, pragmatic aspects specific to figurative interpretation are illustrated that distinguish conflictual metaphors from conventional ones and make readers, analysists and translators aware of the peculiar nature of figurative meaning. Prandi's model also challenges the view of "metaphor" as one phenomenon to encompass other figures, their differentiated objectives and effects. Within this plural consideration of metaphors, the metaphorical swarm stands out as a useful tool to break down complex metaphorical networks and explain their functions, as illustrated by the excerpt from *The Two Gentlemen of Verona*. In conclusion, the adoption of the conflictual concept model for the text analysis preparing translation allows for a more detailed, text- and context-specific notion of metaphors, which offers very useful insights for choosing the translation strategy to be applied and paves the way for an effective rendering of the source text from a pragmatic point of view.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

FUNDING

This article was published with the contribution of the PRIN research project *New perspectives in metaphor research* (2017–2020), 2015 Call, Grant No. 2015YHYWSH.

REFERENCES

Anon (1933). *Oxford English Dictionary*. Oxford: Oxford University Press.

Baker, M. (2011). *In Other Words: A Coursebook on Translation*. London: Routledge.

Black, M. (1954). Metaphor. *New Series* 55, 273–294.

Doherty, B. (2007). *Abela: The Girl Who Saw Lions*. London: Andersen Press.

House, J. (1997). *Translation Quality Assessment. A Model Revisited*. Tübingen: Gunter Narr Verlag.

Kövecses, Z. (2010). *Metaphor: A Practical Introduction*. Oxford: Oxford University Press.

Lakoff, G., and Johnson, M. (1980). *Metaphors we Live by*. Chicago: University of Chicago Press.

Lakoff, G., and Turner, M. (1989). *More Than Cool Reason: A Field Guide to Poetic Metaphor*. Chicago: University of Chicago Press.

Newmark, P. (1981). *Approaches to Translation*. Oxford: Pergamon Press.

Prandi, M. (2004). *The Building Blocks of Meaning: Ideas for a Philosophical Grammar*. Amsterdam: Benjamins.

Prandi, M. (2010). Typology of metaphors: implications for translation. *Mutatis Mutandis* 3, 304–332.

Prandi, M. (2012). A plea for living metaphors: conflictual metaphors and metaphorical swarms. *Met. Symbol* 27, 148–170. doi: 10.1080/10926488.2012.667690

Prandi, M. (2013). "Le metafore nella creazione di terminologia: una tipologia ragionata," in *Il ruolo della metafora nella creazione di terminologie*, eds M. Prandi, A. Giaufret, and M. Rossi (Genova: Genova University Press), 25–39.

Prandi, M. (2015). "Translating metaphors," in *Language Across Languages: New Perspectives on Translation*, eds E. Miola and P. Ramat (Newcastle upon Tyne: Cambridge Scholars Publishing), 83–104.

Prandi, M. (2016). Selection restrictions as ultimate presuppositions of natural ontology. *Topoi* 35, 73–81. doi: 10.1007/s11245-014-9280-5

Prandi, M. (2017). *Conceptual Conflicts in Metaphors and Figurative Language*. New York, NY: Routledge.

Richards, I. A. (1936). *The Philosophy of Rhetoric*. Oxford: Oxford University Press.

Rizzato, I. (2019). "Translating figurative language in Shakespeare's *The Two Gentlemen of Verona*," in Metaphor *and Translation*, eds R. Trim and D. Sliwa (Newcastle upon Tyne: Cambridge Scholars Publishing), 145–156.

Rizzato, I. (2021). "Shakespeare's metaphorical swarms: text functions in The Two Gentlemen of Verona and implications for its translation into Italian," in *Metaphor and Conflict / Métaphore et Conflit*, eds P. Paissa, M. Conoscenti, R. Druetta, and M. Solly (Bern: Peter Lang), 339–357.

Schäffner, C. (2004). Metaphor and translation: some implications of a cognitive approach. *J. Pragmatics* 36, 1253–1269. doi: 10.1016/j.pragma.2003.10.012

Shakespeare, W. (1969). *The Two Gentlemen of Verona,* ed C. Leech (London: Methuen).

Shakespeare, W. (1990). *I due gentiluomini di Verona, Teatro Completo, Vol. I, Le Commedie Eufuistiche*, ed G. Melchiori (Milano: Mondadori), 401–587.

Shakespeare, W. (2005). *The Two Gentlemen of Verona. William Shakespeare. The Complete Works*, eds S. Wells, G. Taylor, J. Jowett, and W. Montgomery (Oxford: Oxford University Press), 1–24.

Shakespeare, W. (2015). *I due gentiluomini di Verona. Tutte le Opere di William Shakespeare, Vol. II, Commedie*, ed F. Marenco (Milano: Bompiani), 3–199.

Sidney, P. (1973). *The Countess of Pembroke Arcadia (The Old Arcadia)*. Oxford: Clarendon Press.

Snell-Hornby, M. (1995). *Translation Studies. An Integrated Approach*. Amsterdam: John Benjamins.

Snell-Hornby, M. (2006). *The Turns of Translation Studies. New Paradigms or Shifting Viewpoints?* Amsterdam: John Benjamins.

Sperber, D., and Wilson, D. (2008). "A deflationary account of metaphor," in *The Cambridge Handbook of Metaphor and Thought*, ed R. W. Gibbs (Cambridge: Cambridge University Press), 84–105.

Trevor, W. (2004). *Cheating at Canasta*. New York, NY: Viking.

Trevor, W. (2009). *Uomini d'Irlanda*. Parma: Guanda.

Venuti, L. (2008). *The Translator's Invisibility. A History of Translation*. London: Routledge.

Conflict of Interest: 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 © 2021 Rizzato. 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) and the copyright owner(s) 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: 06 May 2021
doi: 10.3389/fpsyg.2021.635724

The Role of Working Memory in the Processing of Scalar Implicatures of Patients With Schizophrenia Spectrum and Other Psychotic Disorders

Walter Schaeken[1], Linde Van de Weyer[1], Marc De Hert[2,3] and Martien Wampers[2]*

[1] Brain and Cognition, Faculty of Psychology and Educational Sciences, KU Leuven, Leuven, Belgium, [2] University Psychiatric Center KU Leuven, Leuven, Belgium, [3] Center for Clinical Psychiatry, Department of Neurosciences Antwerp Health Law and Ethics Chair, University of Antwerp, Antwerp, Belgium

A number of studies have demonstrated pragmatic language difficulties in people with Schizophrenia Spectrum and Other Psychotic Disorders. However, research about how people with schizophrenia spectrum and other psychotic disorders understand scalar implicatures (SIs) is surprisingly rare, since SIs have generated much of the most recent literature. Scalar implicatures are pragmatic inferences, based on linguistic expressions like *some*, *must*, *or*, which are part of a scale of informativeness (e.g., *some/many/all*). Logically, the less informative expressions imply the more informative ones, but pragmatically people usually infer that the presence of a less informative term implies that the more informative term was not applicable. In one of the few existing studies with people with schizophrenia spectrum and other psychotic disorders, Wampers et al. (2018) observed that in general, people with schizophrenia spectrum and other psychotic disorders were less likely to derive SIs than controls. The current study has three main aims. First, we want to replicate the original finding with the scalar terms *some-all*. Second, we want to investigate how these patients deal with different scalar terms, that is, we want to investigate if scalar diversity is also observed in this clinical group. Third, we investigate the role of working memory, often seen as another important mechanism to enable inferring SIs. Twenty-one individuals with a psychotic disorder and 21 matched controls answered 54 under-informative statements, in which seven different pairs of scalar terms were used. In addition, working memory capacity was measured. Patients with schizophrenia spectrum and other psychotic disorders did not make more logical interpretations when processing quantifiers, disconfirming Wampers et al. (2018). However, certain scalar scales elicited more pragmatic interpretations than others, which is in line with the scalar diversity hypothesis. Additionally, we observed only partial evidence for the role of working memory. Only for the scalar scale *and-or*, a significant effect of working memory was observed. The implications of these results for patients with schizophrenia spectrum and other psychotic disorders are discussed, but also the role of working memory for pragmatic inferences, as well as the place of SIs in experimental pragmatics.

Keywords: pragmatics, experimental pragmatics, scalar implicature, schizophrenia spectrum and other psychotic disorders, working memory

OPEN ACCESS

Edited by:
Valentina Cuccio, University of Messina, Italy

Reviewed by:
Mikhail Kissine, Université libre de Bruxelles, Belgium
Valentina Cardella, University of Messina, Italy
Anna Papafragou, University of Pennsylvania, United States

****Correspondence:***
Walter Schaeken
walter.schaeken@kuleuven.be

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

Received: *30 November 2020*
Accepted: *06 April 2021*
Published: *06 May 2021*

Citation:
Schaeken W, Van de Weyer L, De Hert M and Wampers M (2021) The Role of Working Memory in the Processing of Scalar Implicatures of Patients With Schizophrenia Spectrum and Other Psychotic Disorders. Front. Psychol. 12:635724. doi: 10.3389/fpsyg.2021.635724

INTRODUCTION

Pragmatics is the study of the use of language in context, whereby one of the key findings is that the meaning of words and sentences can change in the light of the specific context they are used. Important in this domain is the distinction between the literal and the intended meaning (in Grice, 1975). The distinction between literal and intended meaning can be experimentally studied using for instance metaphors, humor, or irony (for a recent overview, see Noveck, 2018). Other linguistic expressions which are widely researched in experimental pragmatics are Scalar Implicatures (SIs). Expressions like *some-many-all may-should-must*, *warm-hot* are part of a scale organized by informativity (Horn, 1972). These scales have a specific characteristic: when the stronger term holds, the weaker term also holds, while the opposite is not true. Consider a professor correcting essays who tells her partner in the evening: "*Some of the essays were thought-provoking*." This expression is true when the professor found all essays thought-provoking and when she did find only some of them thought-provoking but not all. However, in case the professor only found some and not all of the essays thought-provoking, she could not say "*All of the essays were thought-provoking*" if she did not want to lie. One can notice here already the Gricean distinction (1975) between what is said on the one hand (the specific sentence) and what is implicated on the other hand (the speaker's meaning). An implicature is a component of the speaker's meaning, which is not said and therefore should be inferred. How do listeners make the required inference in our first example according to Grice? First of all, listeners adhere to the Cooperation principle and assume that a speaker is trying to be cooperative. More specifically, and translating this to the Gricean maxim of Quantity, a listener expects that the speaker was as informative as possible and also that she gave as much information as was needed (and also not more information than needed). Since our professor used the weaker term of a scale (e.g., *some essays were thought-provoking*) and not the stronger term (e.g., *all essays were thought-provoking*), the listener can infer that the professor found that the stronger term was not appropriate, otherwise she would have used it. Consequently, the listener can infer that, or stated differently, the listener enriches the original expression to the upper-bounded meaning with "*some and not all of the essays were thought-provoking*." Noveck (2001) refers to this line of reasoning as the preference for the pragmatic interpretation above the logical one. Experimental research, often only focusing on *all-some*, clearly demonstrates that adults predominantly prefer the pragmatic upper-bounded *some but not all* interpretation of *some* (e.g., Noveck, 2001; Bott and Noveck, 2004; De Neys and Schaeken, 2007; Marty and Chemla, 2013; Heyman and Schaeken, 2015; van Tiel and Schaeken, 2017). Children, however, prefer more often than adults the logical meaning of *some* which is also compatible with *all* (see e.g., Chierchia et al., 2001; Noveck, 2001; Papafragou and Musolino, 2003; Foppolo et al., 2012; Janssens et al., 2014, 2015; Schaeken et al., 2019), although more adult-like behavior can be elicited (e.g., Papafragou and Musolino, 2003; Pouscoulous et al., 2007; Barner et al., 2011; Katsos and Bishop, 2011; Schaeken et al., 2019).

It is clear that there is abundant experimental research on SIs, and it is sometimes said that *some-all* expressions stand as the poster child of pragmatic inference (Scontras et al., 2018). There are several reasons for the special status of these expressions in experimental pragmatics: the context and content of such expressions are simple to manipulate; potential confounding variables are easy to control; competing theories make clear predictions about these experimental manipulations; different age-groups be tested with similar paradigms (see e.g., Noveck and Sperber, 2007; Katsos and Cummins, 2010). Therefore, it is surprising that few researchers have addressed the issue how clinical populations deal with SIs.

The current study wants to fill this gap in knowledge, linking understanding SIs by people with schizophrenia spectrum and other psychotic disorders with scalar diversity and the role of working memory. The remaining part of this introduction is structured in the following way. First, we will describe pragmatic difficulties of people with schizophrenia spectrum and other psychotic disorders. Next, we will discuss some findings with respect to SIs with clinical populations, more specifically people with Autism Spectrum Disorder (ASD). Then, we will discuss briefly the potential role of Theory of Mind (ToM) and more extensively the role of working memory. Finally, we will introduce the issue of scalar diversity and describe our own research.

According to the American Psychiatric Manual (APA 5) Schizophrenia Spectrum and Other Psychotic Disorders is characterized by for instance delusions, hallucinations, but also by disorganized speech and other symptoms that cause social or occupational dysfunction. Language impairments have always been observed in people with schizophrenia spectrum and other psychotic disorders. Extensive research showed that patients diagnosed with schizophrenia spectrum and other psychotic disorders have difficulties when they have to deal with non-literal expressions or the non-literal parts of expressions. Studies focusing on one or a few aspects, like humor, irony, metaphors, proverbs, . . . showed that these aspects are all difficult to understand for patients with schizophrenia spectrum and other psychotic disorders (see e.g., Langdon et al., 2002a,b; Sponheim et al., 2003; Brüne and Bodenstein, 2005; Linscott, 2005; Bambini et al., 2016, 2020; for an overview, see e.g., Bosco and Parola, 2017). Also, broad assessments of the pragmatic competence showed a deficit in patients with schizophrenia spectrum and other psychotic disorders. Colle et al. (2013) found evidence of a wide dysfunction using the Assessment Battery of Communication; Bambini et al. (2016) showed, using the APACS Test for the assessment of pragmatic abilities and cognitive substrates that the pragmatic abilities were weakened in schizophrenia, with comprehending discourse and non-literal meanings being especially compromised. Bambini et al. (2016) even argue that the high frequency of impairment suggests that the pragmatic deficit is a core feature of schizophrenia. The latter study also evidenced the role of pragmatics for quality of life: overall pragmatic qualities predicted quality of life, while this was not the case for other cognitive variables in their study. In the same line, Adamczyk et al. (2016) showed that selective language and communication skills (inferential meaning, humor

and metaphors, emotional and linguistic prosody) are important for patients with schizophrenia in their social recovery process. Agostoni et al. (2021) show through a mediation analysis that pragmatics has both a direct and an indirect effect on daily functioning, and especially in interpersonal functioning. In other words, recent evidence not only points to pragmatic difficulties in people with schizophrenia spectrum and other psychotic disorders, but also to the important role pragmatics might play for their daily functioning.

As said before, SIs is one of the most widely researched topics in pragmatic with adults and typically developing children. However, research on clinical populations is unexpectedly scarce, with as main exception research about the understanding of SIs by people with ASD. In many of these studies no decrease in pragmatic responses was observed (see e.g., Pijnacker et al., 2009; Chevallier et al., 2010; Su and Su, 2015; Hochstein et al., 2017; see also Antoniou et al., 2016 and Heyman and Schaeken, 2015 for similar findings with participants with higher Autism-Spectrum Quotient scores). However, some other studies did observe differences, albeit sometimes subtle (see e.g., Nieuwland et al., 2010; Zhao et al., 2015; Mazzaggio and Surion, 2018; Schaeken et al., 2018).

Regarding people with schizophrenia spectrum and other psychotic disorders, there is to the best of our knowledge only one published study on the understanding of SIs. Wampers et al. (2018) observed in both a binary and ternary statement-evaluation-task with *some-all* that patients with schizophrenia spectrum and other psychotic disorders derived fewer SIs than matched control participants. Moreover, ToM abilities were positively correlated with deriving SIs.

This significant correlation between ToM abilities and the derivation of SIs in Wampers et al. (2018) added to the mixed evidence on the role of ToM for pragmatics. ToM skills are often seen as an important driver of pragmatic comprehension. Support for this claim comes from work with typically developing adults, for instance showing an important involvement of ToM skills in irony (e.g., Andrés-Roqueta and Katsos, 2017). Even more evidence come from work with clinical populations, like ASD and schizophrenia spectrum and other psychotic disorders, where a relation between the ToM deficit and their difficulty in pragmatics is observed (see e.g., Baron-Cohen, 1988; Happé, 1993; Corcoran et al., 1995, 1997; Langdon et al., 2002a; Janssen et al., 2003; Brüne and Bodenstein, 2005), or from work where impaired pragmatic reasoning is observed in patients with cortical lesions to ToM areas (Champagne-Lavau and Joanette, 2009; Spotorno et al., 2015). However, this picture is far from consistent. Some authors (e.g., Langdon et al., 2002a; Mazza et al., 2008) observed for instance a role of ToM for irony, but not for metaphors, while others (e.g., Brüne and Bodenstein, 2005; Mo et al., 2008) observed the opposite pattern. Similarly, sometimes effects were already observed for first-order ToM (see Happé, 1993), sometimes only for second-order ToM (see Mo et al., 2008; Panzeri and Foppolo, 2016). Finally, the picture of ToM for schizophrenia spectrum and other psychotic disorders is even more complicated, since ToM is not impaired in paranoid schizophrenia, which is often characterized by a hyper-ToM (Abu-Akel and Bailey, 2000; Frith, 2004; Peyroux et al., 2019).

Similarly, the role of ToM for scalars is debatable. Pijnacker et al. (2009) argue that the ToM load for scalars is low: A SI seems to require only first order mental states (e.g., *she knows, or she does not know that*. . .) and not second-order mental states (e.g., *she does not know that he knows that*. . .). Therefore, it is possible that just basic ToM skills are already sufficient for inferring SIs (see also Chevallier et al., 2010). Brüne (2003) argues that the comprehension of this first order ToM is relatively preserved in patients with schizophrenia spectrum and other psychotic disorders. Moreover, Andrés-Roqueta and Katsos (2017) argue that especially in work with children the knowledge needed to infer the SIs is often visually accessible, therefore minimizing the demand on ToM.

One possibility mentioned by Wampers et al. (2018) to explain their observed ToM-effect, is the potential role of working memory. Both ToM and working memory are considered as cognitive substrates underlying pragmatic competence (see e.g., Cummings, 2017). However, they are not independent of each other, since for instance working memory capacity is required to be able to think about other persons thoughts. The role of working memory for pragmatic language understanding is widely discussed. For instance, Chiappe and Chiappe (2007) and Columbus et al. (2015) showed the important role of working memory in metaphor comprehension for young adults. Bambini et al. (2021) observed that working memory skills were crucial for the pragmatic skills tested (comprehension of oral narrative stories, humor, figurative language and implicatures). Flexibility played a role for figurative language and implicatures, while, surprisingly, inhibition was not a robust predictor. Also for clinical populations and definitely also for patients with Schizophrenia Spectrum and Other Psychotic Disorders, an important role of working memory is observed (see Forbes et al., 2009), also in pragmatics. For instance, Bosia et al. (2016) observed a significant correlation of working memory with pragmatic production, Kiang et al. (2007) with comprehension of proverbs and Schettino et al. (2010) with idiomatic expressions. Moreover, the role of working memory in the production of SIs is a popular research topic and a vast amount of experiments evidenced a processing cost associated with processing SIs. When given less time, participants infer less SIs (see e.g., Bott and Noveck, 2004; van Tiel and Schaeken, 2017); similarly, when working memory was burdened, pragmatic inferences dropped (see e.g., De Neys and Schaeken, 2007; Huang and Snedeker, 2009; Dieussaert et al., 2011; Marty et al., 2013). For this reason, we decided to focus solely on working memory in this study. However, to be fair, not all evidence points in the same direction. Grodner et al. (2010) observed no delay for pragmatic inferences from *some* compared to other, non-scalar expressions and in the latent class analysis of Heyman and Schaeken (2015), working memory capacity did not explain the interindividual variability in the interpretation of infelicitous *some* statements (see e.g., also Feeney et al., 2004; Breheny et al., 2013; Janssens and Schaeken, 2016).

Past research almost uniquely focused on *or-and*, *might-must* and especially *some-all*, whereby it was basically assumed that other scales would behave similarly. However, recent research (see e.g., Doran et al., 2009; Geurts, 2010;

Simons and Warren, 2018) questioned this uniformity. In a series of experiments, Van Tiel et al. (2016) showed that different types of scales are not all the same and we cannot use one type as the prototypical type. They tested 43 types of scalar inferences by presenting participants a statement with the weaker scalar term (e.g., *or*), and asking them if they would infer that the corresponding sentence with the stronger scalar term (e.g., *and*) is false. The results showed large differences across different lexical scales. Almost none of the participants made this falsity-inference with pairs as *content-happy* or *tired-exhausted*, while almost all of them made it for pairs like *possible-certain*, and *some-all*. As potentially relevant factors for the scalar diversity closed versus open scales, minimal versus rich contexts, word class and semantic distance are mentioned (see also Gotzner et al., 2018).

EXPERIMENT

The current experiment aims first of all to replicate the observed difficulty of patients with schizophrenia spectrum and other psychotic disorders with SIs with quantifiers, since that was to the best of our knowledge the first observation of it (Wampers et al., 2018). We opted for the use of a more fine-grained scale with a middle option as in the second experiment of Wampers et al. (2018) (see also Katsos and Bishop, 2011; Schaeken et al., 2018) instead of a task with the classic binary answer options. We hypothesize to observe similar effects with respect to the quantifier items, in other words, we expect the patients with schizophrenia spectrum and other psychotic disorders to interpret the quantifiers less pragmatically than the control group. Moreover, we want to investigate if these patients demonstrate scalar diversity as well or whether their difficulty is more uniform. In previous studies (e.g., Van Tiel et al., 2016), typically developing adults produce especially for quantifiers, disjunctions and modals a higher number of pragmatic responses. In Wampers et al. (2018) the clinical group produced fewer pragmatic inferences on the quantifiers, although this significant decrease was not large. Combining these two evidences, we expect our clinical group to produce fewer pragmatic responses for the quantifiers, disjunctions and modals than the control group, but still to a higher degree than for the other items, for which typically developing adults predominantly produce logical responses. Finally, we want to investigate if working memory capacity is related to the number of pragmatic responses given. It is well-documented that the working memory capacity of patients with schizophrenia spectrum and other psychotic disorders is decreased (see e.g., Goldman-Rakic, 1994; Silver, 2003; Lee and Park, 2005; Forbes et al., 2009; Arnsten, 2013; Nielsen et al., 2017). Therefore, it makes it very interesting to investigate for the first time the working memory and scalar diversity observations with this group. We hypothesize that working memory capacity will definitely influence the items on which the control group produces a higher number of pragmatic responses (i.e., quantifiers, disjunctions and modals). In order to obtain these aims, we presented seven different scales to our participants, which consisted of a group of patients with schizophrenia spectrum and other psychotic disorders and a control group, while we also tested their working memory capacity.

Method

Participants

In total, 42 persons participated in the experiment (22 men and 20 women). Half of these participants (11 men and 10 women) with a mean age of 27.5 (SD = 4.99) were diagnosed with Schizophrenia Spectrum and Other Psychotic Disorders according to DSM-V criteria by an experienced psychiatrist. All patients were hospitalized at the moment of testing. The second half of the participants, the control group, was matched to the patient group with respect to age and educational level (see **Table 1**) and consisted of 21 adults (11 men, 10 women) with a mean age of 27.0 (SD = 5.42). All participants were of Dutch literacy and provided written informed consent. The study was granted full ethical approval by the Ethics Committee of the University Psychiatric Hospital KU Leuven.

Material

For the assessment of the sensitivity to SIs, we constructed a questionnaire taken from the Dutch items of Van Tiel et al. (2016) and Zevakhina (2012). The questionnaire contained 54 under-informative sentences subdivided into 35 critical items and 19 filler items. To exclude sequence effects, four randomized versions of the questionnaire were prepared. For each item, a fictional person named Vera made a statement that contained a scalar term and could give rise to a scalar implication. Next, the participants were asked whether it could be deduced that, according to Vera, the statement implied that a stronger scalar term was not involved. The assessment was made by means of a five-point Likert scale, ranging from 1 = totally disagree to 5 = completely agree. An example of a critical item is:

> *Vera says: "Some theater performances are interesting."*
> *Would you infer from this that, according to Vera, not all theater performances are interesting?*

The 35 critical items were subdivided into seven pairs of different scalars, whereby each pair had five critical items: existential quantifier items (*all-some, always-sometimes*), disjunctive items (*and-or*), modal items (*have to-may*), and four pairs of adjective items (*excellent-good*, *hot-warm*, *huge-big*, *terrible-bad*). The critical items can be found in Appendix 1. The questionnaire also contained 19 filler items, of which 13 were valid and six were invalid. The valid and invalid filler items are

TABLE 1 | Demographic variables of patient and control group.

	N	Age	Gender		Education			
			Men	Women	EE	SE	HE	UE
Patient group	21	27.5	11	10	3	12	2	4
Control group	21	27.0	11	10	0	14	3	4
Total	42	27.2	22	20	3	26	5	8

EE is elementary school education only, SE is secondary school education, HE is higher education, and UE is university education.

also listed in Appendix 1. These items are superficially similar to the critical items, but they are pragmatically or semantically clearly wrong. With these items, we could therefore also test whether or not our participants had sufficient language and reasoning capacities. An example of a valid control item is:

Vera says: "The garden is small."
Would you deduce from this that, according to Vera, the garden is not large?

An example of an invalid control item:

Vera says: "The sea is warm."
Would you deduce from this that, according to Vera, the sea is not clear blue?

The 52 statements of each stimulus set were bundled in random order in booklets that displayed one item per page to discourage participants to return to previous responses. The first page of each booklet contained the task instructions. On the last page participants filled in their age, gender and educational level.

Working memory was assessed by means of the Digit Span subtest (with three parts, that is listen to sequences of numbers orally and to repeat them (a) as heard, (b) in reverse order, and (c) in ascending order) of the Dutch version of the fourth edition of the Wechsler Adult Intelligence Scale (WAIS IV). The subtests' scores were converted on the basis of the test manual into a standardized working memory score.

Procedure

Each participant voluntarily participated in the study and signed the information and consent form. The participants were tested individually in a quiet room. The experiment took approximately 30 min per participant. The measurement of the working memory capacity was taken together with the researcher and lasted about 15 min. It took approximately 15 min to complete the implicature questionnaire.

Statistical Analysis

Overall performance on the filler items was good (86% for the control group, 82% for the clinical group). In line with Van Tiel et al. (2016); see also (Pipijn, 2014), participants who answered less than 14 out of 19 of the filler items correctly were excluded from the analyses. This implied that 4 of the control subjects and 5 of the participants with schizophrenia spectrum and other psychotic disorders were excluded from the analyses. Even after the exclusion of these participants, both groups did not differ significantly in terms of age and educational level.

The average performance of the included participants on the filler items was 92% for the control group and 89% for the clinical group.

The responses obtained on the five-point Likert scale were transposed into a tertiary score (1 and 2 were collapsed into "disagree," 3 was "neutral," and 4 and 5 were collapsed into "agree"). Given the ordinal character of tertiary scores, we performed a mixed effect ordinal regression analysis with the tertiary agreement score as the dependent variable. The independent variables were Group (with the levels schizophrenia spectrum and other psychotic disorders group and matched control group), Scalar-Type (with the seven different types of scalar terms) and Working Memory Capacity as measured by the standardized Digit Span score. The latter score was mean-centered. The model was fitted using the clmm() function from the ordinal package in R (Christensen, 2015). All models included random intercepts for participants and items and a random slope for scalar type to capture the extent to which the possible mean differences between scalar types may differ across participants.

We started with the most complex fixed effects structure including the three-way interaction between group, scalar type and working memory capacity besides all two-way interactions and main effects. Subsequently we used backward elimination which involved simplifying the model by removing interaction terms that did not contribute significantly as evaluated through a likelihood ratio test. We verified the final model fitting by evaluating whether Akaike's Information Criterion (AIC) would have led to the same conclusion. In all analyses we used an alpha level of 0.05.

Results

Table 2 presents the percentage of answers in each response category for each scalar type as a function of participant group and working memory capacity. Low and high working memory capacity are defined as a standardized DS score below or above the population mean, respectively.

The final model included one two-way interactions, that is, the interaction between Scalar-Type and Working Memory Capacity. For a complete description of the final model, see **Table 3**.

As can be seen in **Table 3** two scalar types, that is, quantifiers ($Z = -4.18$, $p < 0.000$) and disjunctions ($Z = -1.56$, $p = 0.0303$) differed significantly from the scalar type that acted as the

TABLE 2 | Percentages of each response type for each scalar type as a function of group and working memory capacity.

		Control group		Clinical group	
Scalar type	**Response options**	**Low WM**	**High WM**	**Low WM**	**High WM**
Disjunctions	Disagree	60.00	86.67	40.00	75.00
	Neutral	12.00	6.67	0.00	12.50
	Agree	28.00	6.67	60.00	12.50
Good–excellent	Disagree	48.00	48.33	42.50	30.00
	Neutral	16.00	21.67	30.00	40.00
	Agree	36.00	30.00	27.50	30.00
Big-huge	Disagree	44.00	50.00	57.50	35.00
	Neutral	16.00	11.67	20.00	32.50
	Agree	40.00	38.33	22.50	32.50
Modal	Disagree	60.00	65.00	45.00	47.50
	Neutral	12.00	10.00	12.50	20.00
	Agree	28.00	25.00	42.50	32.50
Quantifier	Disagree	96.00	90.00	87.50	70.00
	Neutral	0.00	10.00	2.50	15.00
	Agree	4.00	0.00	10.00	15.00
Bad-horrible	Disagree	44.00	48.33	57.500	37.50
	Neutral	16.00	13.33	10.00	22.50
	Agree	40.00	38.33	32.50	40.00
Warm-hot	Disagree	36.00	53.33	47.50	45.00
	Neutral	32.00	10.00	22.50	22.50
	Agree	32.00	36.67	30.00	32.50

TABLE 3 | Complete description of the final model*.

Random effects									
Groups	**Name**	**Variance**	**Std.Dev.**	**Correlations**					
Participant	(intercept)	3.62	1.91						
Item	(intercept)	0.37	0.61						
	Big-Huge	1.27	1.13	0.18					
	Modal	3.01	1.74	−0.75	0.36				
	Quantifier	8.42	2.90	0.30	0.54	0.34			
	Bad-horrible	1.31	1.14	−0.14	0.93	0.67	0.56		
	Warm-hot	1.37	1.17	−0.88	0.28	0.89	−0.09	0.57	
	Disjunction	7.33	2.71	−0.63	−0.06	0.57	−0.32	0.21	0.70
Number of groups	**Participant**	**Item**							
	33	35							
Fixed effects									
Coefficients	Estimate	Std.Error	*z*-value	Pr(>| z|)					
Psychosis	0.3217	0.4537	0.709	0.4783					
ST "Big-huge"	−0.1636	0.5060	−0.323	0.7465					
ST "Modal"	−0.4735	0.5560	−0.852	0.3944					
ST "Quantifier"	−4.9910	1.1917	−4.188	0.0000281					
ST "Bad-horrible"	0.0290	0.5043	0.058	0.9541					
ST "Warm-hot"	−0.0650	0.4976	−0.131	0.8961					
ST "Disjunction"	−1.5558	0.7184	−2.166	0.0303					
DS score	−0.0669	0.1278	−0.524	0.6005					
DS score × ST "Big-huge"	0.0371	0.1063	0.349	0.7271					
DS score × ST "Modal"	−0.0827	0.1318	−0.627	0.5304					
DS score × ST "Quantifier"	−0.0170	0.2498	−0.068	0.9457					
DS score × ST "Bad-horrible"	0.0541	0.1062	0.507	0.6118					
DS score × ST "Warm-hot"	0.0751	0.1045	0.718	0.4725					
DS score × ST "Disjunction"	−0.4597	0.1941	−2.369	0.0178					

**The fixed effects are: Group [with levels: Psychosis and age and education matched control subjects (=reference group)], ST = Scalar Type [with levels good-excellent (=reference group), big-huge, modals, quantifiers, bad-horrible, warm-hot, disjunctions], DS score = scaled Digit Span score from the WAIS, mean-centered so that the intercept reflects the results of a participant with the mean Digit Span score in the total study population.*

reference category i.e., "Good-Excellent." Additional pairwise comparisons using emmeans() showed that quantifiers are also interpreted more pragmatically than the adjective items "Big-Huge" ($Z = -4.143$, $p = 0.0007$), "Warm-Hot" ($Z = -4.048$, $p = 0.0010$), "Bad-Horrible" ($Z = -4.324$, $p = 0.0003$), and the modal items ($Z = -3.808$, $p = 0.0027$). No significant interpretative differences were observed between quantifier and disjunctive items ($Z = -2.513$, $p = 0.1546$).

Although the results in **Table 3** show that disjunctive items were interpreted significantly more pragmatically than items from the reference scalar type (Good- excellent), we observed no other significant pairwise differences were between disjunctions and other scalar types. This observation is probably due to the fact that the disjunctive scalar type is involved in a significant interaction with working memory capacity.

The significant interaction between the mean-centered measure of working memory capacity and the scalar type disjunctives ($\beta = -0.46$, $Z = -2.37$, $p = 0.018$) shows that the extent to which disjunctive items are interpreted pragmatically, depends on participants' working memory capacity. Participants with a lower working memory capacity will interpret disjunctive items more logically than participants with a higher working memory capacity. The higher someone's working memory capacity, the more he/she tends to interpret disjunctive items pragmatically. This can also be observed in **Table 2**.

There was no significant interaction between group and scalar type, so people diagnosed with Schizophrenia Spectrum and Other Psychotic Disorders and matched control subjects show the same response pattern when confronted with a diversity of scalar items. The similarity between both study populations is also illustrated in **Figure 1** which shows the boxplots of the fitted values of the final model for the different scalar types for control subjects and subjects diagnosed with Schizophrenia Spectrum and Other Psychotic Disorders.

GENERAL DISCUSSION

As mentioned in the literature review, despite the abundant evidence that people with schizophrenia spectrum and other psychotic disorders show pragmatic difficulties, not much is known about how they deal with SIs. The first aim of the current study was to fill this gap in an attempt to replicate the only study about this topic (Wampers et al., 2018) in which fewer pragmatic responses were given by people with psychosis on SIs with quantifiers when compared with controls. As a second aim, the study broadened the SIs used and investigated if we can

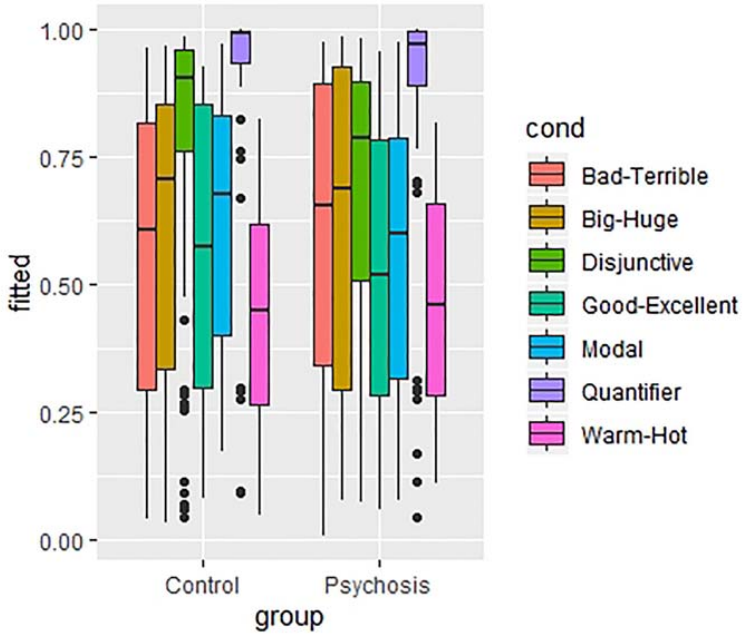

FIGURE 1 | The boxplots of the fitted values of the final model for the different scalar types for control participants and participants diagnosed with Schizophrenia Spectrum and Other Psychotic Disorders.

observe scalar diversity in people with schizophrenia spectrum and other psychotic disorders, as is shown in typically developed adults (see e.g., Van Tiel et al., 2016). As a third aim, the study investigated if working memory capacity is related to the number of pragmatic responses given. Our study found evidence for two of these three aims, but clearly in a nuanced way.

Starting with the results regarding the first aim, our study did not replicate the effect of Wampers et al. (2018): although our clinical group performed less pragmatically on the quantifier items than the control group, this effect was not significant. Hence, we could not confirm the hypothesis that the pragmatic difficulties of people with schizophrenia and other psychotic disorders can be extended to SIs. These results are therefore in contrast with Wampers et al. (2018), who did observe such a significant difference. Regarding the second aim, as hypothesized, we observed scalar diversity in our data: disjunctive and quantifier items were treated more pragmatically than the adjectives. These findings are in line with recent findings (see e.g., Van Tiel et al., 2016), which show that not all scalar items elicit a similar number of pragmatic responses. With respect to the third aim, an effect of working memory was observed, but again only in a subtle way: only the disjunctive items were solved less pragmatically by the people with lower working memory capacity.

How can we explain these results? We observed that especially disjunctive and quantifier items elicited more pragmatic responses and that the adjectives are answered more logically. This result ties well with previous studies, although we used a ternary scale instead of a binary scale. That a different procedure did lead to more or less the same results adds to the robustness of the scalar diversity effect. Interestingly, there was no significant effect of the clinical condition on scalar diversity. In other words, the clinical group showed more or less the same pattern of results as the control group. A potential explanation can be found in the study of Moro et al. (2015). They presented people with schizophrenia sentences in which they had to detect anomalies. Some of these sentences contained syntactic errors (violations of Universal Grammar principles) or semantic errors, resulting from a contradiction in the computation of the whole sentence meaning. The people with schizophrenia had only difficulties in identifying syntactic anomalies, suggesting an impairment of syntactic knowledge in schizophrenia. There were, however, no difficulties observed with semantic anomalies. The absence of difficulties with identifying semantic errors points to the absence of a semantic deficit. This could explain the lack of an effect of the clinical condition on scalar diversity, since the major hypotheses for scalar diversity mentioned in the literature are of a semantic nature (e.g., closed versus open scales, minimal versus rich contexts, word class, positively versus negatively oriented scalar words, and semantic distance) and hence will have a more or less equal influence on the control and the clinical group, as the absence of a main effect of group also shows. This observation is important with respect to the claim in the introduction of Bambini et al. (2016) that the high frequency of impairment suggests that the pragmatic deficit is a core feature of schizophrenia. The fact that there were no differences between our two groups with respect to scalar diversity runs against this idea, or at least, it nuances this thesis in showing that not all domains of pragmatic language are impaired. This important nuance is even more strengthened in the next point.

Contrary to our hypothesis and to Wampers et al. (2018), this study has been unable to demonstrate that for quantifiers there is a difference between the clinical and the control group. What might cause the differences between our results and those of Wampers et al. (2018) There are two important procedural differences between the current study and Wampers et al. (2018). The latter presented 10 quantifier items and only quantifier items, while in the current study only five quantifier items were presented, which were, moreover, mixed with many other scalar terms. Dieussaert et al. (2011) demonstrated that when participants have to change strategy often (in their case, by manipulating the number of filler items; in the current study by presenting seven different scalar items), the number of pragmatic responses decreases. It might therefore be that the performance of our control group was more logically compared to a study where only (and more) quantifiers were presented. Moreover, the tasks for the participants also clearly differed between the two studies. Wampers et al. (2018) used a statement-evaluation-task, where participants were presented with10 underinformative quantifier items like "*Some oaks are trees*" and they were asked to judge them as either true (logical) or false (pragmatic). This paradigm fits quantifiers very well, but it is incompatible with many other scalar expressions. Therefore, as in Pipijn (2014), Van Tiel et al. (2016), and Zevakhina (2012), the current study employs an inference paradigm, which, in general, leads to higher rates of SIs than the statement-evaluation paradigm (Geurts and Pouscoulous, 2009). In other words, it might be that these procedural differences caused the observed difference between the two studies on the quantifiers with respect to the effect of clinical group.

However, these conflicting data also accord with the mixed evidence with participants with ASD. On the one hand, some studies (e.g., Pijnacker et al., 2009; Chevallier et al., 2010; Su and Su, 2015) observed a similar amount of SIs with quantifiers for participants with ASD as for typically developing participants. Also, Hochstein et al. (2017) observed no difference in the amount of SIs between adolescents with ASD and controls. However, despite the ASD-group showing awareness of speakers' mental states, they were not always considering these spontaneously when deriving SIs. On the other hand, there are two recent studies in which children with ASD answered less pragmatically on SIs than typically developing children (Schaeken et al., 2018; Mazzaggio et al., 2021).

Therefore, it is still under debate whether participants from clinical populations, and in this specific study participants with schizophrenia spectrum and other psychotic disorders, are less pragmatic with quantifiers than typically developing control participants. It is clear that more research is needed, not only with the current population, but also with other clinical populations. Therefore, it looks to us that SIs with quantifiers developed from the poster child of pragmatic inference (Scontras et al., 2018) into a capricious teenager. We still see the possibilities, the intrinsic promises of SIs with quantifiers as a key element in experimental pragmatics, but getting them realized is definitely challenging.

Our data regarding the role of working memory are partly in line with the literature (see the introduction for more details on the controversial nature of the role of working memory), finding a small and specific working memory effect, that is, on disjunctive items. Most relevant for our study presumably is van Tiel et al. (2019), which combined investigating scalar diversity with manipulations of working memory load with typically developed adults. Their study revealed an interesting significant interaction between memory load and scalar type. Greater memory load led to fewer pragmatic responses for four scales (*or-and*, *might-must*, *some-all*, and *most-all*), but for three scales there was no working memory effect at all (*low-empty*, *scarce-absent*, and *try-succeed*). Hence, like our study, these findings not only demonstrate scalar diversity, but also a nuanced working memory effect. However, contrary to our results, they did also find a working memory effect on the quantifiers and modals. Three important differences between their and our study might cause this difference: a statement-evaluation-task versus an inference paradigm, the answer options offered to the participants (a binary option versus 5 options) and, definitely important, the fact that we only measured working memory and therefore treated it as a interindividual difference variable, while in van Tiel et al. (2019) working memory load was a manipulated factor (see also Dieussaert et al., 2011 for a discussion of measuring and manipulating working memory). What makes disjunctions special so that in both studies the inference from "*or*" to "*not and*" is cognitively costly? There are different potential accounts, but an intriguing explanation can be found in Singh et al. (2016). They argue that the retrieval of alternatives for disjunctions is peculiar, since there are two mechanisms for generating alternatives for adults (lexical replacement and the possibility of deleting material to generate an alternative), while there is only one for children (lexical replacement), giving rise to the different number of pragmatic responses by children and adults (see also Tieu et al., 2017; Verschueren et al., 2004). It might be that this developmental difference is also linked to working memory capacity, in which more working memory capacity is needed for the two roads to the alternatives. Future research should clarify this possible link.

Given the role of pragmatics for quality of life, intervention studies are critical tools in the rehabilitation process. Recently, some promising intervention or remediation studies have been developed for clinical and older populations (see e.g., Tompkins et al., 2011; Blake et al., 2013; Gabbatore et al., 2015, Gabbatore et al., 2017; Lundgren and Brownell, 2016; Bambini et al., 2020; Parola et al., 2020). For example, the PragmaCom (Bambini et al., 2020) focuses on the use of the Gricean maxims to strengthen the appreciation and knowledge of the pragmatic processes in communication, and uses for instance metaphors, proverbs, humor, and off-topic verbosity. The outcome of our study with respect to SIs suggests that adding them to such training programs, albeit interesting, is not essential. The outcome with respect to disjunctions, however, suggest that adding a working memory component in intervention studies could strengthen them (see Cortese et al., 2014; Danielsson et al., 2015; Spencer-Smith and Klingberg, 2015).

Before concluding, we have to mention some limitations of our study. First, a working memory manipulation would be a stronger indicator of a potential working memory effect than the measurement that we used in the current study. Second, the diagnosis of our rather young group of patients was general. It would be interesting if future research could investigate an older group of patients, and definitely with more specific information

about the diagnosis. This seems especially relevant given the cognitive heterogeneity of people with schizophrenia spectrum and other psychotic disorders (see e.g., Van Rheenen et al., 2017a,b; Buonocore et al., 2021). Related is the absence of direct IQ and language measurements. In our experiment, we matched our participant on educational level, since it is associated with many life outcomes and functions, such as income, occupation, intelligence, and language. There is indeed abundant evidence that education is a significant driver of language proficiency (see e.g., Massing and Schneider, 2017; Rudd and Honkiss, 2020). Therefore, we used educational level as a proxy for language proficiency. Moreover, the filler items used are also an implicit test of basic language and reasoning abilities. Average accuracy of the total group of participants on those items was good (86% for the control group, 82% for the clinical group). Moreover, to be sure of the basic language and reasoning abilities of our participants, those who scored less than 14 out of 19 were excluded, which lead to the exclusion of 9 participants (4 in the control group and 5 in the clinical group). The average accuracy on the filler items for the included participants was 92 and 89%, respectively, clearly indicating good and comparable language and reasoning skills of the participants in our sample. However, given the important role of language proficiency (see e.g., Parola et al., 2020) and verbal IQ (e.g., Chevallier et al., 2010) for pragmatic understanding, a more direct measurement would have been better and is definitely a recommendation for future studies. Third, it would be interesting to add additional measurements apart from working memory. There is not so much work on executive functions and implicatures with typically developing adults and the evidence is mixed. Antoniou et al. (2016) observed that working memory predicts the amount of pragmatic scalar responses, but inhibition did not. Fairchild (2018) reported significant correlations between executive functions and pragmatic, but when factoring ToM, these correlations disappeared. Husband (2014), however, did observe an effect of executive function. In other domains of pragmatics, and especially with clinical populations, executive functions played an important role. Bosia et al. (2016) for instance observed significant correlations between processing of figurative language and verbal memory, while humor was correlated with verbal memory, verbal fluency and processing speed in patients with schizophrenia. With respect to proverbs interpretations by patients with schizophrenia, the role of executive functions is clearly determined: set shifting and planning in Sponheim et al. (2003), divided attention, set-shifting and inhibitory control in Thoma et al. (2009) and cognitive flexibility in Mossaheb et al. (2014). Especially inhibitory control seems to be important in clinical studies (see e.g., Li et al., 2017; Parola et al., 2020). Bambini et al. (2021) found that in the elderly inhibition was not a significant predictor, but cognitive flexibility played a significant role in pragmatic comprehension in the elderly. Hence, future studies could fruitfully explore this issue further with SIs by including executive functions like inhibition, set shifting and cognitive flexibility. Furthermore, Brüne and Bodenstein (2005) and Champagne-Lavau and Stip (2010), investigating cognitive and executive functions and ToM together, both observed that ToM seems to be a better predictor than the cognitive and executive functions. Hence, future research should ideally not use only a measurement of different executive functions but also of ToM.

CONCLUSION

Overall, our study adds new knowledge, both theoretically and clinically, to the field of clinical and experimental pragmatics. From a theoretical point of view, the most obvious implication of the current study is the importance of taking into account scalar diversity, not only when working with typically developed adults, but also with clinical groups: one cannot generalize from *some* scalar expressions. Moreover, the role of working memory has been confirmed, but, importantly, only in a nuanced way, that is, we only observed a reliable effect for the disjunction items. From a clinical point of view, this study does not support the finding of Wampers et al. (2018) that people with schizophrenia spectrum and other psychotic disorders have difficulties with the pragmatic interpretation of quantifiers. However, our study accords with the hypothesis that there is no severe semantic deficit in our clinical group, given the observed scalar diversity effect. Finally, given the link between pragmatic functioning and quality of life, the current results might be used to feed intervention studies. Our study seems to underline the potential role of working memory training in intervention programs for people with schizophrenia spectrum and other psychotic disorders.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ethics Committee of the University Psychiatric Center KU Leuven. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

WS and MW did the data interpretation, wrote the general discussion and designed the study, with the help of MD. LV prepared the experiment, constructed the stimuli, and performed the experiment. MW did the statistical analysis. WS wrote the first draft of the introduction and method section. All authors approved the final version of the manuscript.

ACKNOWLEDGMENTS

We are extremely grateful to Reviewer 1 who pointed out that in an earlier version of the manuscript this random slope was wrongly not included (see Judd et al., 2017).

REFERENCES

Abu-Akel, A., and Bailey, A. L. (2000). Letter. *Psychol. Med.* 30, 735–738.

Adamczyk, P., Daren, A., Sułecka, A., Błądziński, P., Cichocki, Ł, Kalisz, A., et al. (2016). Do better communication skills promote sheltered employment in schizophrenia? *Schizophr. Res.* 176, 331–339. doi: 10.1016/j.schres.2016.08.015

Agostoni, G., Bambini, V., Bechi, M., Buonocore, M., Spangaro, M., Repaci, F., et al. (2021). Communicative-pragmatic abilities mediate the relationship between cognition and daily functioning in schizophrenia. *Neuropsychology* 35, 42–56. doi: 10.1037/neu0000664

Andrés-Roqueta, C., and Katsos, N. (2017). The contribution of grammar, vocabulary and theory of mind in pragmatic language competence in children with autistic spectrum disorders. *Front. Psychol.* 8:996. doi: 10.3389/fpsyg.2017.00996

Antoniou, K., Cummins, C., and Katsos, N. (2016). Why only some adults reject under-informative utterances. *J. Pragmat.* 99, 78–95. doi: 10.1016/j.pragma.2016.05.001

Arnsten, A. F. T. (2013). The neurobiology of thought: the ground-breaking discoveries of Patricia Goldman-Rakic 1937–03. *Cereb. Cortex* 23, 2269–2281. doi: 10.1093/cercor/bht195

Bambini, V., Arcara, G., Bechi, M., Buonocore, M., Cavallaro, R., and Bosia, M. (2016). The communicative impairment as a core feature of schizophrenia: frequency of pragmatic deficit, cognitive substrates, and relation with quality of life. *Compr. Psychiatry* 71, 106–120. doi: 10.1016/j.comppsych.2016.08.012

Bambini, V., Arcara, G., Bosinelli, F., Buonocore, M., Bechi, M., Cavallaro, R., et al. (2020). A leopard cannot change its spots: a novel pragmatic account of concretism in schizophrenia. *Neuropsychologia* 139:107332. doi: 10.1016/j.neuropsychologia.2020.107332

Bambini, V., Van Looy, L., Demiddele, K., and Schaeken, W. (2021). What is the contribution of executive functions to communicative-pragmatic skills? Insights from aging and different types of pragmatic inference. *Cogn. Process.* doi: 10.1007/s10339-021-01021-w [Epub Online ahead of print].

Barner, D., Brooks, N., and Bale, A. (2011). Accessing the unsaid: the role of scalar alternatives in children's pragmatic inference. *Cognition* 188, 87–96.

Baron-Cohen, S. (1988). Social and pragmatic deficits in autism: cognitive or affective? *J. Autism Dev. Disord.* 18, 379–402. doi: 10.1007/BF02212194

Blake, M. L., Frymark, T., and Venedictov, R. (2013). Based systematic review on communication treatments for individuals with right hemisphere brain damage. *Am. J. Speech Lang. Pathol.* 22, 146–160. doi: 10.1044/1058-0360(2012/12-0021)

Bosco, F. M., and Parola, A. (2017). "Schizophrenia," in *Research in Clinical Pragmatics*, ed. L. Cummings (London, UK: Springer).

Bosia, M., Arcara, G., Buonocore, M., Bechi, M., Moro, A., Cavallaro, R., et al. (2016). "Communication in schizophrenia, between pragmatics, cognition, and social cognition," in *Biolinguistic Investigations on the Language Faculty*. ed. A. M. Di Sciullo (Amsterdam: John Benjamins), 213–234 doi: 10.1075/la.235.10bos

Bott, L., and Noveck, I. A. (2004). Some utterances are underinformative: the onset and time course of scalar inferences. *J. Mem. Lang.* 51, 437–457. doi: 10.1016/j.jml.2004.05.006

Breheny, R., Ferguson, H. J., and Katsos, N. (2013). Taking the epistemic step: toward a model of on-line access to conversational implicatures. *Cognition* 126, 423–440. doi: 10.1016/j.cognition.2012.11.012

Brüne, M. (2003). "Social cognition and behaviour in schizophrenia," in *The Social Brain. Evolution and Pathology* eds M. Brüne, H. Ribbert, and W. Schiefenhövel (Chichester, UK: Wiley and Sons), 277–313. doi: 10.1002/0470867221.ch13

Brüne, M., and Bodenstein, L. (2005). Proverb understanding reconsidered: theory of mind and pragmatics in schizophrenia. *Schizophr. Res.* 75, 233–239. doi: 10.1016/j.schres.2004.11.006

Buonocore, M., Inguscioet, E., Spangaro, M., et al. (2021). *Disentangling Cognitive Heterogeneity in Psychotic Spectrum Disorders: The Distribution of Performance on BACS. Manuscript Submitted for Publication*. Amsterdam: Elsevier.

Champagne-Lavau, M., and Joanette, Y. (2009). Pragmatics, theory of mind and executive functions after a right-hemisphere lesion: different patterns of deficits. *J. Neurolinguistics* 22, 413–426. doi: 10.1016/j.jneuroling.2009.02.002

Champagne-Lavau, M., and Stip, E. (2010). Pragmatic and executive dysfunction in schizophrenia. *J. Neurolinguistics* 23, 285–296. doi: 10.1016/j.jneuroling.2009.08.009

Chevallier, C., Wilson, D., Happé, F., and Noveck, I. (2010). Scalar inferences in Autism Spectrum Disorders. *J. Autism Dev. Disord.* 40, 1104–1117. doi: 10.1007/s10803-010-0960-8

Chiappe, D. L., and Chiappe, P. (2007). The role of working memory in metaphor production and comprehension. *J. Mem. Lang.* 56, 172–188. doi: 10.1016/j.jml.2006.11.006

Chierchia, G., Crain, S., Guasti, M. T., Gualmini, A., and Meroni, L. (2001). "The acquisition of disjunction: evidence for a grammatical view of scalar implicatures," in *Proceedings of the 25th Boston University Conference on Language Development*. Somerville, MA: Cascadilla Press.

Christensen, R. H. B. (2015). *Ordinal - Regression Models for Ordinal Data. R package version 2015.1-21*.

Colle, L., Angeleri, R., Vallana, M., Sacco, K., Bara, B. G., and Bosco, F. M. (2013). Understanding the communicative impairments in schizophrenia: a preliminary study. *J. Commun. Disord.* 46, 294–308. doi: 10.1016/j.jcomdis.2013.01.003

Columbus, G., Sheikh, N. A., Côté-Lecaldare, M., Häuser, K., Baum, S. R., and Titone, D. (2015). Individual differences in executive control relate to metaphor processing: an eye movement study of sentence reading. *Front. Hum. Neurosci.* 8:1057. doi: 10.3389/fnhum.2014.01057

Corcoran, R., Cahill, C., and Frith, C. D. (1997). The appreciation of visual jokes in people with schizophrenia: A study of "mentalizing" ability. *Schizophr. Res.* 24, 319–327. doi: 10.1016/s0920-9964(96)00117-x

Corcoran, R., Mercer, G., and Frith, C. D. (1995). Schizophrenia, symptomatology and social inference: Investigating "theory of mind" in people with schizophrenia. *Schizophr. Res.* 17, 5–13. doi: 10.1016/0920-9964(95)00024-g

Cortese, S., Ferrin, M., Brandeis, D., Buitelaar, J., Daley, D., Dittmann, R. W., et al. (2014). Cognitive training for attention-deficit/hyperactivity disorder: meta-analysis of clinical and neuropsychological outcomes from randomized controlled trials. *J. Am. Acad. Child Adolesc. Psychiatry* 54, 164–174. doi: 10.1016/j.jaac.2014.12.010

Cummings, L. (2017). "'Cognitive aspects of pragmatic disorders," in *Research in Clinical Pragmatics*, ed. L. Cummings (Cham, Switzerland: Springer International Publishing AG), 587–616. doi: 10.1007/978-3-319-47489-2_22

Danielsson, H., Zottarel, V., Palmqvist, L., and Lanfranchi, S. (2015). The effectiveness of working memory training with individuals with intellectual disabilities - a meta-analytic review. *Front. psychol.* 6:1230. doi: 10.3389/fpsyg.2015.01230

De Neys, W., and Schaeken, W. (2007). When people are more logical under cognitive load. *Exp. Psychol.* 54, 128–133. doi: 10.1027/1618-3169.54.2.128

Dieussaert, K., Verkerk, S., Gillard, E., and Schaeken, W. (2011). Some effort for some: further evidence that scalar implicatures are effortful. *Q. J. Exp. Psychol.* 64, 2352–2367. doi: 10.1080/17470218.2011.588799

Doran, R., Baker, R. M., McNabb, Y., Larson, M., and Ward, G. (2009). On the non-unified nature of scalar implicature: an empirical investigation. *Int. Rev. Pragmat.* 1, 211–248. doi: 10.1163/187730909x12538045489854

Fairchild, S. (2018). *Speaker and Listener Effects on the Processing of Pragmatic Meaning*. Ph. D. thesis. USA: University of Delaware.

Feeney, A., Scrafton, S., Duckworth, A., and Handley, S. J. (2004). The story of some: everyday pragmatic inference by children and adults. *Can. J. Exp. Psychol.* 58, 121–132. doi: 10.1037/h0085792

Foppolo, F., Guasti, M. T., and Chierchia, G. (2012). Scalar implicatures in child language: give children a chance. *Lang. Learn. Dev.* 8, 365–394. doi: 10.1080/15475441.2011.626386

Forbes, N. F., Carrick, L. A., McIntosh, A. M., and Lawrie, S. M. (2009). Working memory in schizophrenia: a meta-analysis. *Psychol. Med.* 39, 889–905. doi: 10.1017/s0033291708004558

Frith, C. D. (2004). Schizophrenia and theory of mind. *Psychol. Med.* 34, 385–389.

Gabbatore, I., Bosco, F. M., Geda, E., Gastaldo, L., Duca, S., Costa, T., et al. (2017). Cognitive pragmatic rehabilitation program in schizophrenia: a single case fMRI study. *Neural Plast.* 2017:1612078 doi: 10.1155/2017/1612078

Gabbatore, I., Sacco, K., Angeleri, R., Zettin, M., Bara, B. G., and Bosco, F. M. (2015). Cognitive pragmatic treatment: a rehabilitative program for traumatic brain injury individuals. *J. Head Trauma Rehabil.* 30, E14–E28.

Geurts, B. (2010). *Quantity Implicatures*. Cambridge, UK: Cambridge University Press. doi: 10.1017/CBO9780511975158

Geurts, B., and Pouscoulous, N. (2009). Embedded implicatures?!? *Semant. Pragmat.* 2, 1–34.

Goldman-Rakic, P. S. (1994). Working memory dysfunction in schizophrenia. *J. Neuropsychiatry Clin. Neurosci.* 6, 348–357.

Gotzner, N., Solt, S., and Benz, A. (2018). Scalar Diversity, Negative Strengthening, and Adjectival Semantics. *Front. Psychol.* 9:1659. doi: 10.3389/fpsyg.2018.01659

Grice, H. P. (1975). "Logic and conversation," in *Syntax and Semantics, Vol. 3, Speech Acts*, eds P. Cole and J. L. Morgan (New York: Academic Press), 41–58. doi: 10.1163/9789004368811_003

Grodner, D. J., Klein, N. M., Carbary, K. M., and Tanenhaus, M. K. (2010). "Some," and possibly all, scalar inferences are not delayed: evidence for immediate pragmatic enrichment. *Cognition* 116, 42–55. doi: 10.1016/j.cognition.2010.03.014

Happé, F. G. E. (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

Heyman, T., and Schaeken, W. (2015). Some differences in some: examining variability in the interpretation of scalars using latent class analysis. *Psychol. Belg.* 55, 1–18. doi: 10.5334/pb.bc

Hochstein, L., Bale, A., and Barner, D. (2017).). Scalar implicature in absence of epistemic reasoning? The case of autism spectrum disorder. *Lang. Learn. Dev.* 14, 1–17. doi: 10.1016/j.rasd.2018.04.009

Horn, L. (1972). *On the Semantic Properties of Logical Operators in English*. Ph. D. thesis. Los Angeles, CA: University of California.

Huang, Y., and Snedeker, J. (2009). On-line interpretation of scalar quantifiers: insight into the semantics-pragmatics interface. *Cogn. Psychol.* 58, 376–415. doi: 10.1016/j.cogpsych.2008.09.001

Husband, E. M. (2014). A subclinical study of the cognitive resources underlying scalar implicature: a focus on scalar adjectives. *UCLA Work. Pap. Linguist.* 18, 189–211.

Janssen, J„ Krabbendam, L., Jolles, J., and van Os, J. (2003). Alteration in theory of mind in patients with schizophrenia and non-psychotic relatives. *Acta Psychiatr. Scand.* 108, 110–117. doi: 10.1034/j.1600-0447.2003.00092.x

Janssens, L., Drooghmans, S., and Schaeken, W. (2015). But: do age and working memory influence conventional implicature processing? *J. Child Lang.* 42, 695–708. doi: 10.1017/S0305000914000312

Janssens, L., Fabry, I., and Schaeken, W. (2014). 'Some' effects of age, task, task content and working memory on scalar implicature processing. *Psychol. Belg.* 54, 374–388. doi: 10.5334/pb.ax

Janssens, L., and Schaeken, W. (2016). 'But' Implicatures: a Study of the Effect of Working Memory and Argument Characteristics. *Front. Psychol.* 7:1520. doi: 10.3389/fpsyg.2016.01520

Judd, C. M., Westfall, J., and Kenny, D. A. (2017). Experiments with More Than One Random Factor: designs, Analytic Models, and Statistical Power. *Annu. Rev. Psychol.* 68, 601–625. doi: 10.1146/annurev-psych-122414-033702

Katsos, N., and Bishop, D. V. M. (2011). Pragmatic tolerance: implications for the acquisition of informativeness and implicature. *Cognition* 120, 67–81. doi: 10.1016/j.cognition.2011.02.015

Katsos, N., and Cummins, C. (2010). Pragmatics: from Theory to Experiment and Back Again. *Lang. Linguist. Compass* 4, 282–295. doi: 10.1111/j.1749$-$818x.2010.00203.x

Kiang, Michael, Light, Gregory, A., Prugh, Jocelyn, et al. (2007). Cognitive. Neurophysiological, and Functional Correlates of Proverb Interpretation Abnormalities in Schizophrenia. *J. Int. Neuropsychol. Soc.* 13, 653–663.

Langdon, R., Coltheart, M., Ward, P. B., and Catts, S. V. (2002a). Disturbed communication in schizophrenia: the role of poor pragmatics and poor mind-reading. *Psychol. Med.* 32, 1273–1284. doi: 10.1017/s0033291702006396

Langdon, R., Davies, M., and Coltheart, M. (2002b). Understanding minds and understanding communicated meanings in schizophrenia. *Mind. Lang.* 17, 68–104. doi: 10.1111/1468-0017.00189

Lee, J., and Park, S. (2005). Working memory impairments in schizophrenia: a meta-analysis. *J. Abnorm. Psychol.* 114, 599–611. doi: 10.1037/0021-843x.114.4.599

Li, X., Hu, D., Deng, W., Tao, Q., Hu, Y., and Yang, X. (2017). Pragmatic Ability Deficit in Schizophrenia and Associated Theory of Mind and Executive Function. *Front. Psychol.* 8:2164. doi: 10.3389/fpsyg.2017.02164

Linscott, R. J. (2005). Thought disorder, pragmatic language impairment, and generalized cognitive decline in schizophrenia. *Schizophr. Res.* 75, 225–232. doi: 10.1016/j.schres.2004.10.007

Lundgren, K., and Brownell, H. (2016). Figurative language deficits associated with right hemisphere disorder. *Perspect. ASHA Spec. Interest Groups* 1, 66–81. doi: 10.1044/persp1.SIG2.66

Marty, P., and Chemla, E. (2013). Scalar implicatures: working memory and a comparison with 'only'. *Front. Psychol.* 4:403. doi: 10.3389/fpsyg.2013.00403

Marty, P., Chemla, E., and Spector, B. (2013). Interpreting numerals and scalar items under memory load. *Lingua* 133, 152–163. doi: 10.1016/j.lingua.2013.03.006

Massing, N., and Schneider, S. L. (2017). Degrees of competency: the relationship between educational qualifications and adult skills across countries. *Large-scale Assess. Educ.* 5:6

Mazza, M., Di Michele, V., Pollice, R., Roncone, R., and Casacchia, M. (2008). Pragmatic language and theory of mind deficits in people with schizophrenia and their relatives. *Psychopathology* 41, 254–263. doi: 10.1159/000128324

Mazzaggio, G., Foppolo, F., Job, R., and Surian, L. (2021). Ad-hoc and scalar implicatures in children with autism spectrum disorder. *J. Commun. Disord.* 90:106089. doi: 10.1016/j.jcomdis.2021.106089

Mazzaggio, G., and Surion, L. (2018). A diminished propensity to compute scalar implicatures is linked to autistic traits. *Acta Linguist. Hung.* 65, 651–668. doi: 10.1556/2062.2018.65.4.4

Mo, S., Su, Y., Chan, R. C., and Liu, J. (2008). Comprehension of metaphor and irony in schizophrenia during remission: the role of theory of mind and IQ. *Psychiatry Res.* 157, 21–29. doi: 10.1016/j.psychres.2006.04.002

Moro, A., Bambini, V., Bosia, M., Anselmetti, S., Riccaboni, R., Cappa, S., et al. (2015). Detecting syntactic and semantic anomalies in schizophrenia. *Neuropsychologia* 79, 147–157. doi: 10.1016/j.neuropsychologia.2015.10.030

Mossaheb, N., Aschauer, H. N., Stoettner, S., Schmoeger, M., Pils, N., Raab, M., et al. (2014). Comprehension of metaphors in patients with schizophrenia-spectrum disorders. *Compr. Psychiatry* 55, 928–937. doi: 10.1016/j.comppsych.2013.12.021

Nielsen, J. D., Madsen, K. H., Wang, Z., Liu, Z., Friston, K. J., and Zhou, Y. (2017). Working Memory Modulation of Frontoparietal Network Connectivity in First-Episode Schizophrenia. *Cereb. Cortex* 27, 3832–3841. doi: 10.1093/cercor/bhx050

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

Noveck, I., and Sperber, D. (2007). "Why and how of experimental pragmatics: the case of 'scalar inferences'," in *Advances in Pragmatics*. ed. N. Burton-Roberts (Basingstoke: Palgrave), 307–330

Noveck, I. A. (2001). When children are more logical than adults: experimental investigations of scalar implicature. *Cognition* 78, 165–188. doi: 10.1016/S0010-0277(00)00114-1

Noveck, I. A. (2018). *Experimental Pragmatics: The Making of a Cognitive Science*. Cambridge: Cambridge University Press.

Panzeri, F., and Foppolo, F. (2016). "Deafness, Theory of Mind, and figurative language comprehension," in *BUCLD 40 Online Proceedings Supplements*, eds J. Scott and D. Waughtal. Italy: University of Milano-Bicocca.

Papafragou, A., and Musolino, J. (2003). Scalar implicatures: experiments at the semantics-pragmatics interface. *Cognition* 86, 253–282. doi: 10.1016/S0010-0277(02)00179-8

Parola, A., Salvini, R., Gabbatore, I., Colle, L., Berardinelli, L., and Bosco, F. M. (2020). Pragmatics, Theory of Mind and executive functions in schizophrenia: disentangling the puzzle using machine learning. *PLoS One* 15:e0229603. doi: 10.1371/journal.pone.0229603x

Peyroux, E., Prost, Z., Danset-Alexandre, C., Brenugat-Herne, L., Carteau-Martin, I., Gaudelus, B., et al. (2019). From "under" to "over" social cognition in schizophrenia: is there distinct profiles of impairments according to negative and positive symptoms? *Schizophr. Res. Cogn.* 15, 21–29. doi: 10.1016/j.scog.2018.10.001

Pijnacker, J., Hagoort, P., Buitelaar, J., Teunisse, J.-P., and Geurts, B. (2009). Pragmatic inferences in high-functioning adults with autism and Asperger syndrome. *J. Autism Dev. Disord.* 39, 607–618. doi: 10.1007/s10803-008-0661-8

Pipijn, K. (2014). *Adults, children and hummingbirds. An investigation into the development of scalar implicatures and the variables that influence them*. Ph.D. thesis Leuven: KU Leuven.

Pouscoulous, N., Noveck, I. A., Politzer, G., and Bastide, A. (2007). A developmental investigation of processing costs and implicature production. *Lang. Acquis.* 14, 347–375. doi: 10.1080/10489220701600457

Rudd, M., and Honkiss, L. (2020). Analysing the Correlation between English Proficiency and Academic Performance among Thai University Students. *Athens J. Educ.* 7, 122–138.

Schaeken, W., Schouten, B., and Dieussaert, K. (2019). Development of Quantitative and Temporal Scalar Implicatures in a Felicity Judgment Task. *Front. Psychol.* 9:2763. doi: 10.3389/fpsyg.2018.02763

Schaeken, W., Van Haeren, M., and Bambini, V. (2018). The understanding of scalar implicatures in Children with Autism Spectrum Disorder: dichotomized responses to violations of informativeness. *Front. Psychol.* 9:1266. doi: 10.3389/fpsyg.2018.01266

Schettino, A., Lauro, L. R., Crippa, F., Anselmetti, S., Cavallaro, R., and Papagno, C. (2010). The comprehension of idiomatic expressions in schizophrenic patients. *Neuropsychologia* 48, 1032–1040. doi: 10.1016/j.neuropsychologia.2009.11.030

Scontras, G., Tessler, M. H., and Franke, M. (2018). *Probabilistic Language Understanding: An Introduction to the Rational Speech Act Framework*. Available online at: https://www.problang.org (accessed November 29, 2020).

Silver, H. (2003). Working memory deficit as a core neuropsychological dysfunction in schizophrenia. *Am. J. Psychiatry* 160, 1809–1816. doi: 10.1176/appi.ajp.160.10.1809

Simons, M., and Warren, T. (2018). A closer look at strengthened readings of scalars. *Q. J. Exp. Psychol.* 71, 272–279. doi: 10.1080/17470218.2017.1314516

Singh, R., Wexler, K., Astle-Rahim, A., Kamawar, D., and Fox, D. (2016). Children interpret disjunction as conjunction: consequences for theories of implicature and child development. *Nat. Lang. Semant.* 24, 305–352. doi: 10.1007/s11050-016-9126-3

Spencer-Smith, M., and Klingberg, T. (2015). Benefits of a Working Memory Training Program for Inattention in Daily Life: a Systematic Review and Meta-Analysis. *PLoS One* 10:e0119522. doi: 10.1371/journal.pone.0119522

Sponheim, S. R., Surerus-Johnson, C., Leskela, J., and Dieperink, M. E. (2003). Proverb interpretation in schizophrenia: the significance of symptomatology and cognitive processes. *Schizophr. Res.* 65, 117–123. doi: 10.1016/s0920-9964(02)00525-x

Spotorno, N., McMillan, C. T., Rascovsky, K., Irwin, D. J., Clark, R., and Grossman, M. (2015). Beyond words: pragmatic inference in behavioral variant of frontotemporal degeneration. *Neuropsychologia* 75, 556–564. doi: 10.1016/j.neuropsychologia.2015.07.002

Su, Y. E., and Su, L. Y. (2015). Interpretation of logical words in Mandarin-speaking children with autism spectrum disorders: uncovering knowledge of semantics and pragmatics. *J. Autism Dev. Disord.* 45, 1938–1950. doi: 10.1007/s10803-014-2350-2350

Thoma, P., Hennecke, M., Mandok, T., Wähner, A., Brüne, M., Juckel, G., et al. (2009). Proverb comprehension impairments in schizophrenia are related to executive dysfunction. *Psychiatry Res.* 170, 132–139. doi: 10.1016/j.psychres.2009.01.026

Tieu, L., Yatsushiro, K., Cremers, A., Romoli, J., Sauerland, U., and Chemla, E. (2017). On the role of alternatives in the acquisition of simple and complex disjunctions in French and Japanese. *J. Semant.* 34, 127–152.

Tompkins, C. A., Blake, M. T., Wambaugh, J., and Meigh, K. (2011). A novel, implicit treatment for language comprehension processes in right hemisphere brain damage: phase I data. *Aphasiology* 25, 789–799. doi: 10.1080/02687038.2010.539784

Van Rheenen, T. E., Cropley, V., Zalesky, A., Bousman, C., Wells, R., Bruggemann, J., et al. (2017a). Widespread volumetric reductions in schizophrenia and schizoaffective patients displaying compromised cognitive abilities. *Schizophr. Bull.* 44, 560–574. doi: 10.1093/schbul/sbx109

Van Rheenen, T. E., Lewandowski, K., Tan, E., Ospina, L., Ongur, D., Neill, E., et al. (2017b). Characterizing cognitive heterogeneity on the schizophrenia–bipolar disorder spectrum. *Psychol. Med.* 47, 1848–1864. doi: 10.1017/S0033291717000307

van Tiel, B., Pankratz, E., and Sun, C. (2019). Scales and scalarity: processing scalar inferences. *J. Mem. Lang.* 105, 93–107. doi: 10.1016/j.jml.2018.12.002

van Tiel, B., and Schaeken, W. (2017). Processing conversational implicatures: alternatives and counterfactual reasoning. *Cog. Sci.* 41, 1119–1154. doi: 10.1111/cogs.12362

Van Tiel, B., van Miltenburg, E., Zevakhina, N., and Geurts, B. (2016). Scalar diversity. *J. Semant.* 33, 137–175.

Verschueren, N., Schaeken, W., De Neys, W., and d'Ydewalle, G. (2004). The difference between generating counterexamples and using them during reasoning. *Q. J. Exp. Psychol. A* 57, 1285–1308. doi: 10.1080/02724980343000774

Wampers, M., Schrauwen, S., De Hert, M., Gielen, L., and Schaeken, W. (2018). Patients with psychosis struggle with scalar implicatures. *Schizophr. Res.* 195, 97–102. doi: 10.1016/j.schres.2017.08.053

Zevakhina, N. (2012). "Strength and similarity of scalar alternatives," in *Proceedings of Sinn und Beedeutung (MIT Working Papers in Linguistics)*, Vol. 16 (Cambridge: MIT Press), 647–658.

Zhao, M., Liu, T., Chen, G., and Chen, F. (2015). Are scalar implicatures automatically processed and different for each individual? A mismatch negativity (MMN) study. *Brain Res.* 1599, 137–149. doi: 10.1016/j.brainres.2014.11.049

Conflict of Interest: 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 © 2021 Schaeken, Van de Weyer, De Hert and Wampers. 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) and the copyright owner(s) 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 June 2021
doi: 10.3389/fpsyg.2021.662940

Exposure to Literary Fiction Is Associated With Lower Psychological Essentialism

Emanuele Castano[1]*, *Maria Paola Paladino*[1], *Olivia G. Cadwell*[2], *Valentina Cuccio*[3] and *Pietro Perconti*[3]

[1] Department of Psychology and Cognitive Science, University of Trento, Trento, Italy, [2] Department of Psychology, The New School, New York City, NY, United States, [3] COSPECS, University of Messina, Messina, Italy

We investigated the impact of exposure to literary and popular fiction on psychological essentialism. Exposure to fiction was measured by using the Author Recognition Test, which allows us to separate exposure to authors of literary and popular fiction. Psychological essentialism was assessed by the discreteness subscale of the psychological essentialism scale in Study 1, and by the three subscales of the same scale (such as discreteness, informativeness, and biological basis) in Study 2 that was pre-registered. Results showed that exposure to literary fiction negatively predicts the three subscales. The results emerged controlling for political ideology, a variable that is commonly associated with psychological essentialism, and level of education.

Keywords: literary fiction, linguistic inferences, theory of mind, social cognition, psychological essentialism, fiction (narrative), cognitive literary theory

OPEN ACCESS

Edited by:
Francesco Ferretti,
Roma Tre University, Italy

Reviewed by:
Steven Brown,
McMaster University, Canada
Peter Verkoeijen,
Erasmus University
Rotterdam, Netherlands

***Correspondence:**
Emanuele Castano
emanuele.castano@unitn.it

Specialty section:
This article was submitted to Theoretical and Philosophical Psychology, a section of the journal Frontiers in Psychology

Received: *01 March 2021*
Accepted: *11 May 2021*
Published: *08 June 2021*

Citation:
Castano E, Paladino MP, Cadwell OG, Cuccio V and Perconti P (2021) Exposure to Literary Fiction Is Associated With Lower Psychological Essentialism. Front. Psychol. 12:662940. doi: 10.3389/fpsyg.2021.662940

INTRODUCTION

Oscar Wilde suggested, or more likely stated, that "You are what you read." This sweeping statement is hardly disprovable, but various aspects of it have been applied for the test through empirical studies in a variety of disciplines ranging from linguistics to developmental psychology, and from literary studies to cognitive science. Reading affects both *what* we learn and *how* we learn about the world (Heyes, 2012).

One of the most fascinating aspects of reading is that when reading, we go well-beyond what is literally said. Guided by pragmatic competence, we fill the gaps chiefly through the process of inference (Grice, 1957; Recanati, 2004). Inference-making occurs when we read all kinds of texts, but it is perhaps at its fullest when we engage with narrative texts (Bruner, 1986), the comprehension of which is aided by the creation and integration of mental representations of the characters, events, and context of the story (Graesser et al., 1991; Zwaan et al., 1995; Zwaan and Radvansky, 1998; Zwaan, 1999, 2004). In human development, the comprehension of narrative texts appears earlier than the comprehension of other expository text and other genres (Kaplan, 2013), which is consistent with the fact that stories played a significant role in the evolution of *Homo sapiens* well before the practice of writing and reading (e.g., Boyd, 2010). According to anthropologist Polly Wiessner, stories played such a role by "evoking higher orders of theory of mind *via* the imagination, conveying attributes of people in broad networks (virtual communities), and transmitting the 'big picture' of cultural institutions that generate regularity of behavior, cooperation, and trust at the regional level" (Wiessner, 2014, p. 1).

The theoretical and empirical findings that have emerged over the last two decades support the conclusion of the study by Wiessner, particularly with regard to the relationship between

engagement with narrative fiction and the development of Theory of Mind (ToM; Dodell-Feder and Tamir, 2018). This article is based on this line of inquiry. Building on the existent theory based on the difference between literary and popular fiction, and on research showing their differential impact on social cognition (Kidd and Castano, 2013), we investigated whether engagement with literary, but not popular fiction, reduces psychological essentialism.

NARRATIVE FICTION AND THEORY OF MIND

The capacity of narrative fiction to transport the reader into the mind of characters has long been noted, but only in the work of cognitive literary theorist Lisa Zunshine this phenomenon has been thoroughly described. The traditional literary critical analysis of the practice of reading and writing describes this ability in terms of *imagination* and *pretense*. Zunshine emphasized, and convincingly showed, that our ability to make sense of fiction relies heavily on cognitive processes such as mind-reading and meta-representationality (Zunshine, 2006). This work in the tradition of literary studies is similar to psychological research.

Building on early insight demonstrating that when we read narrative fiction, we experience thoughts and emotions congruent with that of the fictional characters (Gerrig, 1993; Oatley, 1999), and researchers began exploring the relationship between reading fiction and ToM (Premack and Woodruff, 1978). Also known as taking an intentional stance (Dennett, 1989), or mentalizing (Frith, 1989), ToM, in its advanced form, can be defined as the capacity to infer and represent the mental states of other people. Is ToM enhanced by fiction reading? Mar et al. measured exposure to fiction by using an adapted version of the Author Recognition Test (ART) suggested by Stanovich and West (1989), which consisted of names of authors of narrative fiction and authors of non-fiction, and asked participants to select the name of authors that they recognized. This allowed the computation of exposure to fiction and non-fiction. Subsequently, participants performed a series of tests, including the Reading the Mind in the Eyes Test (RMET; Baron-Cohen et al., 2001). The RMET, among the most widely used measures of ToM, is a performance measure, which assesses the accuracy in inferring the mental states from the photographs of the eye region. Results showed that exposure to narrative fiction, but not non-fiction, is positively related to ToM (Mar et al., 2006).

The results of these studies are interpreted as stemming from the fact that narrative fiction simulates social life: readers are transported into the fictional world, and they identify with characters, feel their emotions, and imagine their thoughts and desires. In doing so, they engage in mentalizing and thus train their ToM skills (Oatley, 2016). Recent research has qualified this conclusion and proposed a complementary account of the effect of reading fiction on mind-reading.

In a series of experiments, participants were first randomly assigned to read one among many excerpts of novels or short stories that were categorized either as literary or as popular fiction, and they were then asked to complete the RMET. Results showed that participants in the literary fiction condition scored higher on the test compared with those in the popular fiction condition. The literary fiction condition participants also scored higher than those who were either in a non-fiction reading condition or in a condition in which they did not read anything at all (Kidd and Castano, 2013; see also Black and Barnes, 2015; Kidd et al., 2016; Pino and Mazza, 2016; Kidd and Castano, 2018; van Kuijk et al., 2018; but see Panero et al., 2016; Kidd and Castano, 2017, 2018; for a meta-analysis, see Dodell-Feder and Tamir, 2018).

The differential impact of exposure to literary vs. popular fiction is further supported by the research using the ART, the same instrument originally used by Mar et al. (2006) when exploring the relation between fiction reading and ToM. When Kidd and Castano (2017) factor-analyzed the answers to the version of the ART suggested by Acheson et al. (2008), which are provided by a large number of participants, they identified two factors that correspond to the recognition of literary vs. popular genre authors (see also Moore and Gordon, 2015). The literary fiction factor comprises authors such as Michael Ondaatje, Thomas Pynchon, Margaret Atwood, and Alice Walker. The popular fiction factor comprises authors such as Tom Clancy, Nelson DeMille, Danielle Steel, and James Patterson (for a complete list of authors and their loadings on the two factors, see Kidd and Castano, 2017). After assigning a separate score representing familiarity with literary fiction and familiarity with popular fiction to each participant, they used both of these scores in a multiple regression analysis with performance on the RMET as a criterion. Resembling these experimental findings, only exposure to literary fiction emerged as a predictor and this relation was not accounted for by the differences in gender, age, undergraduate major, level of education, or self-reported trait empathy. Kidd and Castano (2017) also computed a variety of checks with regard to literary and popular factors of the ART, notably checking for the number of authors, the publication date of the work by the authors, specific genres, and the presence of classic authors. These checks did not alter the findings, and the two factors, namely, literary, and popular, were also used in the research that followed, in which they differentially predicted thinking styles and social cognition biases (Castano et al., 2020). Thus, it appears that, notwithstanding the imperfection of the ART as a measure of exposure to fiction, the instrument can be considered a valid proxy for exposure to fiction and computing the two separate scores is meaningful and statistically significant (Kidd and Castano, 2017). The ART is further discussed in the following section.

NARRATIVE FICTION AND SOCIAL CATEGORIES

The rationale behind the studies reviewed earlier is that literary fiction revolves around complex, *round* (Forster, 2002) characters, whose mental life is not explicitly revealed. This requires a mentalizing effort, which results in priming and training of ToM processes. The simpler, *flat* (Forster, 2002)

characters of popular fiction do not require the same mentalizing effort. Recent experimental research confirms that characters of literary fiction are perceived as more complex than those of popular fiction (Kidd and Castano, 2019).

Another way to consider the differences between popular and literary fiction is how these differences map onto the distinction made in social cognition research between the category-based and individuated perceptions (Brewer, 1988; Fiske and Neuberg, 1990). In his analysis of characterization in fiction, Culpeper (2001) proposed that the flat characters of popular fiction are easily recognized as fitting certain social categories (adolescent or elderly, Latino or WASP; Tajfel, 1981) a social role (Goffman, 1956), or, we suggested, as exemplars of a certain personality type (such as the extroverted type and the neurotic type) (McCrae and Costa, 1999). The characters of literary fiction, on the other hand, are category-resistant; these characters do not easily fit the social categories that we routinely use to make sense of the social world (Eder et al., 2010; Keen, 2011; Kidd and Castano, 2017), and those characters are thus more likely to be appraised through individuated perception. From the social cognition research, we also learn that we appraise others in a person-based, individuated manner only when we are motivated to do so (e.g., because we are in a competitive interpersonal relation; Ruscher and Fiske, 1990) or when the information about the person challenges the group stereotype (Rubinstein et al., 2018).

In a novel by, say, Tony Hillerman or James Patterson, the category-based appraisal may work, and the schema for the appropriate social categories might even be reinforced. In contrast, if we are engaging with the characters in a novel by Thomas Pynchon or Virginia Woolf, because characters are more complex and resist social categorization, individuated appraisal is more likely. Being confronted with these literary fiction characters may, in the long run, diminish our propensity to appraise the world in terms of social categories. If people keep encountering, either in real life or in the fictional world of novels, who are overly emotional British (or unemotional Italians), emotionally attuned engineers, and empathic bankers, then they may come to perceive British (and Italians), engineers, and bankers categories as loose associations, rather than discrete, entitative categories (Castano et al., 2002; Yzerbyt et al., 2004). Similarly, if the behavior of a fictional character suggests extraversion in one context but introversion in another, the belief that people are either extroverted or introverted may be undermined.

Since characters of literary fiction undermine the reification of social categories and of social types, such as personality types, exposure to literary fiction should reduce psychological essentialism.

PSYCHOLOGICAL ESSENTIALISM

Being an essentialist means believing that behind the variety of observable phenomena, there are essences that explain their systematic nature; essentialism means belief in the metaphysical claim that there are natural essences that explain the surface similarity. This stance is rooted in the history of thought, at least since the time of Plato and Aristotle, and is still very much alive in contemporary thought (e.g., Putnam, 1975). How do we speak of "cats" given the manifold manifestations of this phenomenon in our ordinary experience? The answer is that among all the manifestations of "catness" in the world, there is the essence of "cat," a set of unavoidable properties that something must possess in order to be a cat, and that in fact instantiate all cats.

Unlike its philosophical variant, which consists of a metaphysical thesis, psychological essentialism emphasizes representations of the world, rather than claims about the foundations on which the world is made. It is descriptive rather than normative. Psychological essentialism is less about the claim that things have essences and more about how representations of things of people can reflect this tendency, even if it is wrong (Medin and Ortony, 1989). The main advantage of psychological essentialism is its cognitive frugality. It is a quick and inexpensive heuristic that provides a basis (and a justification) for social categorization (Yzerbyt et al., 1997; Newman and Knobe, 2019). It does not matter whether it is true, as long as it is successful/adaptive under most social circumstances.

Psychological essentialism has been studied from two main perspectives. In the field of education, particularly with regard to the notion of personhood, the study by Carol Dweck on entity vs. incremental mindset has shown the negative consequences which the belief in the non-malleability of self-attributes has on motivation in terms of lack of persistence, low importance attributed to efforts and to learning goals (Dweck, 1999, 2008; Yeager and Dweck, 2012). Social psychologists have focused on social categories, showing that psychological essentialism enhances stereotyping (Bastian and Haslam, 2006), prejudice (e.g., Haslam et al., 2002; Chen and Ratliff, 2018), the acceptance, and the justification of social inequalities (Mandalaywala et al., 2018), and it lessens the desire to enter in contact with racial out-group members (Williams and Eberhardt, 2008). In spite of significant bad press, as noted earlier, psychological essentialism has strategic value for both the individual and the group (Ryazanov and Christenfeld, 2018). Such value has also been discussed in the specific context of fiction by Zunshine (2008) who observed that essentialist thinking not only allows us to place entities into categories but also, and above all, allows us to make inferences (Zunshine, 2008). Without denying their many potential negative consequences, Zunshine (2008) pointed out that such inferences greatly contribute to our ability to understand and deal with the intentions of others. This holds true in both circumstances when we make inferences about natural kinds, for example, other people or a tiger (i.e., when we make inferences about the behaviors of others and behave accordingly) and when we make inferences about artifacts.

Irrespective of its relative value, what is of interest for our purposes is that psychological essentialism is culturally transmitted (Rhodes et al., 2012; Rhodes and Moty, 2020) and remains malleable throughout the lifetime both in terms of mindset (e.g., Levy and Dweck, 1999; Blackwell et al., 2007) and with regard to social categories. Of particular relevance for our hypothesis on the impact of literary fiction on psychological essentialism is a recent study by Pauker et al. (2018), in which white American first-year college students who moved to Hawaii,

a racially diverse environment with a high multiracial population, were followed over a 9-month period. They found that the essentialist beliefs of these students about race decreased over time and that this was due to an increase in the ethnic diversity of the acquaintances of students (see also, Young et al., 2013). According to the authors of these studies, exposure to diversity creates uncertainty and challenges category boundaries, hence reducing psychological essentialism. As noted earlier, when the social world is complex and not easy to categorize, essentialism ceases to be a valid heuristic to construe and make sense of it.

Our hypothesis concerning the impact of literary fiction is based on the same premises. Literary fiction exposes readers to the diversity and complexity of humanity, and as such it reduces the extent to which readers engage in psychological essentialism. To test this hypothesis, we relied on the ART as a measure of exposure to literary vs. popular fiction (Kidd and Castano, 2017; Castano et al., 2020) and on the psychological essentialism scale developed by Bastian and Haslam (2006). A two-step procedure was followed. First, in Study 1, we analyzed the data from three different samples in which the ART, one of the psychological essentialism subscales, a measure of political ideology, and the level of education were collected. The latter two variables are important to exclude possible confounds of the ART scores, as detailed in the following section. These samples were collected for different purposes and at different times. The heterogeneity of the samples means that there is considerable noise that might interfere with the relationship under investigation in this study, resulting in a rather conservative test of the same. Second, we conducted a pre-registered study (Study 2), in which the ART, the three subscales of the essentialism scale, a measure of political ideology, and level of education were measured. Other variables were also included in this data collection, but as per the pre-registration, this was done for exploratory purposes and the data are not presented in this study.

STUDY 1

Participants (residents of the United States) were recruited *via* Amazon Mechanical Turk (MTurk) and paid from US$2 to US$5, depending on the length of the specific study they took part in. MTurk is a crowdsourcing marketplace for work which is used extensively in behavioral research and has been proven to be a reliable source of good quality data (Crump et al., 2013). Out of a total sample of 795 participants, 37 participants were excluded for failing to select at least one author or selecting more foils than authors on the ART, leaving a sample of 758 (415 females, 1 undeclared; age: $M = 34$, range: 18–82; education: some high school [3], high school [94], some college [271], college graduate [325], and graduate degree [65]).

Measures

The ART suggested by Acheson et al. (2008) was used. While self-reports of reading habits may be unduly influenced by the desire to appear well-read/knowledgeable, the ART is less likely to suffer from this shortcoming because it includes names of non-authors, which can then be used to assess (and statistically correct for) the tendency of the participants to inflate their knowledge of fiction authors. Also, the ART predicts the actual engagement of the participants with fiction (Stanovich and Cunningham, 1992; Stanovich et al., 1995; Rain and Mar, 2014), and the modified versions have been used in the previous work to compare exposure to non-fiction and fiction (e.g., Mar et al., 2006), and different genres of fiction (e.g., Fong et al., 2013, 2015). Most importantly for present purposes, and as discussed earlier, the ART of Acheson et al. (2008) has been shown to have a two-factor structure that corresponds to exposure to literary and popular fiction authors (Moore and Gordon, 2015; Kidd and Castano, 2017) and that separate scores computed for literary and popular fiction differentially predict social cognition processes and thinking styles (Kidd and Castano, 2017; Castano et al., 2020).

We thus followed exactly the same procedure used in the earlier research (Kidd and Castano, 2017; Castano et al., 2020) to compute ART Literary, ART Popular, and ART Foil scores as proportion of the selection of each category: ART Literary $M = 0.29$, $SD = 0.23$; ART Popular $M = 0.28$, $SD = 0.23$; ART foils $M = 0.01$; $SD = 0.04$. The discreteness subscale of the essentialism scale developed by Bastian and Haslam (2006) was used, which includes eight items (e.g., "People can behave in ways that seem ambiguous, but the central aspects of their character are clear-cut," "People can have many attributes and are never completely defined by any particular one" (reversed), "Everyone is either a certain type of person or they are not"). An average score was computed ($\alpha = 0.85$), so that high scores mean a strong tendency to see individuals as falling into discrete categories ($M = 3.54$; $SD = 0.86$). Participants further answered three items asking participants to what extent they self-identified as liberal (1) or conservative (7), in general, from a fiscal and social point of view. A composite score was created by averaging responses to the three items ($\alpha = 0.93$), with high values indicating more conservative ($M = 3.33$; $SD = 1.68$).

Results

We used General Linear Model (GLM) (SAS Institute Inc., 2020) with discreteness as the criterion and the ART Literary and ART Popular scores as predictors, and ART Foil, political ideology, and education as covariates. As noted earlier, the data were collected from three different samples. To account for this source of variance, we included Sample as a class covariate. Due to their significant skewness, ART Literary and ART Popular were square-root transformed before being entered in the analysis. The results are shown in **Table 1**. ART Literary negatively predicted discreteness, while ART Popular did not. Among the covariates, ART Foil was not significant, but Sample was, indicating that discreteness levels varied from one Sample to another. While it is expected, and it does not *per se* impact on the interpretation of the main effects of the ART variables, we conducted further analyses in which the interaction effect between Sample and both ART Literary and ART Popular was included. Neither was significant, and the main effect of ART Literary was unchanged.

TABLE 1 | Multiple regression results with discreteness as criterion (Study 1).

Predictors		β		95% CI		SE	*F*	*p*	sr
ART-Lit		−0.18		−0.30	−0.08	0.05	11.17	<0.001	−0.11
ART-Pop		0.09		−0.01	0.18	0.05	2.78	0.10	0.06
ART-Foil		0.03		−0.04	0.10	0.03	0.90	0.34	0.03
Political ideology		0.16		0.10	0.23	0.04	21.12	<0.001	0.16
Education		−0.02		−0.09	0.05	0.04	0.30	0.58	−0.02
		Means			**SDs**				
Sample	3.12	3.57	3.73	0.76	0.88	0.81	23.88	<0.001	0.24

Model R-square = 0.12.

STUDY 2

Study 2 was pre-registered. Participants (residents of the United States) were recruited *via* MTurk and paid US$1.5 for their participation. Expecting the data loss, we over-recruited ($N = 300$) participants compared with the pre-registered ($N = 218$) participants. After deletion of participants following the pre-registered criteria ($N = 57$), the sample included 243 participants (105 females, 2 undeclared; age: $M = 40.93$; range 18–75). Education was also recorded: elementary school [0]; middle school [0]; some high school [3]; high school [24]; some college [47]; college [121]; and postgraduate [48].

Measures

The ART was included and scored as for Study 1, and resulted in similar scores: ART Literary $M = 0.23$, $SD = 0.21$; ART Popular $M = 0.22$, $SD = 0.21$; and ART foils $M = 0.03$; $SD = 0.05$. To assess psychological essentialism, in addition to the discreteness subscale ($M = 3.17$; $SD = 0.70$), the two other subscales, namely, informativeness and biological basis, developed by Bastian and Haslam (2006) were used. Informativeness (seven items) assesses beliefs that differences among people allow many inferences to be drawn about them (e.g., "Generally speaking, once you know someone in one or two contexts, it is possible to predict how they will behave in most other contexts"). Biological bases (including eight items) assess beliefs that human attributes are biologically grounded (e.g., "With enough scientific knowledge, the basic qualities that a person has could be traced back to, and explained by, their biological make-up"). A composite score was computed for both constructs, by averaging the items in each scale (informativeness: $M = 3.45$; $SD = 0.67$; biological basis: $M = 3.31$; $SD = 0.85$). Participants also answered one item indicating their self-identification as liberal (1) or conservative (7) with high values indicating more conservative ($M = 4.16$; $SD = 2.19$). They also indicated their level of education on an improved scale comprising seven levels (i.e., elementary school; middle school; some high school; high school; some college; college; and postgraduate).

Results

We used the same GLM (SAS Institute Inc., 2020) as for Study 1, with ART Literary and ART Popular scores as predictors, and ART Foils, political ideology, and education as covariates, predicting the three subscales of the psychological essentialism scale suggested by Bastian and Haslam (2006), namely, discreteness, informativeness, and biology. As for Study 1, and as pre-registered, variables with considerable skewness were square-root transformed, before being entered in the model. Results (shown in **Table 2**) replicate the pattern observed for discreteness in Study 1 and further indicate the same effect for the subscale informativeness. Both these effects had been pre-registered. The biology subscale, pre-registered without a strong hypothesis, followed the same pattern.

In the analysis reported earlier, an unexpected effect of ART foil emerged. The more foils a participant selected as author names, the stronger his/her essentialism score. ART Foil was not transformed because square-root or log transformation did not improve its strongly and positively skewed distribution. Therefore, we conducted further analyses in which instead of using ART Foil as a covariate, we removed it from both ART Literary and ART Popular, to create two "foils-adjusted" versions of both of these variables, and we entered them as a predictor in the same model described earlier. We did this for both Study 1 and Study 2. Results are presented in **Tables 3**, **4**. The predicted effect of (foils-adjusted) ART Literary remains significant in both cases. Interestingly, the (foils-adjusted) ART Popular also emerges as a significant predictor for discreteness, but contrary to ART Literary it positively predicts it. **Supplementary Figure 1** shows the residual plots for the original model and the results are shown in **Tables 1**, **2**, and the residual plots for the model presented as supplementary analyses are reported in **Tables 3**, **4**. **Supplementary Table 1** reports the bivariate correlations for both Study 1 and Study 2.

DISCUSSION

In this article, we put forward the hypothesis that exposure to literary, but not popular fiction, is associated with lower psychological essentialism. We tested this hypothesis in two large studies, in one of which (Study 2) we pre-registered the details of data collection and management, as well as the deletion criteria and statistical analyses. The hypothesis found clear support from the results of the analyses of the two studies: exposure to literary, but not popular, fiction was associated with reduced essentialism in terms of perceived discreteness (Study 1) and both discreteness and informativeness (Study 2). The same result emerged on the third subscale suggested by Bastian and Haslam (2006), i.e.,

TABLE 2 | Multiple regression results (Study 2).

	Discreteness						Informativeness						Biological basis					
Predictors	**β**	**95% CI**		**SE**	***F***	***p***	**sr**	**β**	**95% CI**		**SE**	**F**	***p***	**sr**	**β**	**95% CI**		**SE**	**F**	***p***	**sr**
Lit	−0.24	−0.37	−0.10	0.07	12.01	0.001	−0.20	−0.15	−0.29	−0.01	0.07	4.77	0.030	−0.13	−0.23	−0.40	−0.05	0.09	6.41	0.010	−0.15
Pop	0.04	−0.09	0.16	0.06	0.34	0.560	0.03	0.01	−0.11	0.14	0.06	0.04	0.850	0.01	0.06	−0.10	0.22	0.08	0.51	0.470	0.04
Foil	0.13	0.04	0.22	0.04	8.54	0.004	0.17	0.11	0.02	0.20	0.05	5.77	0.020	0.14	0.11	0.00	0.23	0.06	3.92	0.050	0.12
Political Ideology	0.13	0.04	0.22	0.05	7.93	0.000	0.16	0.07	−0.02	0.17	0.04	2.30	0.130	0.09	0.07	−0.05	0.19	0.06	1.26	0.260	0.06
Education	−0.06	−0.14	0.02	0.04	1.94	0.160	−0.08	−0.04	−0.02	−0.01	0.04	1.07	0.300	−0.06	0.01	−0.09	0.12	0.05	0.06	0.800	−0.04

Model R-square: discreteness = 0.2102; informativeness = 0.10; biological bases = 0.08.

TABLE 3 | Multiple regression results using discreteness as criterion (Study 1).

Predictors	**β**		**95% CI**		**SE**	***F***	***p***	**sr**
Lit-adj	−0.20		−0.31	−0.09	0.06	13.10	<0.001	−0.12
Pop-adj	0.11		0.00	0.21	0.05	4.17	0.04	0.07
Political ideology	0.16		0.09	0.23	0.04	19.94	<0.001	0.15
Education	−0.02		−0.09	0.05	0.04	0.23	0.63	−0.02
	Means			**SDs**				
Sample	3.12	3.57	3.73	0.76	0.88	0.81	23.88	<0.001	0.24

Model R-square = 0.13.

TABLE 4 | Multiple regression results (Study 2).

	Discreteness							Informativeness							Biological basis						
Predictors	**β**	**95% CI**		**SE**	***F***	***p***	**sr**	**β**	**95% CI**		**SE**	***F***	***p***	**sr**	**β**	**95% CI**		**SE**	***F***	***p***	**sr**
Lit-adj	−0.34	−0.49	−0.18	0.08	17.63	< 0.001	−0.24	−0.21	−0.37	−0.05	0.08	6.34	0.01	−0.15	−0.23	−0.44	−0.02	0.11	4.83	0.03	−0.13
Pop-adj	0.14	−0.01	0.29	0.07	3.55	0.06	0.11	0.05	−0.10	0.20	0.08	0.48	0.49	0.04	0.07	−0.12	0.26	0.10	0.48	0.49	0.04
Political ideology	0.14	0.05	0.23	0.05	8.64	0.00	0.17	0.08	−0.02	0.02	0.05	2.67	0.10	0.10	0.09	−0.03	0.21	0.06	1.96	0.16	0.09
Education	−0.03	−0.11	0.05	0.04	0.59	0.44	−0.03	−0.02	−0.11	0.06	0.04	0.31	0.58	−0.04	0.03	−0.08	0.14	0.05	0.29	0.59	−0.03

Model R-square: discreteness = 0.21; informativeness = 0.10; biological bases = 0.074.

biology. The expected effect on this variable was pre-registered as exploratory, but the fact that it emerges and that it parallels that found on the other subscales provides further support for our rationale regarding the effect of exposure to literary fiction on psychological essentialism. To our knowledge, this is the first study of the relation between exposure to literary/popular fiction and psychological essentialism, and thus we cannot compare the strength of the effect we observed in the prior research. We noted, however, that the strength of the relationship between exposure to literary fiction and psychological essentialism is similar to the earlier research findings with regard to the relationship between exposure to literary fiction and ToM (Kidd and Castano, 2017; Castano et al., 2020).

We also conducted supplementary analyses that were not pre-registered, in which a different strategy was adopted for casual or self-aggrandizing responding to the ART—i.e., selection of foils as authors. These analyses revealed the same pattern, i.e., exposure to literary fiction negatively predicted essentialism. They also showed the opposite pattern for exposure to popular fiction, i.e., the more exposure to popular fiction, the greater psychological essentialism.

We first discussed the main hypothesis tested in this study and then discussed the results of the supplementary analyses.

Significance of the Finding

The main finding confirming our hypothesis adds to the growing literature on the effect of exposure to fiction on social cognition. This work has mostly focused on ToM, but recent findings indicate that exposure to literary fiction is also uniquely associated with attributional complexity for social events, i.e., increased social accuracy and, to a lesser extent, with reduced egocentric bias (Castano et al., 2020). Whether or not psychological essentialism can be considered as inaccurate or leading to biased perception, we perceived a conceptual similarity between these recent findings and those presented in this study: in both cases, literary fiction exposure is associated with a decreased use of reasoning heuristics.

Given the above-mentioned correlates of psychological essentialism, exposure to literary fiction seems to provide the same benefits as those gained through exposure to diversity in the real world (Pauker et al., 2018). Research on the contact hypothesis (Allport, 1954) has shown that entering into contact with actual members of an out-group results in less prejudicial attitudes and stereotyping (Pettigrew and Tropp, 2006, 2008), even when a social contact is mass-mediated (Schiappa et al., 2005). For instance, reduced prejudice and conflict may result from parasocial intergroup contact in virtual environments, i.e., the Internet, TV, movies, and radio (Schiappa et al., 2005; Ramiah and Hewstone, 2013). These studies, however, show context-specific effects, i.e., the stereotype is reduced with regard to the out-group whose members the person enters into contact with. In this study, we suggested that literary fiction undermines one of the very heuristics that supports stereotyping and prejudice, namely, psychological essentialism. The results presented in this study may also thus contribute to the understanding of the dynamics of stereotype change and prejudice reduction through parallel mechanisms to those already highlighted in the social psychological literature.

Does Exposure to Popular Fiction Matter?

We predicted and found that literary fiction has an inhibiting effect on psychological essentialism. We did not predict that popular fiction would have the opposite effect, and the results of the main analyses suggest that it does not. Why not? If, as it has been suggested (e.g., Kidd and Castano, 2013; Castano et al., 2020), popular fiction reifies social categories, one could expect that greater exposure to it would result in greater psychological essentialism. The reason why we did not predict such a pattern (see pre-registration) is that the kind of category-based perception that is associated with essentialism is likely to be engaged in the absence of a specific motivation to individuate (Ruscher and Fiske, 1990) or in the absence of information that challenges the group stereotype (Rubinstein et al., 2018). In fact, it has also been argued that the individuated appraisal-like process of ToM is not something we engage in with high frequency in our daily life, and that we rather rely on schematic information to navigate and make sense of our social world (see Theory of Society, Hirschfeld, 2006). Furthermore, we would conjecture that psychological essentialism, be it in the form of the essentialization of social categories such as nationality or gender, personality types, or the self, stems from social practices that start early in life, notably through the use of generic language. Gelman and Hirschfeld (1999) wrote that "It seems plausible [. . .] that children learn their essentialist beliefs from the messages directed toward them by mass media (including educational books and TV programs as well as *popular fiction*) and by parents" (p. 423–424; emphasis added). In other words, at least in the Western world, we teach our children to think in essentialistic terms, probably because of the strategic advantage that such thinking provides (Ryazanov and Christenfeld, 2018). Important cultural differences may exist between the Western world and other areas of the world, such as Southeast Asian, where the research has shown that through storybooks, different models of agency are conveyed (Goyal et al., 2019).

While Gelman and Hirschfeld (1999) also noted that powerful counter-essentialist imagery is provided in fiction for children such as *Horton Hears a Who* by Dr. Seuss, it is rather clear that essentialistic mental training is more pervasive. After all, popular fiction is *popular*. Be it with regard to children or an adult audience, narrative fiction that is considered popular is typically perceived as more enjoyable and easier to access, precisely because it can be read using heuristics—which, in turn, are reinforced by its reading. Literary fiction, on the other hand, primes and requires (and thus exercises) a different set of social cognition processes, in which the research shows to be more akin to individuated perception. Literary fiction, therefore, might be undermining, or providing a counterpoint to, the default mode of social perception. For this reason, we believed that while theorizing is consistent with the theoretical perspective proposed in this study, a positive effect of exposure to popular fiction on category-based perception and thus also on essentialism might be difficult to prove empirically. Further research, however, may well find conceptually similar findings, possibly using correlates of psychological essentialism.

Notwithstanding the above, the reanalysis of both studies presented in this study, in which a different strategy to control for the effect of careless or self-aggrandizing responses on the ART was used, revealed, for discreteness, an effect of exposure to popular fiction that is consistent with the rationale concerning the possible effects of popular fiction presented in this study and in other publications (e.g., Kidd and Castano, 2013, 2017; Castano et al., 2020). The alternative analytical strategy that results in this pattern is just as valid as the main one used in this study, and in fact, has been used in earlier work on the ART (e.g., Acheson et al., 2008). Our choice to use the covariate approach, and pre-register it for Study 2, was dictated by the fact that it is the approach used in earlier work which has distinguished, as we did in this study, ART scores for literary and popular fiction (e.g., Kidd and Castano, 2017). Further work on the psychometric properties of the ART will improve our understanding of the implications of using different strategies. In this study, we decided to present both.

Correlational vs. Experimental Research

Earlier research showing the differential impact of exposure to literary vs. popular fiction on ToM is complemented by experimental work (e.g., Kidd and Castano, 2013). Future research may also provide experimental, rather than correlational, evidence for the relationship between exposure to literary fiction and psychological essentialism reported in this study. Experimental research on the impact of fiction on ToM utilizes, however, performance measures, which we suspected are more sensitive to manipulation in the context of an experiment. The measure of psychological essentialism that we used, just as other measures, is the self-report indications about beliefs, rather than performance measures. The effect of an experimental manipulation on this type of measure might be more difficult to prove empirically, but our rationale, of course, predicts it. The advantage of experimental research, aside from providing a stronger basis for claims about causality, is that it allows for the investigation of mediating factors. An interesting

mediating hypothesis to investigate concerns the presence of generic language, which the research has shown to be directly linked to the development of essentialistic thinking in children (Rhodes et al., 2012; Segall et al., 2015; Gelman and Roberts, 2017). It could be that literary fiction makes lesser use of generic language than popular fiction. Other possible mediators are the evaluations provided by participants of the complexity of the characters and especially of their typicality with regard to their main social identities. Although it would not allow to test its mediating effect, analyzing the perception of fictional characters already available on platforms online might provide indirect evidence in support of the idea that category-resistant characters populate literary with greater frequency than popular fiction.

Limitations

We have already mentioned the limitation stemming from the correlational character of the results presented. In this study, we drew attention to other limitations. One limitation is that evidence from Study 1 has been obtained from a collection of three different samples, rather than an *ad hoc* collected sample of participants. The same sample-aggregation technique was used in the other two articles that factor-analyzed the ART (Moore and Gordon, 2015; Kidd and Castano, 2017). As noted earlier, however, this might also be considered a strength, rather than a weakness—especially in light of the fact that the expected effect emerges while controlling for sample effects. Furthermore, the fact that the pattern was replicated in the pre-registered study using Study 2 is reassuring in this regard. Nonetheless, more data, ideally the cross-cultural data, are needed. The second limitation is the number of covariates we included. Earlier research using the ART, also differentiating literary and popular fiction scores, has ruled out the confounding roles of variables such as personality traits, intelligence, empathic tendencies, or college major (e.g., Mar et al., 2009; Kidd and Castano, 2017; Castano et al., 2020). In this study, we further controlled for political ideology, which is known to be associated with psychological essentialism, and level of education, which may be loosely associated with reading habits. The impact of exposure to literary fiction proved robust to the influence of these variables, but future research may identify and test the impact of other correlates of either exposure to fiction or psychological essentialism, and refine or refute the pattern that we reported in this study.

CONCLUSION

In this study, we presented an evidence that exposure to literary fiction is associated with lower psychological essentialism. This finding is consistent with theorizing about the role of fiction in shaping not only *what* we think about the social world, but also *how* we think about it. It complements and extends the emerging body of empirical research on the impact of fiction, both in written form and in other formats, and it further shows the contributions to the debate around the cultural transmission of social cognition processes and thinking styles.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Comitato Etico—Dipartimento COSPECS. The patients/participants provided their informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

EC developed the hypothesis, designed study, and wrote first draft. MPP, OGC, VC, and PP contributed to preparation of material, data interpretation, and writing. All authors have approved the final version.

ACKNOWLEDGMENTS

We thank Wanda Orme for proofreading the article, and Rachel Smith and David Kidd for their help with the data collection.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.662940/full#supplementary-material

Image 1 | Figure 1A.

Image 2 | Figure 1B.

Image 3 | Figure 1C.

Image 4 | Figure 1D.

REFERENCES

Acheson, D. J., Wells, J. B., and MacDonald, M. C. (2008). New and updated tests of print exposure and reading abilities in college students. *Behav Res Methods* 40, 278–289. doi: 10.3758/BRM.40.1.278

Allport, G. W. (1954). *The Nature of Prejudice.* Cambridge, MA Perseus Books.

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., and Plumb, I. (2001). The "Reading the Mind in the Eyes" test revised version: a study with normal adults, and adults with Asperger syndrome or high-functioning autism. *J. Child Psychol. Psychiatry* 42, 241–251. doi: 10.1111/1469-7610.00715

Bastian, B., and Haslam, N. (2006). Psychological essentialism and stereotype endorsement. *J. Exp. Soc. Psychol.* 42, 228–235. doi: 10.1016/j.jesp.2005.03.003

Black, J., and Barnes, J. L. (2015). Fiction and social cognition: The effect of viewing award-winning television dramas on theory of mind. *Psychol. Aesthet. Creat. Arts* 9:423. doi: 10.1037/aca0000031

Blackwell, L. S., Trzesniewski, K. H., and Dweck, C. S. (2007). Implicit theories of intelligence predict achievement across an adolescent transition: a longitudinal study and an intervention. *Child Dev.* 78, 246–263. doi: 10.1111/j.1467-8624.2007.00995.x

Boyd, B. (2010). *On the Origin of Stories: Evolution, Cognition, and Fiction*. Cambridge: Harvard University Press. doi: 10.2307/j.ctvjf9xvk

Brewer, M. B. (1988). "A dual process model of impression formation," in *Advances in Social Cognition, Vol. 1. A Dual Process Model of Impression Formation*, eds T. K. Srull and R. S. Wyer, Jr. (Mahwah, NJ: Lawrence Erlbaum Associates, Inc.), 1–36.

Bruner, J. (1986). *Actual Minds, Possible Worlds*. Cambridge: Harvard University Press. doi: 10.4159/9780674029019

Castano, E., Martingano, A. J., and Perconti, P. (2020). The effect of exposure to fiction on attributional complexity, egocentric bias and accuracy in social perception. *PLoS ONE* 15:e0233378. doi: 10.1371/journal.pone.0233378

Castano, E., Yzerbyt, V., Paladino, M. P., and Sacchi, S. (2002). I belong, therefore, I exist: Ingroup identification, ingroup entitativity, and ingroup bias. *Pers. Soc. Psychol. Bull.* 28, 135–143. doi: 10.1177/0146167202282001

Chen, J. M., and Ratliff, K. A. (2018). Psychological essentialism predicts intergroup bias. *Soc. Cogn.* 36, 301–323. doi: 10.1521/soco.2018.36.3.301

Crump, M. J. C., McDonnell, J. V., and Gureckis, T. M. (2013). Evaluating Amazon's mechanical Turk as a tool for experimental behavioral research. *PLoS ONE* 8:e57410. doi: 10.1371/journal.pone.0057410

Culpeper, J. (2001). *Language and Characterisation: People in Plays and Other Texts*. London: Longman.

Dennett, D. C. (1989). *The Intentional Stance*. Cambridge: MIT Press.

Dodell-Feder, D., and Tamir, D. I. (2018). Fiction reading has a small positive impact on social cognition: a meta-analysis. *J. Exp. Psychol. Gen.* 147:1713. doi: 10.1037/xge0000395

Dweck, C. S. (1999). *Self-Theories: Their Role in Motivation, Personality and Development*. Philadelphia, PA: Taylor and Fran-cis/Psychology Press.

Dweck, C. S. (2008). Can personality be changed? The role of beliefs in personality and change. *Curr. Direct. Psychol. Sci.* 17, 391–394. doi: 10.1111/j.1467-8721.2008.00612.x

Eder, J., Jannidis, F., and Schneider, R. (Eds.). (2010). "Characters in fiction: An introduction," in *Revisionen: Characters in Fictional Worlds: Under- Standing Imaginary Beings in Literature, Film, and Other Media* (Berlin: Walter de Gruyter), 3–64. doi: 10.1515/9783110232424

Fiske, S. T., and Neuberg, S. L. (1990). A continuum of impression formation, from category-based to individuating processes: influences of information and motivation on attention and interpretation. *Adv. Exp. Soc. Psychol.* 23, 1–74. doi: 10.1016/S0065-2601(08)60317-2

Fong, K., Mullin, J. B., and Mar, R. A. (2013). What you read matters: the role of fiction genre in predicting interpersonal sensitivity. *Psychol. Aesthet. Creat. Arts* 7:370. doi: 10.1037/a0034084

Fong, K., Mullin, J. B., and Mar, R. A. (2015). How exposure to literary genres relates to attitudes toward gender roles and sexual behavior. *Psychol. Aesthet. Creat. Arts* 9, 274–285. doi: 10.1037/a0038864

Forster, E. M. (2002). *Aspects of the Novel*. New York, NY: RosettaBooks.

Frith, U. (1989). Autism and "theory of mind," in *Diagnosis and Treatment of Autism* (Boston, MA: Springer), 33–52. doi: 10.1007/978-1-4899-0882-7_4

Gelman, S. A., and Hirschfeld, L. A. (1999). How biological is essentialism. *Folkbiology* 9, 403–446.

Gelman, S. A., and Roberts, S. O. (2017). How language shapes the cultural inheritance of categories. *Proc. Natl. Acad. Sci. U.S.A.* 114, 7900–7907. doi: 10.1073/pnas.1621073114

Gerrig, R. J. (1993). *Experiencing Narrative Worlds: On the Psychological Activities of Reading*. New Haven, CT: Yale University Press.

Goffman, E. (1956). The nature of deference and demeanor. *Am. Anthropol.* 58, 473–502. doi: 10.1525/aa.1956.58.3.02a00070

Goyal, N., Wice, M., Aladro, A., Kallberg-Shroff, M., and Miller, J. G. (2019). Culture and the development of views of agency: perspectives from storybooks, parents, and children. *Dev. Psychol.* 55, 1096–1110. doi: 10.1037/dev0000671

Graesser, A., Golding, J. M., and Long, D. L. (1991). "Narrative representation and comprehension," in *Handbook of Reading Research, Vol. 2,* eds R. Barr, M. L. Kamil, P. Mosenthal, and P. D. Pearson (White Plains, NY: Longman), 171–204.

Grice, H. P. (1957). Meaning. *Philosoph Rev.* 66, 377–388. doi: 10.2307/2182440

Haslam, N., Rothschild, L., and Ernst, D. (2002). Are essentialist beliefs associated with prejudice? *Br. J. Soc. Psychol.* 41, 87–100. doi: 10.1348/014466602165072

Heyes, C. (2012). Grist and mills: on the cultural origins of cultural learning. *Philos. Trans. R. Soc. B Biol. Sci.* 367, 2181–2191. doi: 10.1098/rstb.2012.0120

Hirschfeld, L. (2006). "Who needs a theory of mind?" in *Biological and Cultural Bases of Human Inference,* eds R. Viale, D. Andler, and L. Hirschfeld (Mahwah, NJ: Erlbaum), 131–160.

Kaplan, D. (2013). Development of reading comprehension from middle childhood to adolescence: distributional and qualitative analyses of two genres. *Written Lang. Literacy* 16, 208–240. doi: 10.1075/wll.16.2.04kap

Keen, S. (2011). Readers' temperaments and fictional character. *N. Liter. History* 42, 295–314. doi: 10.1353/nlh.2011.0013

Kidd, D., and Castano, E. (2017). Different stories: How levels of familiarity with literary and genre fiction relate to mentalizing. *Psychol. Aesthet. Creat. Arts* 11, 474–486. doi: 10.1037/aca0000069

Kidd, D., and Castano, E. (2018). Reading literary fiction can improve theory of mind. *Nat. Hum. Behav.* 2, 604–604. doi: 10.1038/s41562-018-0408-2

Kidd, D., and Castano, E. (2019). Reading literary fiction and theory of mind: three preregistered replications and extensions of Kidd and Castano (2013). *Soc. Psychol. Pers. Sci.* 10, 522–531. doi: 10.1177/1948550618775410

Kidd, D., Ongis, M., and Castano, E. (2016). On literary fiction and its effects on theory of mind. *Sci. Study Liter.* 6, 42–58. doi: 10.1075/ssol.6.1.04kid

Kidd, D. C., and Castano, E. (2013). Reading literary fiction improves theory of mind. *Science* 342, 377–380. doi: 10.1126/science.1239918

Levy, S. R., and Dweck, C. S. (1999). The impact of children's static versus dynamic conceptions of people on stereotype formation. *Child Dev.* 70, 1163–1180. doi: 10.1111/1467-8624.00085

Mandalaywala, T. M., Amodio, D. M., and Rhodes, M. (2018). Essentialism promotes racial prejudice by increasing endorsement of social hierarchies. *Soc. Psychol. Pers. Sci.* 9, 461–469. doi: 10.1177/1948550617707020

Mar, R. A., Oatley, K., Hirsh, J., Dela Paz, J., and Peterson, J. B. (2006). Bookworms versus nerds: exposure to fiction versus non-fiction, divergent associations with social ability, and the simulation of fictional social worlds. *J. Res. Pers.* 40, 694–712. doi: 10.1016/j.jrp.2005.08.002

Mar, R. A., Oatley, K., and Peterson, J. B. (2009). Exploring the link between reading fiction and empathy: Ruling out individual differences and examining outcomes. *Communications* 34, 407–428. doi: 10.1515/COMM.2009.025

McCrae, R. R., and Costa, P. T. Jr. (1999). A five-factor theory of personality. *Handbook Pers. Theor. Res.* 2, 139–153.

Medin, D. L., and Ortony, A. (1989). "Psychological essentialism," in *Similarity and Analogical Reasoning,* eds S. Vosniadou and A. Ortony (Cambridge: Cambridge University Press), 179–195. doi: 10.1017/CBO9780511529863.009

Moore, M., and Gordon, P. C. (2015). Reading ability and print exposure: item response theory analysis of the author recognition test. *Behav. Res. Methods* 47, 1095–1109. doi: 10.3758/s13428-014-0534-3

Newman, G. E., and Knobe, J. (2019). The essence of essentialism. *Mind Lang.* 34, 585–605. doi: 10.1111/mila.12226

Oatley, K. (1999). Why fiction may be twice as true as fact: fiction as cognitive and emotional simulation. *Rev. Gen. Psychol.* 3, 101–117. doi: 10.1037/1089-2680.3.2.101

Oatley, K. (2016). Fiction: simulation of social worlds. *Trends Cogn. Sci.* 20, 618–628. doi: 10.1016/j.tics.2016.06.002

Panero, M. E., Weisberg, D. S., Black, J., Goldstein, T. R., Barnes, J. L., Brownell, H., et al. (2016). Does reading a single passage of literary fiction really improve theory of mind? An attempt at replication. *J. Pers. Soc. Psychol.* 111:e46. doi: 10.1037/pspa0000064

Pauker, K., Carpinella, C., Meyers, C., Young, D. M., and Sanchez, D. T. (2018). The role of diversity exposure in Whites' reduction in race essentialism over time. *Soc. Psychol. Pers. Sci.* 9, 944–952. doi: 10.1177/1948550617731496

Pettigrew, T. F., and Tropp, L. R. (2006). A meta-analytic test of intergroup contact theory. *J. Pers. Soc. Psychol.* 90, 751–783. doi: 10.1037/0022-3514.90.5.751

Pettigrew, T. F., and Tropp, L. R. (2008). How does intergroup contact reduce prejudice? Meta-analytic tests of three mediators. *Eur. J. Soc. Psychol.* 38, 922–934. doi: 10.1002/ejsp.504

Pino, M. C., and Mazza, M. (2016). The use of "literary fiction" to promote mentalizing ability. *PLoS ONE* 11:e0160254. doi: 10.1371/journal.pone.0160254

Premack, D., and Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behav. Brain Sci.* 1, 515–526. doi: 10.1017/S0140525X00076512

Putnam, H. (1975). *The Meaning of "Meaning"*. Minneapolis, MN: University of Minnesota Press. Retrieved from: https://hdl.handle.net/11299/185225

Rain, M., and Mar, R. A. (2014). Measuring reading behavior: examining the predictive validity of print-exposure checklists. *Empiric. Stud. Arts* 32, 93–108. doi: 10.2190/EM.32.1f

Ramiah, A., and Hewstone, M. (2013). Intergroup contact as a tool for reducing, resolving, and preventing intergroup conflict: evidence, limitations, and potential. *Am. Psychol.* 68, 527–542. doi: 10.1037/a0032603

Recanati, F. (2004). *Literal Meaning*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511615382

Rhodes, M., Leslie, S. J., and Tworek, C. M. (2012). Cultural transmission of social essentialism. *Proc. Natl. Acad. Sci. U.S.A.* 109, 13526–13531. doi: 10.1073/pnas.1208951109

Rhodes, M., and Moty, K. (2020). What is social essentialism and how does it develop? *Adv. Child Dev. Behav.* 59, 1–30. doi: 10.1016/bs.acdb.2020.05.001

Rubinstein, R. S., Jussim, L., and Stevens, S. T. (2018). Reliance on individuating information and stereotypes in implicit and explicit person perception. *J. Exp. Soc. Psychol.* 75, 54–70. doi: 10.1016/j.jesp.2017.11.009

Ruscher, J. B., and Fiske, S. T. (1990). Interpersonal competition can cause individuating processes. *J. Pers. Soc. Psychol.* 58, 832–843. doi: 10.1037/0022-3514.58.5.832

Ryazanov, A. A., and Christenfeld, N. J. (2018). The strategic value of essentialism. *Soc. Pers. Psychol. Compass* 12:e12370. doi: 10.1111/spc3.12370

SAS Institute Inc. (2020). *SAS Studio Help and Documentation*. Cary, NC: SAS Institute Inc.

Schiappa, E., Gregg, P. B., and Hewes, D. E. (2005). The parasocial contact hypothesis. *Commun. Monogr.* 72, 92–115. doi: 10.1080/03637750520003042544

Segall, G., Birnbaum, D., Deeb, I., and Diesendruck, G. (2015). The intergenerational transmission of ethnic essentialism: how parents talk counts the most. *Dev. Sci.* 18, 543–555. doi: 10.1111/desc.12235

Stanovich, K. E., and Cunningham, A. E. (1992). Studying the consequences of literacy within a literate society: the cognitive correlates of print exposure. *Memory Cogn.* 20, 51–68. doi: 10.3758/BF03208254

Stanovich, K. E., and West, R. F. (1989). Exposure to print and orthographic processing. *Read. Res. Q.* 402–433. doi: 10.2307/747605

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

Tajfel, H. (1981). *Human Groups and Social Categories: Studies in Social Psychology*. Cambridge: Cambridge University Press.

van Kuijk, I., Verkoeijen, P., Dijkstra, K., and Zwaan, R. A. (2018). The effect of reading a short passage of literary fiction on Theory of Mind: a replication of Kidd and Castano (2013). *Collabra Psychol.* 4:7. doi: 10.31234/osf.io/ht2ej

Wiessner, P. W. (2014). Embers of society: firelight talk among the Ju/'hoansi Bushmen. *Proc. Natl. Acad. Sci. U.S.A.* 111, 14027–14035. doi: 10.1073/pnas.1404212111

Williams, M. J., and Eberhardt, J. L. (2008). Biological conceptions of race and the motivation to cross racial boundaries. *J. Pers. Soc. Psychol.* 94:1033. doi: 10.1037/0022-3514.94.6.1033

Yeager, D. S., and Dweck, C. S. (2012). Mindsets that promote resilience: when students believe that personal characteristics can be developed. *Educ. Psychol.* 47, 302–314. doi: 10.1080/00461520.2012.722805

Young, D. M., Sanchez, D. T., and Wilton, L. S. (2013). At the crossroads of race: racial ambiguity and biracial identification influence psychological essentialist thinking. *Cult. Divers. Ethnic Minor. Psychol.* 19, 461–467. doi: 10.1037/a0032565

Yzerbyt, V., Rocher, S., and Schadron, G. (1997). "Stereotypes as explanations: a subjective essentialistic view of group perception," in *The Social Psychology of Stereotyping and Group Life,* eds R. Spears, P. J. Oakes, N. Ellemers, and S. A. Haslam (Oxford: Blackwell Publishing), 20–50.

Yzerbyt, V. E., Judd, C. M., and Corneille, O. E. (2004). *The Psychology of Group Perception: Perceived Variability, Entitativity, and Essentialism*. Psychology Press. doi: 10.4324/9780203644973

Zunshine, L. (2006). *Why We Read Fiction: Theory of Mind and the Novel*. Columbus, OH: Ohio State University Press.

Zunshine, L. (2008). *Strange Concepts and the Stories They Make Possible: Cognition, Culture, Narrative*. Baltimore, MD: JHU Press.

Zwaan, R. A. (1999). Situation models: the mental leap into imagined worlds. *Curr. Direct. Psychol. Sci.* 8, 15–18. doi: 10.1111/1467-8721.00004

Zwaan, R. A. (2004). The immersed experiencer: toward an embodied theory of language comprehension. *Psychol. Learn. Motiv.* 44, 35–62. doi: 10.1016/S0079-7421(03)44002-4

Zwaan, R. A., Langston, M. C., and Graesser, A. C. (1995). The construction of situation models in narrative comprehension: an event-indexing model. *Psychol. Sci.* 6, 292–297. doi: 10.1111/j.1467-9280.1995.tb00513.x

Zwaan, R. A., and Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychol. Bull.* 123, 162. doi: 10.1037/0033-2909.123.2.162

Conflict of Interest: 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 © 2021 Castano, Paladino, Cadwell, Cuccio and Perconti. 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) and the copyright owner(s) 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 June 2021
doi: 10.3389/fpsyg.2021.556120

Child-Like Adults: Dual-Task Effects on Collective vs. Distributive Sentence Interpretations

Anna M. B. de Koster[1]*, *Petra Hendriks*[1] and *Jennifer K. Spenader*[2]

[1] Center for Language and Cognition Groningen (CLCG), Faculty of Arts, University of Groningen, Groningen, Netherlands,
[2] Department of Artificial Intelligence, Faculty of Science and Engineering, University of Groningen, Groningen, Netherlands

In this work, we consider a recent proposal that claims that the preferred interpretation of sentences containing definite plural expressions, such as "The boys are building a snowman," is not determined by semantic composition but is pragmatically derived via an implicature. Plural expressions can express that each member of a group acts individually (distributive interpretation) or that the group acts together (collective interpretation). While adults prefer collective interpretations for sentences that are not explicitly marked for distributivity by the distributive marker *each*, children do not show this preference. One explanation is that the adult collective preference for definite plurals arises due to a conversational implicature. If implicature calculation requires memory resources, children may fail to calculate the implicature due to memory limitations. This study investigated whether loading Dutch-speaking adults' working memory, using a dual task, would elicit more child-like distributive interpretations, as would be predicted by the implicature account. We found that loading WM in adults did lead to response patterns more similar to children. We discuss whether our results offer a plausible explanation for children's development of an understanding of distributivity and how our results relate to recent debates on the role of cognitive resources in implicature calculation.

Keywords: conversational implicature, distributivity, dual task, language development, pragmatics, quantification, semantics, working memory

OPEN ACCESS

Edited by:
Valentina Cuccio, University of Messina, Italy

Reviewed by:
Stephen Politzer-Ahles, Hong Kong Polytechnic University, China
Xiaoming Jiang, McGill University, Canada
Xiaolin Zhou, Peking University, China

***Correspondence:**
Anna M. B. de Koster
a.m.b.de.koster@rug.nl

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

Received: *27 April 2020*
Accepted: *05 May 2021*
Published: *10 June 2021*

Citation:
de Koster AMB, Hendriks P and Spenader JK (2021) Child-Like Adults: Dual-Task Effects on Collective vs. Distributive Sentence Interpretations. Front. Psychol. 12:556120. doi: 10.3389/fpsyg.2021.556120

INTRODUCTION

An essential feature of language is the ability to refer to groups of individuals. We can talk about these individuals performing actions individually, or together as a group. Consider sentence (1), which contains the plural definite description *the boys*:

(1) The boys are building a snowman.

Sentences with plural definite subjects like (1) are compatible with more than one interpretation according to semantic theories (e.g., Landman, 2000; Champollion, 2017). Are the boys in sentence (1) building one snowman together (the collective interpretation, see **Figure 1**)? Or are they building individual snowmen (the distributive interpretation, see **Figure 2**)? A collective interpretation simply requires that the predicate applies to the set denoted by the plural expression. For example, in a situation with three boys, Al, Ben, and Chris, the collective interpretation of the sentence "Every boy is building a snowman" means that several of the boys are building one snowman together. The distributive interpretation, on the other hand, requires that the predicate applies to each member of the set denoted by the plural expression individually. This would

FIGURE 1 | Collective interpretation.

FIGURE 2 | Distributive interpretation.

necessarily entail that Al is building his own snowman, Ben is building his own snowman, and Chris is building his own snowman.

Experimental studies have shown that adults disprefer distributive interpretations, unless an overt distributive marker like *each* is present (Gil, 1982; Brooks and Braine, 1996; Frazier et al., 1999; Kaup et al., 2002), as in (2). When *each* is present, adults prefer distributive interpretations almost exclusively.

(2) Each boy is building a snowman.

The preference for a collective or distributive interpretation is affected by multiple other factors as well, such as whether or not the action tends to be done with others, e.g., carrying a piano, or tends to be done individually, e.g., drinking an espresso [see also Geurts (2010); de Koster et al. (2020); Kursat and Degen (2020) for more discussion]. However, the presence or absence of distributive marking seems to be the most influential feature for lexically ambiguous predicates [see (Dotlačil and Brasoveanu, 2021) for recent results].

Studies in various languages have shown that children have different preferences than adults when interpreting plural NPs like (1). Whereas, adults prefer collective interpretations, children seem to initially prefer distributive interpretations across the board, even when no distributive marker is present (Miyamoto and Crain, 1991; Avrutin and Thornton, 1994; Brooks and Braine, 1996; Syrett and Musolino, 2013). In fact, we can identify two milestones in children's development of adult-like preferences. First, around age-seven children begin to consistently reject distributive markers like *each* with collective situations (Pagliarini et al., 2012; de Koster et al., 2017, 2018), suggesting that it is not until that age that children fully grasp the distributive import of distributive markers. Second, around the age of nine children begin to reject distributive interpretations if there is no distributive marking, like adults. However, they initially reject these cases at a much lower rate than adults. The rate of rejection increases steadily with age, but existing studies found that at age 11 [de Koster et al. (2018) for Dutch] and even at age 14 [Pagliarini et al. (2012) for Italian], the rates of rejection were still lower than those of adults in the same study.

In this paper, we experimentally investigate a proposal that would simultaneously explain adults' rejection of distributive interpretations with distributively unmarked sentences and, perhaps, offer an account for children's development of adult-like preferences: The implicature account of distributivity preferences developed by Dotlačil (2010). This account proposes that the preference to interpret distributively unmarked sentences as collective results from a conversational implicature. To investigate this proposal, we focus on whether or not adult preferences require working memory resources. We focus on the potential role of working memory resources in adults' distributivity preferences for two reasons. First, many studies of more established conversational implicatures have found evidence that calculating an implicature requires working memory resources. Second, in a study of children's distributivity preference development, de Koster et al. (2018) found a positive correlation between working memory capacity and adult-like preferences, suggesting that working memory plays a role in children's distributivity preferences. We thus investigate the role of working memory in distributivity preferences by limiting adults' working memory with a dual-task design. If limited working memory leads to responses more similar to children, we will have found evidence supporting an implicature account of distributivity and a role for working memory in children's non-adult preferences.

IMPLICATURES AND PROCESSING

Conversational implicatures are language-based inferences about a speaker's intended meaning that listeners make by considering alternative forms the speaker could have used (Grice, 1975). The most researched implicature is triggered by the scalar term "some." Consider (3):

(3) Teacher: "Some of my students passed the exam."

Semantically, (3) is consistent with several students passing the exam, but also with all the students passing the exam, because *some* literally means "at least one." However, most listeners understand (3) to mean that not all students passed, because they will not simply interpret what the speaker literally says, but they will compare it with alternatives that the speaker could have said but did not. Because listeners expect speakers to use the most informative expression consistent with the situation, the choice by the speaker to use a weaker expression will suggest to the listener on the comparison that the stronger expression did not hold. Because *all* semantically entails *some, all* is informationally stronger than *some*, so an utterance with the weaker form *some* then implicates that the speaker believes that the stronger version with *all* does not hold (Horn, 1973). Because this explanation relies on the recognition of a scale of informativity (e.g., <some, all>, <might, must>, <or, and>), these are also often called scalar implicatures.

Conversational implicatures are pervasive in language: Strengthened meanings are associated with specific scalar lexical items like *some* or *most* but also can be calculated on the fly. E.g., in the following short exchange, A: *Did you read War and Peace?* B: *I read the first chapter*, B's statement implicates that B did not read the rest of the book.

Dotlačil's (2010) proposal that sentences with definite plurals are interpreted collectively due to an implicature builds on the accepted view that distributively unmarked sentences, such as sentences with definite plural subjects, can semantically express both a collective and a distributive interpretation (Frazier et al., 1999). In contrast, an utterance with the distributive marker *each*, such as sentence (2), signals the more specific distributive interpretation and is thus informationally stronger than a distributively unmarked utterance like (1), which is ambiguous. A distributively unmarked sentence such as (1) is not specified for collective or distributive meaning, but will be interpreted as collective, because a hearer will reason that if the speaker had intended a distributive interpretation, the speaker would have used the informationally stronger form with the distributive marker *each*. Through the reasoning process underlying implicatures, unmarked sentences are biased to be interpreted collectively.

Note that this account proposes that the collective and the distributive interpretations have different sources: Distributive interpretations arise due to distributive marking, such as *each*, while collective interpretations arise because of the absence of a distributive marker due to an implicature. Because the implicature calculation requires comparing alternatives and, perhaps, others steps not necessary for a literal interpretation, it may involve greater cognitive effort. For this reason, the implicature account offers an explanation of children's non-adult interpretation preferences in terms of processing difficulties.

There are two aspects of the implicature proposal that benefit from a closer examination. First, can distributively unmarked sentences with plural definite descriptions and distributively marked sentences with *each* be analyzed analogously to more traditional implicatures such as *some-not all*? Second, what evidence is there that implicature calculation requires additional cognitive resources, in particular working memory resources?

To see if parallels exist between the well-studied implicature with *some* and *all* and the proposed implicature with plural definites and *each*, let us consider the sentence meanings involved in the implicature calculation in both cases. The set of situations where sentences with *all*, expressing an exhaustive meaning, are true is a proper subset of the set of situations where sentences with *some* under its literal interpretation, "some and possibly all," are true. Thus, literal *some* can be considered to be less informative than *all*: It allows the "some but not all" meaning as well as the exhaustive "all" meaning. More informative sentences with *all* will block the exhaustive meaning with ambiguous *some* so that it is preferably interpreted only as "some but not all," resulting in a pragmatically strengthened non-exhaustive meaning for sentences with *some*.

In a similar fashion, plural definite descriptions and distributively quantified DPs can be analyzed. Plural definite descriptions have a so-called "maximality requirement": the maximal set of the plurality modified by the definite description needs to participate in the predicated action, e.g., in (1) all members of the set of boys must build a snowman. *Each* as a universal quantifier also requires that all members of the restrictor set modified by the quantifier participate exhaustively in the predicated action, e.g., in (2), all members of the set of boys must also participate in snowman building. As a distributive quantifier, *each* imposes an additional distributive requirement, and the result is that the set of situations where sentences with the distributive quantifier *each* (expressing a distributive meaning) are true is a proper subset of the set of situations where sentences with the definite article *the* are true, and the proposition in (2) entails the less specific (1). As such, sentences with *the* can be considered to be less informative than sentences with *each*: They allow a collective as well as a distributive interpretation (e.g., Maldonado et al., 2019). The resulting scale (<plural *the, each*>) also fulfills an additional requirement, emphasized by van Tiel et al. (2019) and originating from Horn (1989), that members of the scale must have the same polarity in that both are positive. According to the proposed distributivity implicature, the more informative sentences with *each* will block the distributive meaning for the less informative sentences with *the*, resulting in a pragmatically strengthened collective meaning for sentences with *the*.

Essential to this analysis is that an implicature with plural definites would only be evoked if a speaker was aware of the unambiguous, distributive import of distributive markers and treat <plural *the*, *each*> as forming a scale. We expect adults to know what lexical items signal distributivity, but children have to learn this. Therefore, the proposal predicts that only children that understand the distributive meaning and the way this meaning can be marked have the prerequisite knowledge for the implicature. Children are generally considered to understand the meaning of distributive markers if they consistently reject them with collective situations, which has been found experimentally to be around age seven (Pagliarini et al., 2012; de Koster et al., 2017, 2018). Only after this age will they be able to infer that the absence of distributive marking conversationally implicates collectivity. This prediction has been previously experimentally investigated by Pagliarini et al. (2012) in Italian,

de Koster et al. (2017) in Dutch, and Padilla-Reyes (2018) in Spanish. All three studies found a correlation between children's rejection of collective situations with distributive marking and children's rejection of distributive situations when distributive marking is not present, consistent with the predictions of the implicature account.

Why would implicature calculation be effortful? And what experimental evidence supports this claim? Implicature accounts differ as to whether they see it as a primarily pragmatic (e.g., Gricean) process, where implicature is the product of the listener's expectation that speakers will be as informative as possible, or as a semantic process (Chierchia, 2004, 2006), where strengthened meaning is argued to originate from an unpronounced operator, O, that signals that stronger alternatives do not apply [i.e., "[O]nly" the literal meaning is intended]. Despite these different assumptions, both pragmatic and semantic accounts of implicature identify at least two steps in implicature calculation that could be potentially cognitively demanding: The decision whether or not to calculate an implicature and the derivation and comparison of alternatives, both considered to be pragmatically driven processes (see, Geurts, 2010; e.g., Chemla and Singh, 2014).

Multiple studies have, in fact, found evidence that verifications of sentences with *some-not all* implicatures take more time than the lower-bounded, literal interpretation of *some*. This has been shown with timed sentence verification (Bott and Noveck, 2004), self-paced reading (Breheny et al., 2006; Chemla and Bott, 2013), eye-tracking (e.g., Huang and Snedeker, 2009), mouse tracking (Tomlinson et al., 2011), and looking at response times, with speed accuracy trade-off (SAT) methods (e.g., Bott et al., 2012). Most relevant for our research, however, are the experiments that focused on determining if there is a memory cost involved in the calculation of scalar implicature, using dual-task designs. These experiments ask the participants to judge sentences while their working memory is loaded (De Neys and Schaeken, 2007; Dieussaert et al., 2011; Marty and Chemla, 2013; Marty et al., 2013; van Tiel et al., 2019; Ryzhova and Demberg, 2020). Most of these studies had participants memorize dot patterns on a 3 × 3 matrix, which they then had to recreate after judging sentences that could invoke an implicature. Most studies compared linguistic performance on low-working memory load patterns, with a systematic three-dot pattern (all horizontal or all vertical), to performance with high-load patterns, which had four dots. De Neys and Schaeken (2007) found that participants under a high-working memory load made significantly fewer *some-not all* implicatures (around 10% less) compared to a low-working memory load. Dieussaert et al. (2011) used an identical design but investigated the role of individual working memory capacity by also measuring participants' working memory spans in a separate task. They found fewer implicature calculations under a high load but only for participants with low-working memory capacity. Marty and Chemla (2013) also used a similar design and found that loading working memory decreased the rate of implicatures but had no effect on semantically equivalent *only some* sentences where the negation of the alternative is made explicit, a result that is unexpected under a semantic account of implicature[1].

The type of implicature may also influence whether or not working memory resources are involved. Marty et al. (2013) used a dual-task design with a backward letter sequence reproduction task. They found a decrease in implicature interpretations with *some-not all* implicatures under a high memory load but found no effect with number items, which should implicate an exact interpretation (e.g., *three* means *three and no more*)[2]. In a recent study, van Tiel et al. (2019) used a dual-task design to investigate several scalar words, using a between-subjects design to compare participants under no load, a low working memory load, and a high working memory load. They found that the influence of loading working memory varied by implicature type: Some implicatures, such as *some-not all*, showed a lower rate of calculation already in the low-load condition compared to the no-load condition, whereas others only showed an effect between the low-load and the high-load conditions. Ryzhova and Demberg (2020) also carried out a dual-task study, using dot-tracking as the secondary task. Participants made fewer particularized conversational implicatures under a high memory load compared to a low-memory load. In summary, the existing dual-task studies have all found that loading working memory with a dual task lowers the rate of implicature calculation.

In contrast to the dual-task studies, which focused on working memory resources, a number of other studies, which are primarily eye-tracking studies, have failed to find evidence that implicature calculation requires additional cognitive resources. In a visual-world study, Grodner et al. (2010) failed to find that *some-not all* implicature calculations required greater processing times when they modified the materials from (Huang and Snedeker, 2011). Politzer-Ahles and Fiorentino (2013), Hartshorne and Snedeker (2014), in an eye-tracking reading experiment, and Politzer-Ahles and Husband (2018), all also failed to find evidence of additional processing costs for implicatures. For implicatures based on the *<not all, none>* scale, Cremers and Chemla (2014) (Exp. 1) and Romoli and Schwarz (2015) both found that participants were faster at implicature calculation than the literal interpretation. In a visual world eye-tracking study, Degen and Tanenhaus (2016) found no difference between pragmatic and literal interpretations of *some*. Kursat and Degen (2020), investigating reaction times with *some-not all* implicatures found evidence for two populations: pragmatic responders who tended to always calculate implicatures and who were faster at these interpretations than at lower-bounded interpretations, and literal responders. Like the results from Dieussaert et al. (2011), this

[1]Because *only some* is semantically equivalent with the some-not all implicature it offers a useful control condition to determine whether the implicature itself requires working memory resources rather than, e.g., verification. Unfortunately, for definite descriptions there is not an obvious semantic equivalent that could be used as a control expression. Future research will have to investigate this question with other methods.

[2]Note however that it is currently a matter of debate whether or not numerals should be expected to give rise to implicatures. See Marty and Chemla (2013) for a discussion.

shows that there may be individual differences in implicature calculation tendencies.

Another issue in experimental implicature research is that many studies focus on participants' end-state judgments without further information about the interpretation process. Because end-state judgments are a culmination of multiple interpretation processes, both semantic and pragmatic,[3] it can be hard to determine what influences the participants' final judgment.

If implicatures do require additional resources, exactly what resources and at what step in their interpretation these become relevant is still a topic of investigation. Dotlačil's (2010) implicature account of distributivity preferences claims to both explain adults' preferences and offer an account for children's non-adult preferences by attributing them to children's difficulties in calculating implicatures. Dotlačil does not make specific claims about what aspect of distributivity interpretation preferences might require processing resources. But because many experimental studies did find a role for working memory in implicature calculation, and because de Koster et al. (2018) found that children's tendency to reject distributive situations without distributive marking was found to be positively correlated with these children's working memory capacity, we decided to investigate the role of working memory in distributivity preference in adults.

If we find that limiting working memory capacity in adults decreases their rates of acceptance of distributive readings with distributively unmarked sentences, then children's tendency to allow distributive readings with distributively unmarked sentences might be explained as a consequence of their lower working memory capacity. This result would then be consistent with the predictions of Dotlačil's implicature account of distributive preferences. If we fail to find an effect of limiting working memory capacity, then this would not rule out an implicature calculation if there is no processing cost for implicature (as some research has found). However, this would make the explanation less attractive in that it would fail to offer a processing explanation for children's non-adult interpretation preferences.

We designed our experiment along the lines of previous dual-task experiments, investigating the influence of working memory on implicature calculation. To our knowledge, we are the first to study distributivity interpretations while limiting working memory capacity. Our study thus provides novel empirical evidence, illuminating the role of memory in the interpretation of plural definites.

For practical reasons, we carried out our experiment in Dutch. English has two distributive quantifiers: *each* and *every*. *Each* and *every* are both universal quantifiers that are compatible with distributivity. Whereas *every* only requires partial distributivity, *each* requires full distributivity (Tunstall, 1998), in that a distributive sentence like (2) must entail that for each individual member of the set of boys, it must hold that he was building a snowman. Dutch also has two distributive quantifiers, *elke* and *iedere*. However, experimentally, these quantifiers have been shown to have the exact same interpretation with respect to distributivity preferences (van der Ziel, 2012; Spenader and Bosnic, 2018), so we will simply use *elke*. Research also suggests that both are more similar to *every* than *each*, being partially distributive (Tunstall, 1998) and thus compatible with collective situations in some cases (Rouweler and Hollebrandse, 2015; de Koster et al., 2017). For this reason, we expect that we will see greater acceptance rates of *elke* with collective situations than has been found for *each*, which is fully distributive.

[3]We thank an anonymous reviewer for stressing this point.

METHOD

Participants

Fifty-eight students from the University of Groningen were paid to participate. They were divided into two groups: a WM Load group (42 participants; 13 men; mean age, 21.9; age range, 18–27) and, to establish a baseline for performance, a No WM load group (16 participants: six men; mean age, 23.9; age range, 20–28). All participants were native speakers of Dutch. This study was carried out in accordance with the recommendations of the Research Ethics Committee (CETO) of the Faculty of Arts of the University of Groningen, and they approved the protocol. We also obtained written informed consent from all the participants prior to testing.

Design

The WM Load group carried out a dual-task experiment, consisting of two tasks, a linguistic task and a memory load task: while participants interpreted a sentence (the linguistic task), we manipulated their WM load by asking them to memorize a sequence of digits (the digit-span task). For this group, the experiment had a 2 × 2 × 2 design with the factors PICTURE [collective (**Figure 1**) vs. distributive (**Figure 2**)], SENTENCE (*de* "the" vs. *elke* "each") and WM LOAD (low vs. high).

The No WM load group received the linguistic task without the digit span task. This group was not tested in a WM Load condition and, therefore, received the experiment in a 2 × 2 design, with the factors PICTURE and SENTENCE but without the factor WM LOAD. The remaining procedure was the same for the two groups.

Linguistic Task

The linguistic task was a sentence–picture verification task. Participants saw a picture on the computer screen and had to judge whether it matched a recorded sentence by pressing a key on the keyboard.

Digit-Span Task

At the start of each trial, participants had to memorize a sequence of three or six digits (low and high WM load conditions, respectively), presented on screen for 1 s each. Digits were randomly chosen from 1 to 9, and consecutive digits always differed. After each linguistic item, participants had to recall the digits by typing them in the same order as they appeared.

Materials

The materials consisted of four practice items, 64 test items, 48 implicature control items and 16 task control items (eight true and eight false items), resulting in a total of 132 items.

The experiment was divided into two blocks (preceded by four practice trials), with 64 items per block. The low (three digits) and high (six digits) WM load conditions were presented in different blocks. Block order (either low or high WM load in the first block) was a between-subject factor. The Latin-square design of the test items, together with the factor block order, resulted in eight lists. Item order was randomized for each participant, and the participants were randomly assigned to a list.

Test Items

The 64-test items tested the factors PICTURE [collective (**Figure 1**) vs. distributive (**Figure 2**)] and SENTENCE [*de* "the" (4) vs. *elke* "each" (5)].

(4) De jongens bouwen een sneeuwpop.
The boys are building a snowman.
(5) Elke jongen bouwt een sneeuwpop.
Each boy is building a snowman.

Eight different transitive verbs were used: build, wash, push, pull, carry, lift, hold or paint (in Dutch: *bouwen*, *wassen*, *trekken*, *duwen*, *dragen*, *tillen*, *vasthouden*, *verven*), and the grammatical subjects and objects of these verbs varied across the items. The design resulted in four conditions: The-Collective, The-Distributive, Each-Collective, and Each-Distributive. The items of condition The-Distributive test proposed implicature. Each participant saw 16 items in each of the four conditions, resulting in 64 test items (32 items per block).

Implicature Control Items

In addition to the test items, participants also received 48 implicature control items to mask the goal of the experiment and to be able to compare the results of the test items to the results of the well-investigated *some-not all* implicature in a dual-task setting.

(6) Sommige jongens vissen.
Some boys are fishing.
(7) Enkele meisjes dansen.
Some girls are dancing.

The implicature control items consisted of two sentence types with the scalar expressions *sommige* "some$_1$" (6) or *enkele* "some$_2$" (7) (24 items per sentence type). In contrast to *enkele*, which merely expresses existential quantification ("there are some…"), *sommige* additionally indicates that the individuals introduced by the quantifier have something in common that distinguishes them from other individuals (de Hoop and Kas, 1989; Banga et al., 2009). For each item, a different intransitive verb was used (e.g., fishing, dancing, singing, and sleeping).

Both sentence types were combined with four picture types where either zero, one, two or three (i.e., all) of the three actors are performing the action denoted by the sentence. This resulted in eight implicature control combinations. The participants received six items per combination (three items per block). The implicature control items were not constructed via a Latin-square design: All the participants received the same sentence-picture combinations as implicature control items. The examples of the implicature control items are presented in the **Supplementary Material**.

The implicature control items with a picture with three actors serve as a control to test whether the participants generate a *some-not all* implicature and whether or not this implicature generation is affected by the dual-task setting.

Task Control Items

The participants also received 16 task control items. These control items were straightforwardly true or false items and were used to check the participants' attention as well as general task effects such as a possible "yes"-bias. Examples of a true (8) and a false (9) task control item are presented below. The corresponding pictures for items (8) and (9) can be found in the **Supplementary Material**. The experiment contained eight true task control items and eight false task control items. If the participants answered more than 25% of the task control items incorrectly, they were excluded from the analysis.

(8) De jongen drinkt een pakje melk.
The boy is drinking a carton of milk.
(9) Het meisje drinkt een glas limonade.
The girl is drinking a glass of lemonade.

Procedure

The participants performed the experiment in a quiet room at the University of Groningen. Participants were shown the pictures on the computer screen while the sentences were played via a speaker. The experimenter was present during the entire experiment.

The experiment started with instructions and four practice trials. For each trial, the participants first saw a digit sequence on screen, followed by a picture and a recorded sentence. The recorded sentence was played only once. They then had to judge sentence acceptability by pressing a green (accept) or red (reject) key. Finally, they had to type in the memorized digits. Participants had 10 s to judge the sentences, with a visual warning message after 7 s. Next, they had 5 s to recall the digits in the low WM load condition and 10 s in the high WM load condition. Pilot testing had shown that this provided the participants with sufficient time.

Each trial ended with feedback to participants on how many digits were recalled correctly. A waiting penalty ensured that participants focused on the WM task and prevented rushing: One incorrect digit resulted in a 1-s waiting penalty, two incorrect digits in a 2-s waiting penalty, etcetera. Self-paced breaks were provided after every 16 items, and the participants had a forced break of at least 2 min in between the two blocks.

The procedure for the two participant groups was similar, including breaks, with the exception that the No WM load group only received the linguistic task. Per trial, the following measures were collected: Accuracy of reproducing the digits in the digit-span task, and yes/no responses and response times for the test items, implicature control items, and task control items in the linguistic task.

Predictions

The linguistic task tests four conditions: The-Distributive, Each-Collective, The-Collective, and Each-Distributive.

Condition The-Distributive tests whether working memory limitations play a role in children's non-adult acceptance of distributive readings with unmarked sentences. This condition, therefore, also tests the implicature account of distributivity preferences. A "yes" response (acceptance) in this condition would be consistent with a literal interpretation of the distributively unmarked sentence, and a "no" response (rejection) would be consistent with the derivation of an implicature. If there is an effect, we also expect the participants in the No WM load condition and the participants under a low WM load to show a higher rate of rejection than the participants under a high WM load. This result would then be parallel to previous findings for the "some-not all" implicature (De Neys and Schaeken, 2007; Dieussaert et al., 2011; Marty and Chemla, 2013; Marty et al., 2013). This result would also be consistent with the child language data, showing that rejection of the The-Distributive condition in children correlates with working memory capacity (de Koster et al., 2018). Importantly, this is the only condition we expect to be affected by WM load. In addition, we also expect to see higher response times for pragmatic responses compared to literal responses, following the findings of Bott and Noveck (2004).

We do not expect condition Each-Collective to be affected by WM load, and we expect the items of this condition to be rejected due to the distributive character of *each*. Experimental results show that from around age 7;0, children are adult-like in their responses to distributively marked sentences with collective situations. Earlier non-adult acceptance is attributed to children being unaware of the distributive nature of *each*. However, it should be noted that the Dutch *elke* (tested in the current study), contrary to its English counterpart *each*, is only partially distributive and has been found to be more acceptable in collective situations (Rouweler and Hollebrandse, 2015; de Koster et al., 2017). Acceptances of items in this condition are, therefore, not unexpected but are predicted to be independent of WM load.

Condition The-Collective is predicted to be unaffected by WM load too. Sentences with *the* are semantically ambiguous between a collective and a distributive interpretation, so the collective interpretation is a semantically appropriate interpretation for sentences with plural definite subjects. This prediction is supported by child language data, showing that children fully accept *the* sentences with a collective interpretation from age 4 and onward (Italian: Pagliarini et al., 2012; Dutch: de Koster et al., 2017; Spanish: Padilla-Reyes, 2018). The implicature account of distributivity preferences does not predict that a WM load would have an effect on this condition either.

Finally, condition Each-Distributive is also predicted to be unaffected by a WM load. The items in this condition are predicted to be fully accepted since the distributive quantifier *each* is fully compatible with the distributive interpretation. Child language data show children from age 4 until age 14 fully accept distributive interpretations for sentences with *each* (Pagliarini et al., 2012; de Koster et al., 2017; Padilla-Reyes, 2018), which is expected because *each* is semantically distributive.

The implicature control items, testing the "some-not all" implicature, can be used to make a comparison with the results of previous studies testing this implicature (De Neys and Schaeken, 2007; Dieussaert et al., 2011; Marty and Chemla, 2013; Marty et al., 2013). The items of our implicature control "Some-All" condition consist of sentences with *sommige* "some" in combination with a picture in which all the actors are performing the action denoted by the predicate. A "yes" response in this condition indicates a literal interpretation, in which the "*at least one and possible all*" meaning of "*some*" is accepted. A "no" response, on the other hand, indicates a pragmatic interpretation in which *some* is interpreted as "*some but not all*." We expect to see fewer pragmatic "no" responses under a high WM load, similar to the findings of previous studies (De Neys and Schaeken, 2007; Dieussaert et al., 2011; Marty and Chemla, 2013; Marty et al., 2013). Following the findings of Bott and Noveck (2004) and others, we also expect to observe longer response times for these responses because calculating an implicature comes at a cost.

RESULTS

Two participants of the WM Load group were excluded from the analysis: One participant did not complete the experiment due to technical problems, and one participant gave incorrect answers to more than 25% of the task control items. All remaining participants were included in the analysis.

Digit-Span Task

The WM Load group participants remembered 94% of the digits correctly in the low WM load condition (three digits) and 75% of the digits in the high WM load condition (six digits). This drop in performance is significant [paired-$t_{(39)} = 13.873$; $p < 0.001$], indicating that the high WM load condition was, indeed, more difficult. Furthermore, the linguistic condition had no effect on the percentage of correctly recalled digits. This shows that the participants focused on digit recall performance throughout the experiment, irrespective of linguistic condition.

Linguistic Task

Responses to Test Items

Figure 3 shows the mean acceptance rates for all four linguistic conditions per WM load. The collected data were analyzed, using generalized mixed effect logistic modeling [function glmer(): lme4 package in R, version 3.6.3].

The models were constructed via an iterative forward fitting procedure with model comparisons (cf. Baayen et al., 2008; Wieling et al., 2011; Ko et al., 2014) based on the evaluation of Akaike's information criterion (AIC) (cf. Burnham and Anderson, 2002; Akaike, 2011; Ko et al., 2014; Wieling et al., 2014). An AIC decrease of more than two indicates that the goodness of a fit of the model improves significantly (Akaike, 2011). The AIC values were obtained via model comparisons, using ANOVA testing [function anova() in R]. We determined whether the following fixed-effect factors improved the goodness of fit of the model: SENTENCE (each, the), PICTURE (collective, distributive), BLOCK (first, second), WM LOAD (no load, low, high), and VERB. The dependent variable was the response (0 for rejection, 1 for acceptance). The final model (**Table 1**) included the fixed factors SENTENCE, PICTURE, WM LOAD, and BLOCK.

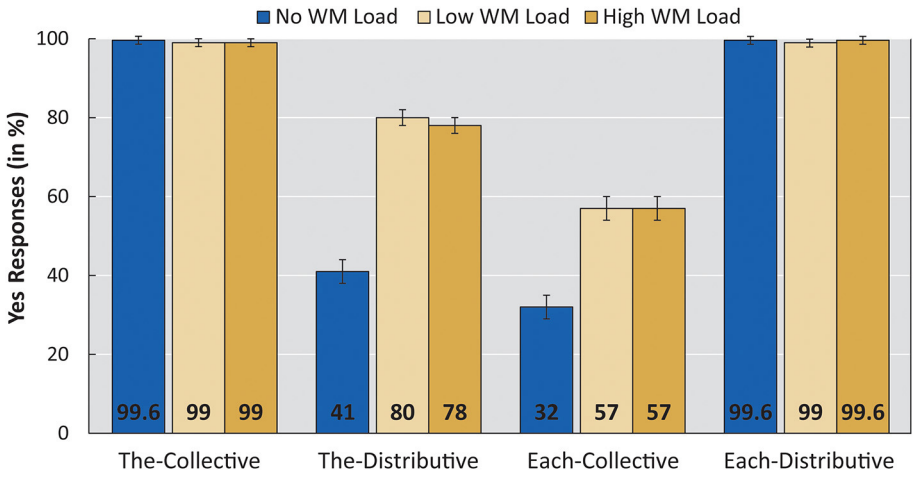

FIGURE 3 | Mean acceptance rates per linguistic condition per WM load (No load, Low, High). Error bars show standard error. The WM Load group was tested with Low and High WM load; the No WM load group was tested on the linguistic task only. Sentences contained either *de* "the" or *elke* "each" and pictures showed either a collective action or a distributive action.

TABLE 1 | Overview of the final model for the responses to the test items, with reference levels: Sentence: "De" "The," Picture: Distributive, WM Load: No Load, and Block: 1.

Predictors	Estimate (β)	SE	*z*	*p*
Intercept	0.3307	0.844	0.695	< 0.001
Sentence "Elke" "Each"	13.316	2.528	5.267	< 0.001
Picture collective	9.470	2.193	4.319	< 0.001
WM load low	2.809	1.049	2.679	< 0.01
WM load high	2.730	1.048	2.605	< 0.01
Block 2	−0.419	0.356	−1.180	0.238
Sentence "Elke" × Picture Collective	−24.392	3.584	−6.807	< 0.001

The factor VERB did not significantly improve the model fit and was left out. The maximal random-effects structure licensed by the data included a random intercept for participants, items and by-participant random slopes for SENTENCE, PICTURE, and BLOCK.

A main effect of SENTENCE and PICTURE was found, as well as an interaction between the two. Crucial to our findings is the significant difference between the No WM load group and each of the low and high WM load conditions of the WM Load group (low: $\beta = 2.809$; $z = 2.679$; $p = 0.007$, high: $\beta = 2.730$; $z = 2.605$; $p = 0.009$). In line with our predictions, participants accepted significantly more items of condition The-Distributive in the WM load group compared with the No WM load group.

A releveled model revealed no significant difference between the low and high WM load conditions within the WM load group ($\beta = -0.079$; $z = -0.235$; $p = 0.814$). The low and high WM load conditions had acceptance rates of condition The-Distributive to the same degree.

To check for an influence of the factor BLOCK-ORDER (low or high WM load in the first block), we performed a separate analysis on the data of the WM load group, with a similar model. In this model, BLOCK-ORDER did not improve the model fit. This shows that participants' acceptance of The-Distributive items was not influenced by whether they received the low or high WM load condition first. Note that it is not possible to add the factor BLOCK-ORDER to the final model presented in **Table 1** (analyzing the WM load and the No WM load group together) for reasons of collinearity.

Responses to Implicature Control Items

We also analyzed the responses of the implicature control items with *sommige* "some$_1$" and *enkele* "some$_2$" in combination with pictures where three (i.e., all) actors are performing the action denoted by the sentence, to check how our WM manipulation affected the "some-not all" implicature. These items are compared to implicature control items with pictures where two actors are performing the action denoted by the sentence, since these latter items do not give rise to an implicature and should be considered true. **Figure 4** presents the results for these items. The remaining implicature control combinations (with pictures with zero actors or one actor) were omitted from the figure and further analysis, since they are expected to be judged as false, because sentences with Dutch *sommige* "some$_1$" and *enkele* "some$_2$" require reference to at least two individuals (Broekhuis and den Dikken, 2012, p. 895).

Models were again constructed via an iterative forward fitting procedure. Using model comparisons (cf. Baayen et al., 2008) based on the AIC values, we determined which fixed-effect factors would improve the model fit. The final model (**Table 2**) included the fixed factors SENTENCE, PICTURE, and WM LOAD. The dependent variable was the response (0 for rejection, 1 for acceptance). The maximal random-effects structure licensed by the data included a random intercept for participants and by-participant random slopes for SENTENCE and PICTURE.

A main effect of SENTENCE and PICTURE was found, but an interaction between the two did not improve the model fit.

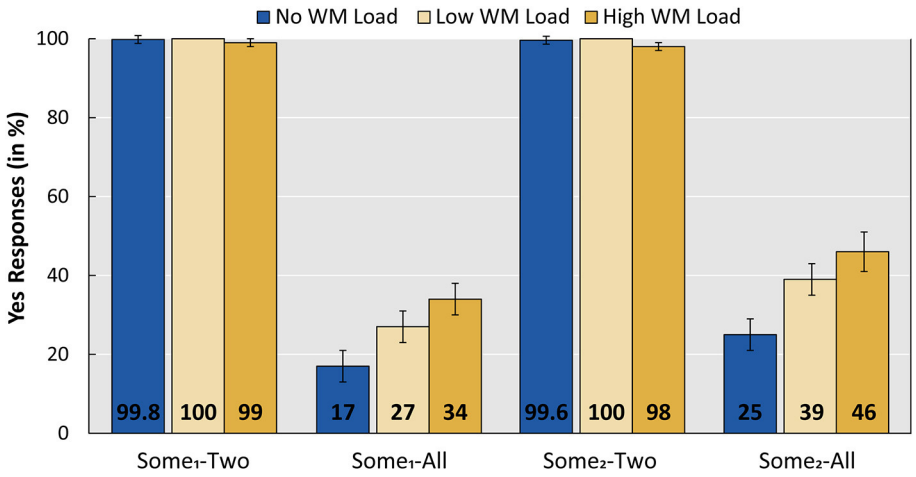

FIGURE 4 | Mean acceptance rates per implicature control condition per WM load (No load, Low, High). Error bars show standard error. The WM Load group was tested with Low and High WM load; the No WM load group was tested on the linguistic task only. Sentences contained either *sommige* "some1" or *enkele* "some2," and pictures showed either two out of three actors or all three actors performing the action denoted by the sentence.

TABLE 2 | Overview of the final model for the responses to the implicature control items, with reference levels: Sentence: "Sommige" "Some$_1$," Picture: All, WM Load: high.

Predictors	Estimate (β)	SE	*z*	*p*
Intercept	−1.849	0.755	−2.447	< 0.05
Sentence "Enkele" "Some$_2$"	1.448	0.456	3.177	< 0.01
Picture two	9.163	1.636	5.601	< 0.001
WM load no load	−2.188	1.089	−2.009	< 0.05
WM load low	−0.653	0.311	−2.099	< 0.05

Crucial, however, is the significant difference between the high WM load condition (six digits) and both the low WM load condition (three digits) and the No WM load group (low: $\beta = -0.653$; $z = -2.099$; $p = 0.0358$, no load: $\beta = -2.188$; $z = -2.009$; $p = 0.0445$). In line with the predictions regarding the "some-not all" implicature, participants accepted significantly more items of condition Some$_1$-All (accepting the literal interpretation of "some") under a high WM load. A releveled model with "some$_2$" *enkele* as the reference level revealed that a similar pattern holds for condition Some$_2$-All.

An additional releveled model (with the low WM load as a reference level) revealed no significant difference between the low WM load condition and the No WM load group ($\beta = 1.535$; $z = 1.421$; $p = 0.155$). This means that the participants' calculation of "some-not all" implicatures was only affected by the high WM load condition of six digits.

Responses to Task Control Items

The task control items (straightforwardly true or false items) were included to check for participants' attention to the linguistic task. Overall, participants answered 95% of all task control items correctly, which shows that they paid sufficient attention to the linguistic task.

To investigate whether the difference we found in acceptance rates between the WM load group and the No WM load group for the test items could be attributed to a general tendency to more readily accept items when WM is loaded, we also analyzed participants' performance on the false task control items (that required a "no" response).

The WM load group participants answered 89% of the false control items correctly in the low WM load condition (three digits) and 88% of the false control items in the high WM load condition (six digits). This difference was not significant [paired-$t_{(39)} = 0.443$; $p = 0.660$].

Participants of the No WM load group answered 92% of the false control items correctly. This did not differ from the false control item performance of the WM load group [unpaired-$t_{(36)} = -1.516$; $p = 0.138$].

The results from the false task control items indicate that participants were not simply more accepting of experimental items because of the cognitive burden of the secondary task, but, instead, the difference in acceptance rates between the WM load group and the No WM load group must have a different explanation such as e.g., the costs associated with calculating the interpretation.

Response Times of Test Items

To test the assumption that implicature calculations require more time (in addition to memory resources), we also analyzed response times (RTs). **Figure 5** presents boxplots of RTs for all four linguistic conditions for "yes" and "no" responses separately and per WM load. RTs were measured from the onset of picture and sentence presentation until button press. Outliers were excluded following the interquartile range rule, excluding data points that are more than one and a half times the interquartile range below the first and above the fourth quartile (5.4% of the data was removed).

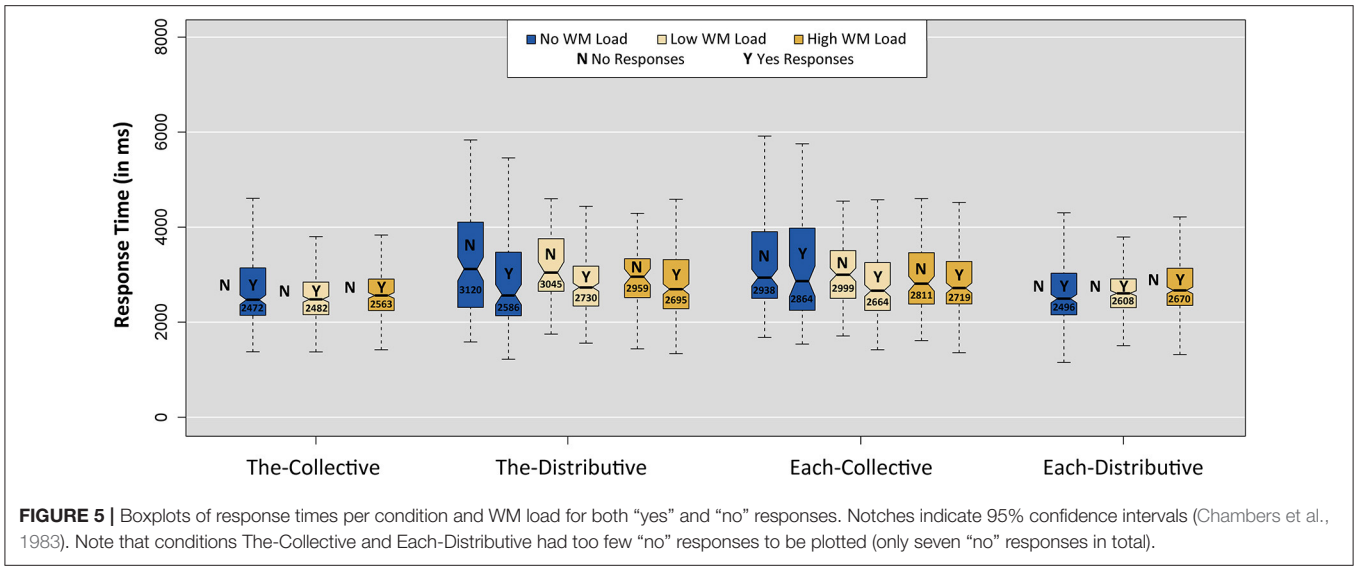

FIGURE 5 | Boxplots of response times per condition and WM load for both "yes" and "no" responses. Notches indicate 95% confidence intervals (Chambers et al., 1983). Note that conditions The-Collective and Each-Distributive had too few "no" responses to be plotted (only seven "no" responses in total).

We performed three different analyses. We first wanted to find out whether it was, indeed, the case that the pragmatic interpretation of the proposed implicature (a "no" response in condition The-Distributive) takes more time than the literal interpretation (a "yes" response in condition The-Distributive). This difference was previously observed by Bott and Noveck (2004) for the "some-not all" implicature.

Second, we wanted to examine whether or not the response times of the pragmatic interpretation were influenced by a WM load. In their dual-task study, De Neys and Schaeken (2007) found that pragmatic interpretations under a high WM load took significantly longer than pragmatic interpretations in the control condition involving only a low WM load.

We also looked more closely into the response pattern of condition Each-Collective. The responses indicated that loading adults' WM not only increased their acceptance of The-Distributive items but also their acceptance of Each-Collective items (see **Figure 3**). This is unexpected, since we do not predict an implicature in this condition. The verification of this condition is expected to be based on the semantics of Dutch *elke* and should, therefore, not be influenced by a limited WM capacity. Analyzing the response times of condition Each-Collective can show if a similar process underlies the interpretations of the items in conditions The-Distributive and Each-Collective.

We did not analyze the response times of the implicature control items because there were too few items for a proper analysis.

Pragmatic vs. Literal Interpretations

To find out whether or not the proposed pragmatic interpretations, indeed, took more time than the proposed literal interpretations, we analyzed the log-transformed RTs of condition The-Distributive, using linear mixed effect modeling [function lmer(): lme 4 package in R, version 3.6.3]. We included the factor RESPONSE, separating the RTs of the "no' responses from the "yes" responses, since a "no" response indicates a pragmatic interpretation and a "yes" response indicates a literal interpretation.

TABLE 3 | Overview of the model comparing the response times of the pragmatic and literal interpretations, with reference levels: Response: "No," and Block: 1.

Predictors	Estimate (β)	SE	df	*t*	*p*
Intercept	8.127	0.040	47	200.159	< 0.001
Response "Yes"	−0.168	0.039	36	−4.288	< 0.001
Block 2	−0.165	0.033	92	−4.988	< 0.001
Response "Yes" × Block 2	0.124	0.036	99	3.395	< 0.001

Based on model comparisons, using the Akaike Information Criterion, the final model (**Table 3**) included the fixed factors RESPONSE and BLOCK. We also included random intercepts for participants, items, and by-participant random slopes for RESPONSE and BLOCK. The dependent variable was the response time in milliseconds (log-transformed).

A main effect of RESPONSE and BLOCK was found, as well as an interaction between the two. Crucially, the main effect of RESPONSE ($\beta = -0.168$; $t = -4.288$; $p < 0.001$) indicates that "no" responses in condition The-Distributive were significantly slower than "yes" responses, following the prediction that pragmatic interpretations take more time than literal interpretations.

The significant effect of the factor BLOCK ($\beta = -0.165$; $t = -4.988$; $p < 0.001$) indicates that participants were generally faster in block 2 compared to block 1. The significant interaction between RESPONSE and BLOCK ($\beta = 0.124$; $t = 3.395$; $p < 0.001$) indicates that the difference between the responses ("yes" and "no") is smaller in block 2 compared to block 1. These findings could be explained as an effect of task experience. Similar block effects have been found in other dual-task studies (e.g., van Rij et al., 2013).

To be sure that the difference in RTs between the pragmatic and literal interpretation was not caused by the possibility that

TABLE 4 | Overview of the model comparing the response times of the "no" responses in conditions The-Distributive and Each-Collective, with reference levels: Condition: "The-Distributive," and Block: 1.

Predictors	Estimate (β)	SE	df	*t*	*p*
Intercept	8.159	0.035	49	234.035	< 0.001
Condition Each-Collective	−0.061	0.027	28	−2.279	< 0.01
Block 2	−0.140	0.018	55	−7.926	< 0.001

TABLE 5 | Overview of the model examining the influence of WM load on the pragmatic interpretation, with reference levels: WM Load: "No Load," and Block: 1.

Predictors	Estimate (β)	SE	df	*t*	*p*
Intercept	8.219	0.059	32	139.32	< 0.001
WM load low	−0.086	0.077	38	−1.120	0.270
WM load high	−0.099	0.077	38	−1.283	0.207
Block 2	−0.155	0.031	224	−4.946	< 0.001

it would take longer in general to provide a "no" response than a "yes" response, regardless of the tested condition, we also compared RTs on "no" responses in the The-Distributive condition (requiring the hypothesized implicature) to RTs on "no" responses in the Each-Collective condition (not requiring an implicature).

Based on model comparisons, using the Akaike Information Criterion, the final model (**Table 4**) included the fixed factors CONDITION and BLOCK. We also included random intercepts for participants, items, and by-participant random slopes for CONDITION. The dependent variable was the response time in milliseconds (log-transformed).

We found a main effect of CONDITION ($\beta = -0.061$; $t = -2.279$; $p < 0.01$), as well as BLOCK ($\beta = -0.140$; $t = -7.926$; $p < 0.001$). The main effect of BLOCK shows that participants' "no" responses were faster in block 2 than in block 1, similar to the previously found BLOCK effect. The main effect of CONDITION shows that "no" responses in the Each-Collective condition were significantly faster than "no" responses in the The-Distributive condition. This finding indicates that the difference in RTs between the pragmatic interpretation ("no" response) and the literal interpretation ("yes" response) in condition The-Distributive was not caused by a general difference in RTs between "yes" and "no" responses.

Implicature Interpretations Under WM Load

We also examined the influence of a WM load on the response time for rejections of the The-Distributive condition. We, therefore, analyzed the log-transformed RTs of the pragmatic interpretations ("no" responses) of condition The-Distributive, using linear mixed effect modeling [function lmer (): lme 4 package in R, version 3.6.3].

Based on model comparisons, using the Akaike Information Criterion, the final model (**Table 5**) included the fixed factors WM LOAD and BLOCK. We also included random intercepts for participants, items, and by-participant random slopes for BLOCK. The dependent variable was the response time in milliseconds (log-transformed).

The factor WM LOAD was not a significant predictor, showing that there was no difference in RTs between the no WM load condition and both the low WM load condition ($\beta = -0.086$; $t = -1.120$; $p = 0.270$) and the high WM load condition ($\beta = -0.099$; $t = -1.283$; $p = 0.207$). A releveled model with the low WM load condition as the reference level showed that there was no difference between the low and high WM load conditions either ($\beta = -0.01$; $t = -0.273$; $p = 0.708$). These findings are different from the findings by De Neys and Schaeken (2007), who found that pragmatic interpretations under a high load took longer than pragmatic interpretations in the low-load condition. Although not significant, in our model, the estimates of the factor WM LOAD were negative, suggesting that participants became faster under a WM load. This is the opposite direction as the results found by De Neys and Schaeken (2007) where the participants became slower. One reason may be the differences between the tasks. De Neys and Schaeken (2007) used a dot-pattern task, which might have been easier than our digit-span recall task. In our task, the participants may attempt to decrease the WM load by speeding up their responses, thus reducing the period of time during which they need to remember the digits.

The final model did include a main effect of the factor BLOCK ($\beta = -0.155$; $t = -4.946$; $p < 0.001$), again showing that the participants' response times were lower in block 2, probably as a result of task experience.

Response Times in Condition Each-Collective

To check for a difference in RTs within condition Each-Collective, we analyzed the log-transformed RTs of condition Each-Collective, using linear mixed effect modeling [function lmer (): lme 4 package in R, version 3.6.3]. We included the factor RESPONSE, separating the RTs of the "no" responses from the "yes" responses to be able to find out whether there is a difference in RTs between the different response types, like we found in condition The-Distributive.

Based on step-wise model comparisons, using the Akaike Information Criterion, the final model (**Table 6**) included the fixed factors RESPONSE, WM LOAD, and BLOCK. We also included random intercepts for the participants, items, and by-participant random slopes for RESPONSE and BLOCK. The dependent variable was the response time in milliseconds (log-transformed).

Although the final model based on model comparisons included a three-way interaction between the factors RESPONSE, WM LOAD, and BLOCK, we only found a main effect of BLOCK ($\beta = -0.185$; $t = -4.313$; $p < 0.001$) and an interaction between RESPONSE and BLOCK ($\beta = 0.258$; $t = 3.380$; $p < 0.001$). Crucially, the factor RESPONSE did not turn out to be significant, showing that there was no difference in RTs between "yes" and "no" responses in condition Each-Collective. The participants reacted similarly to "no" responses as to "yes" responses. This is in contrast with response latencies in the The-Distributive condition, where we did find a difference between the response types. The difference in response times between conditions The-Distributive and Each-Collective points to a different process underlying the interpretations of the items in both conditions.

TABLE 6 | Overview of the model comparing the response times of the "yes" and "no" responses in condition Each-Collective, with reference levels: Response: "No," WM Load: "No Load," and Block: 1.

Predictors	Estimate (β)	SE	df	*t*	*p*
Intercept	8.148	0.047	41	171.948	< 0.001
Response "Yes"	−0.044	0.058	34	−0.747	0.460
WM load low	−0.103	0.066	38	−1.558	0.012
WM load high	−0.086	0.071	38	−1.212	0.233
Block 2	−0.185	0.043	65	−4.313	< 0.001
Response "Yes" × WM Load low	−0.090	0.081	37	−1.124	0.269
Response "Yes" × WM Load high	−0.057	0.085	40	−0.672	0.505
Response "Yes" × Block 2	0.258	0.076	151	3.380	< 0.001
WM Load low × Block 2	0.090	0.084	74	1.070	0.288
WM Load high × Block 2	0.033	0.085	79	0.387	0.700
Response "Yes" × WM load low × Block 2	−0.118	0.117	99	−1.009	0.316
Response "Yes" × WM load high × Block 2	−0.147	0.118	102	−1.247	0.215

DISCUSSION

In the present study, we investigated whether the adult preference for collective interpretations for distributively unmarked sentences with plural NPs requires working memory resources. We found an effect of loading WM on interpretation preferences.

The-Distributive items were accepted 41% of the time in the No WM load group, a rate comparable to previous adult findings [de Koster et al. (2017) for Dutch; Pagliarini et al. (2012) for Italian]. Crucially, when WM was loaded, participants accepted items of condition The-Distributive at a significantly higher rate (80% for the low WM load condition and 78% for the high WM load condition). These general results are as predicted by the implicature account of distributivity preferences (Dotlačil, 2010). Loading adults' WM elicits a higher rate of acceptance for the The-Distributive condition, that is, the condition argued to involve an implicature.

In the rest of the discussion, we focus on three main issues in interpreting our results. First, we did not find a difference between high and low WM loads, only between the no load and load conditions. Does this matter? And what might explain this result? Second, we unexpectedly found an increase in acceptance for the Each-Collective items in the WM load conditions. What could explain this result? And is there evidence that distinguishes it from the increase in acceptance for the The-Distributive items? Finally, a major advantage of the implicature account compared to other explanations for collective and distributive preferences is that it offers an explanation for children's non-adult preferences, since children are known to be less likely to calculate implicatures. But can our results plausibly explain children's very late acquisition of adult preferences?

Different Load Conditions Effects

Contrary to our expectations, we did not find a difference between the low WM load condition and the high WM load condition. We only found a difference between the two WM load conditions on the one hand the no WM load condition on the other hand. Adults showed greater acceptance of distributive readings without distributive marking in both WM load conditions compared to the no WM load condition.

Similar results have actually been found in another study. van Tiel et al. (2019), who treated no WM load, low WM load, and high WM load, all as between subject conditions, also found a difference between the no load condition and the load conditions for *some-not all*. Note, however, that the other dual task studies that tested only the *some-not all* implicature (De Neys and Schaeken, 2007; Dieussaert et al., 2011; Marty and Chemla, 2013, Marty et al., 2013) did not include a no WM load condition, so we cannot be sure whether or not they might have found a difference between a no load and the low load condition.

There is also some evidence that different implicatures show different sensitivities to cognitive load. We know from previous studies that, even without a cognitive load, implicatures vary in rates of calculation (see, e.g., van Tiel et al., 2016; Sun, 2018). In our study, the low WM load condition was already sufficient to lead to more acceptance of The-Distributive items, and this may be because the proposed scale maybe less common or less automatized than the <some, all> scale. In fact, van Tiel et al. (2019) also found that different implicature types showed varying degrees of sensitivity (or lack of) to WM load. It could be that the less frequent or familiar a scale is, the more sensitive that a scale will be to working memory limitations.

Another factor that may have influenced the effect found was our choice of secondary task. Most other dual-task studies cited [except for Marty et al. (2013) that used backwards letter sequence retrieval, and Ryzhova and Demberg (2020) that used a dot-tracking task] used the dot memory task, where participants had to either recall a very simple dot-matrix pattern of three dots in a vertical or horizontal row, or a more complex pattern of four dots. While the four-dot matrix pattern has been shown to tap executive working memory (see e.g., Miyake et al., 2001) it is not clear to what degree the simpler three-dot matrix pattern actually requires WM resources. In comparison with this three-dot matrix pattern task, our three-digit memory task may load working memory more than the three-dot matrix patterns. Further investigations are needed to know to what degree these different secondary tasks load working memory.

There are, however, two issues we should discuss related to our No WM load group. The number of participants in our no-load group was relatively small (namely, 16, compared with 58 in the load group). While the response rates to the four different conditions tested in the No WM load group are similar to what we have found in previous experiments with adults, using the same visual materials (e.g., de Koster et al., 2018), it is well-known from the implicature literature that the rate of implicature can vary widely, that individuals also may differ in their tendency to calculate implicatures (see, e.g., Feeney and Bonnefon, 2013), and that this can be affected by, e.g., individual differences in working memory capacity (Dieussaert et al., 2011). Despite the large effect size, we have to be cautious about interpreting this comparison, because it could be that our No WM load group was made up of individuals who were particularly predisposed to calculate the implicature, which, in turn, could have made the comparison with the load group particularly significant.

The second issue is that, while the no-load group and load groups were between subjects, the two load groups were within subjects. A stronger comparison could be made if all conditions were run within the participants. However, there are also two problems with doing so. First, the experiment with two conditions was already quite long: Running our two load conditions within the participants already took ∼1 h, so running all three conditions as a between-participants study might introduce practice and fatigue effects. Second, exposure to so many items might also lead to unwanted influences and make comparisons between reaction times less valid. This is one of the reasons why van Tiel et al. (2019) ran all three load conditions between subjects. Future work should carefully consider these design issues.

Each-Collective Items More Acceptable Under WM Load

An unexpected result of our experiment was that, under WM load, adults increased their acceptance of Each-Collective items as well. The effect size is smaller than the increase in the acceptance rate of The-Distributive items (which doubled from 40% to around 80%), but it was still substantial and significant (from 32% to 57%).

In fact, given the literature on quantifiers, which suggests that distributive marking is semantically incompatible with most collective situations, the acceptance rate of 32% for Each-Collective items is actually unexpected. However, our experiments were run in Dutch, and Dutch *elke* has been shown to be closer in interpretation to English *every* than to English *each*. Several experimental studies on Dutch have shown that, while participants strongly prefer distributive meanings with Dutch *elke* in a preference task, they will accept a collective interpretation with *elke* in a picture verification task at relatively high rates (around 35%), contrasting sharply with results with English *each* (Rouweler and Hollebrandse, 2015; de Koster et al., 2017). If Dutch *elke* is better understood as being, in some cases, compatible with both distributive and collective interpretations but with a bias to a distributive interpretation, then one explanation for the effect of WM load may be that the ambiguity resolution process is affected by limited WM capacity, leading some participants to simply abandon disambiguation and simply accept all presented situations with *elke*. Thus, this finding could be similar to the finding of van Rij et al., 2013 that the resolution of ambiguous pronouns is affected by limited WM capacity due to the listener's decreased ability to integrate contextual information needed for the disambiguation.

Note, however, that this explanation is not simply a proposal that a WM load leads to greater acceptance across the board. Instead, the idea is that ambiguity resolution, specifically, may be more sensitive to WM capacity. Recall that participants were not more likely to accept interpretations in general under a WM load, and, for the false task control items, there was no difference in acceptance between the No WM load group and the WM load groups.

Additionally, the RTs for conditions The-Distributive and Each-Collective point to different underlying interpretation processes. Recall that we found out that "no" (pragmatic) responses for The-Distributive items were significantly slower than "yes" (literal) responses. Similar findings have been found in several studies of *some-not all* implicatures, including Bott and Noveck (2004), who found that rejecting upper-bound readings with *some* took longer than the literal, semantic interpretation. If the rejection of distributive readings with unmarked sentences is due to an implicature, it would be consistent with these other results, suggesting implicature calculation also takes longer. However, we did not find any difference in RTs between "yes" and "no" responses for the Each-Collective condition, suggesting that the increased acceptance rate under a WM load is due to a different underlying process more than the increased acceptance found with The-Distributive.

Children's Non-adult Preferences and the Role of WM

The implicature account of distributive preferences argues that children fail to interpret distributively unmarked sentences as collective because they fail to calculate the implicature. Young children's lower working memory capacity is often used to explain their failure to compute implicatures, so does finding a role for working memory in distributivity preferences offer an explanation for children's late acquisition?

Studies of implicature acquisition for different scales have often found gaps of several years between when children have the lexical knowledge required for an implicature and when they actually compute the implicature. Even for the well-studied *some-not all* implicature, acquisition results seem to suggest that children at age 4 already possess the lexical knowledge necessary for implicature calculation, but many studies find that they do not calculate implicatures consistently until around age seven (e.g., Noveck, 2001; Pouscoulous et al., 2007; Foppolo et al., 2012). Furthermore, previous findings indicate that different implicatures are acquired at different ages. Noveck (2001), for example, examined implicatures based on the <might, must> scale in which the modal *might* implies that the stronger *must* does not hold. He found out that 7-year-olds were the youngest

to demonstrate modal competence overall, but that 7- and 9-year-olds still interpreted *might* logically (not implicating "not must") more often than adults did. This shows that 9-year-old children did not fully master the *might-not must* implicature, yet, though presumably, they already had the lexical knowledge and working memory capacity at age 9 to calculate the *some-not all* implicature. For distributivity, children are non-adult-like in their distributivity preferences until comparatively late. Recall that Dutch children at age 11 were still not adult-like (de Koster et al., 2018) and Italian children at age 14 were also not adult-like (Pagliarini et al., 2012). However, experimental research has shown that, by age nine, children know that lexically distributive markers signal distributivity. At this age, we would expect that children have sufficient working memory capacity for the implicature calculations needed. With such a large gap between acquisition of the lexical scale and adult-like performance in the calculation of the proposed implicature, the role of working memory in the acquisition process is unclear.

One possibility is that working memory capacity is not really the bottleneck in children's late development of collective interpretation preferences. The difficulty is not in the decision to calculate the implicature, or the comparison of alternatives (which may require memory resources) but in recognizing collective and distributive interpretations as comparable alternatives on an informativity scale. Children may need much more verbal experience than what they have at age 9 (or age 11–14) and need to encounter many more examples where the distinction is relevant before they will begin to interpret the two meanings and their potential marking as alternatives on a scale. While many expressions satisfy the requirements to create a scale, only in a context in which the contrast is relevant do implicatures arise. Thus, <car, Honda civic> is a scale, but if a speaker said that a car almost hit them on their morning bike ride, this is unlikely to give rise to the implicature that the car was not a Honda civic, because, in that context, the specific make of the car is irrelevant [see Matsumoto (1995) and Geurts (2010) for more discussion of the contextual constraints on implicatures]. But this also means that, in addition to the recognition of the scale, experience with a weaker term being used in contexts where the contrast with the stronger is relevant is also important, and, for some scales, this might not be all that frequent. Even *some-not all* implicatures, which many researchers believe to be so frequent as to be (almost) a default interpretation, have been found in corpus studies to be much less frequent than previous believed [e.g., see Degen (2015) and Eiteljoerge et al. (2018), who found that only about 15% of uses of *some* in child-directed speech were intended with an implicature meaning]. An additional difficulty could be that, unlike many other scales, the expressions *the* and *each* require different inflectional morphology (e.g., *each* requires a singular verb, and plural definite descriptions require plural verbal morphology in English and in Dutch) and cannot simply be substituted for each other. Even though it is known that substitutability is not a requirement for scalar expressions [e.g., because it does not work in many contexts, e.g., downward entailing environments, see Geurts (2010) for a discussion], it still may influence how easy it is to acquire the scale and associated implicatures. If frequency and experience, indeed, explain children's late acquisition of adult-like preferences, then the lower rate of implicature in adults in our study and children in other studies has different origins: Adults under a working memory load do not calculate the implicature because it requires too many resources. Children do not calculate the implicature because they do not have sufficient experience with the competing alternatives until quite late. The gradual acquisition that we see in children from age nine onwards could then be reflective of a gradual increase in an experience that translates into greater awareness of the scale and thus a greater tendency to recognize the contrast and calculate the implicature.

The alternative explanation is that working memory capacity still does play a role in children's development, and that greater WM capacities, in combination with greater experience, only comes together quite late (e.g., 14+). The advantage of this proposal is that it offers an explanation for the correlation found between working memory and the rate of implicature calculation in de Koster et al. (2017), a relationship that would otherwise be hard to explain if working memory does not play a role at all in children's interpretation processes. More research can, perhaps, help distinguish between these two possible explanations.

CONCLUSIONS AND DIRECTIONS FOR FUTURE RESEARCH

Summarizing, we found that loading adults' WM leads them to accept distributive interpretations without distributive marking, a result that is predicted by the theoretical proposal that adult collective preferences for distributively unmarked sentences originate from a pragmatic implicature. This study thus makes a novel contribution to our understanding of the semantic and pragmatic processes underlying distributivity and their interaction with cognitive resources such as working memory.

Many open questions remain. First, in general, we need more studies looking at other proposed quantity implicatures. Most research has focused on the <some, all> scale and, to a certain degree, the <or, and> scale, but, within the research that has examined the processing of other proposed scales, it does seem that implicatures differ widely in their tendency to be calculated, and their tendency to be sensitive to processing limitations. But we need to confirm this variation experimentally. Second, for distributivity preferences, in particular, it still remains unclear what role working memory capacity plays in children's non-adult interpretation preferences. While working memory capacity was shown to correlate with the rejection of distributive readings without distributive marking in children (de Koster et al., 2018), the very late age at which children begin to be adult-like in their interpretation preferences suggests that other factors, such as experience with the scale or with distributive and collective situations, might play an even bigger role. Investigating this further would help clarify children's development. Another issue is the question of where working memory resources actually come into play in interpretation. Studies such as Marty and Chemla (2013) have found some evidence suggesting that the decision to calculate an implicature may be what requires cognitive resources,

but more studies are needed. Future research should develop experiments to try to pinpoint where in the interpretation process resources are required. With more experimental investigations, we can hopefully develop a fuller picture of distributive interpretation preferences.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The study in this article, which involves human participants, was reviewed and approved by the Research Ethics Committee (CETO) of the University of Groningen. The participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

AK, JS, and PH conceived the experiment, analyzed the results, read, commented, and revised the manuscript. AK carried out the experiment and wrote the first draft of the manuscript. All the authors approved the final version of the manuscript.

FUNDING

This research was supported by a grant from the Netherlands Organization for Scientific Research (NWO) (Grant No. 322-75-010).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.556120/full#supplementary-material

REFERENCES

Akaike, H. (2011). "Akaike's information criterion," in *International Encyclopedia of Statistical Science*, ed M. Lovric (Berlin; Heidelberg: Springer). doi: 10.1007/978-3-642-04898-2_110

Avrutin, S., and Thornton, R. (1994). Distributivity and binding in child grammar. *Lingu. Inquiry*. 25, 165–171.

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

Banga, A., Heutinck, I., Berends, S. M., and Hendriks, P. (2009). "Some implicatures reveal semantic differences," in *Linguistics in the Netherlands*, eds B. Botma and J. van Kampen (Amsterdam: John Benjamins), 1–13. doi: 10.1075/avt.26.02ban

Bott, L., Bailey, T. M., and Grodner, D. (2012). Distinguishing speed from accuracy in scalar implicatures. *J. Mem. Lang*. 66, 123–142. doi: 10.1016/j.jml.2011.09.005

Bott, L., and Noveck, I. A. (2004). Some utterances are underinformative: the onset and time course of scalar inferences. *J. Mem. Lang*. 51, 437–457. doi: 10.1016/j.jml.2004.05.006

Breheny, R., Katsos, N., and Williams, J. (2006). Are generalised scalar implicatures generated by default? An on-line investigation into the role of context in generating pragmatic inferences. *Cognition* 100, 434–463. doi: 10.1016/j.cognition.2005.07.003

Broekhuis, H., and den Dikken, M. (2012). *Syntax of Dutch: Nouns and Noun Phrases. Vol. 2*. Amsterdam: Amsterdam University Press. doi: 10.1017/9789048517602

Brooks, P. J., and Braine, M. D. S. (1996). What do children know about the universal quantifiers all and each? *Cognition* 60, 235–268. doi: 10.1016/0010-0277(96)00712-3

Burnham, K. P., and Anderson, D. R. (2002). *A Practical Information-Theoretic Approach. Model Selection and Multimodel Inference, 2nd ed*. New York: Springer.

Chambers, J. M., Cleveland, W. S., Kleiner, B., and Tukey, P. A. (1983). *Graphical Methods for Data Analysis*. Pacific Grove, CA: Wadsworth & Brooks/Cole, 158–162.

Champollion, L. (2017). *Parts of a Whole: Distributivity as a Bridge Between Aspect and Measurement. Vol. 66*. Oxford: Oxford University Press. doi: 10.1093/oso/9780198755128.003.0008

Chemla, E., and Bott, L. (2013). Processing presuppositions: Dynamic semantics vs pragmatic enrichment. *Lang. Cogn. Processes*. 28, 241–260. doi: 10.1080/01690965.2011.615221

Chemla, E., and Singh, R. (2014). Remarks on the experimental turn in the study of scalar implicature, Part 1. *Langu. Lingu. Compass* 8/9, 373–386. doi: 10.1111/lnc3.12081

Chierchia, G. (2004). Scalar implicatures, polarity phenomena, and the syntax/pragmatics interface. *Struct. Beyond* 3, 39–103.

Chierchia, G. (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

Cremers, A., and Chemla, E. (2014). "Direct and indirect scalar implicatures share the same processing signature," in *Pragmatics, Semantics and the Case of Scalar Implicatures,* ed S. Pistoia Reda (London: Palgrave Macmillan), 201–227. doi: 10.1057/9781137333285_8

de Hoop, H., and Kas,. M. (1989). Sommige betekenisaspecten van enkele kwantoren, oftewel: enkele betekenisaspecten van sommige kwantoren. *Interdisc. Taal &Tekstwetenschap*. 9, 31–49.

de Koster, A. M. B., Dotlačil, J., and Spenader, J. K. (2017). "Children's understanding of distributivity and adjectives of comparison," in *Proceedings of the 41st Annual Boston University Conference on Language Development*, eds M. LaMendola and J. Scott (Cascadilla Press), 373–386.

de Koster, A. M. B., Spenader, J. K., Dotlačil, J., and Hendriks, P. (2020). "A multiple cue analysis of collective interpretations with 'Each,'" in *Proceedings of the 44th Annual Boston University Conference on Language Development*, eds M. M. Brown and A. Kohut (Cascadilla Press), 252–265.

de Koster, A. M. B., Spenader, J. K., and Hendriks, P. (2018). "Are Children's Overly Distributive Interpretations and Spreading Errors Related?" in *Proceedings of the 42nd Annual Boston University Conference on Language Development*, eds A. B. Bertolini and M. J. Kaplan (Cascadilla Press), 413–426.

De Neys, W., and Schaeken, W. (2007). When people are more logical under cognitive load: Dual task impact on scalar implicature. *Exp. Psych*. 54, 128–133. doi: 10.1027/1618-3169.54.2.128

Degen, J. (2015). Investigating the distribution of some (but not all) implicatures using corpora and web-based methods. *Sem. Pragm*. 8, 11–11. doi: 10.3765/sp.8.11

Degen, J., and Tanenhaus, M. K. (2016). Availability of alternatives and the processing of scalar implicatures: a visual world eye-tracking study. *Cogn. Sci*. 40, 172–201. doi: 10.1111/cogs.12227

Dieussaert, K., Verkerk, S., Gillard, E., and Schaeken, W. (2011). Some effort for some: further evidence that scalar implicatures are effortful. *Q. J. Exp. Psych*. 64, 2352–2367. doi: 10.1080/17470218.2011.588799

Dotlačil, J. (2010). *Anaphora and distributivity: a study of same*, different, reciprocals and others (Doctoral dissertation). Utrecht University, Netherlands.

Dotlačil, J., and Brasoveanu, A. (2021). The representation and processing of distributivity and collectivity: ambiguity vs. underspecification. *Glossa J. Gen. Linguist.* 6:6. doi: 10.5334/gjgl.1131

Eiteljoerge, S. F., Pouscoulous, N., and Lieven, E. V. (2018). Some pieces are missing: implicature production in children. *Front. Psych.* 9:1928. doi: 10.3389/fpsyg.2018.01928

Feeney, A., and Bonnefon, J. F. (2013). Politeness and honesty contribute additively to the interpretation of scalar expressions. *J. Lang. Soc. Psychol.* 32, 181–190.

Foppolo, F., Guasti, M. T., and Chierchia, G. (2012). Scalar implicatures in child language: Give children a chance. *Language learning and development*. 8, 365–394. doi: 10.1080/15475441.2011.626386

Frazier, L., Pacht, J. M., and Rayner, K. (1999). Taking on semantic commitments, II: collective versus distributive readings. *Cognition* 70, 87–104. doi: 10.1016/S0010-0277(99)00002-5

Geurts, B. (2010). *Quantity Implicatures*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511975158

Gil, D. (1982). Quantifier scope, linguistic variation, and natural language semantics. *Lingu. Philos.* 5, 421–472. doi: 10.1007/BF00355582

Grice, H. P. (1975). "Logic and conversation," in *Syntax and Semantics,* Vol. 3: speech acts, eds P. Cole and J. J. Morgan (New York, NY: Academic Press), 41–58.

Grodner, D. J., Klein, N. M., and Carbary, K. M., and Tanenhaus, M. K. (2010). "some," and possibly all, scalar inferencesare not delayed: evidence for immediate pragmatic enrichment. *Cognition* 116, 42–55. doi: 10.1016/j.cognition.2010.03.014

Hartshorne, J. K., and Snedeker, J. (2014). The speed of inference: evidence against rapid use of context in calculation of scalar implicatures (Manuscript submitted for publication).

Horn, L. R. (1973). *On the semantic properties of logical operators in English* (Doctoral dissertation*)*. University of California, United States.

Horn, L. R. (1989). *A Natural History of Negation*. Chicago, IL: University of Chicago Press.

Huang, Y. T., and Snedeker, J. (2009). Online interpretation of scalar quantifiers: Insight into the semantics–pragmatics interface. *Cogn. Psych.* 58, 376–415. doi: 10.1016/j.cogpsych.2008.09.001

Huang, Y. T., and Snedeker, J. (2011). Cascading activation across levels of representation in children's lexical processing. *J. Child Lang.* 38, 644–661.

Kaup, B., Kelter, S., and Habel, C. (2002). Representing referents of plural expressions and resolving plural anaphors. *Lang. Cogn. Processes*. 17, 405–450. doi: 10.1080/01690960143000272

Ko, V., Wieling, M., Wit, E., Nerbonne, J., and Krijnen, W. (2014). Social, geographical, and lexical influences on Dutch dialect pronunciations. *Comput. Ling. Netherl. J.* 4, 29–38.

Kursat, L., and Degen, J. (2020). "Probability and processing speed of scalar inferences is context-dependent," in *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*, 1236–1242.

Landman, F. (2000). *Events and Plurality: The Jerusalem Lectures. Vol. 76.* Boston, MA: Kluwer Academic Pub. 10, 978–994. doi: 10.1007/978-94-011-4359-2

Maldonado, M., Chemla, E., and Spector, B. (2019). Revealing abstract semantic mechanisms through priming: the distributive/collective contrast. *Cognition* 182, 171–176. doi: 10.1016/j.cognition.2018.09.009

Marty, P., Chemla, E., and Spector, B. (2013). Interpreting numerals and scalar items under memory load. *Lingua*, 133: 152–163. doi: 10.1016/j.lingua.2013.03.006

Marty, P. P., and Chemla, E. (2013). Scalar implicatures: working memory and a comparison with only. *Front. Psych.* 4:403. doi: 10.3389/fpsyg.2013.00403

Matsumoto, Y. (1995). The conversational condition on horn scales. *Linguis. Philos.* 18, 21–60. doi: 10.1007/BF00984960

Miyake, A., Friedman, N. P., Rettinger, D. A., Shah, P., and Hegarty, M. (2001). How are visuospatial working memory, executive functioning, and spatial abilities related? A latent-variable analysis. *J. Exp. Psychol. Gen.* 130, 621–640. doi: 10.1037/0096-3445.130.4.621

Miyamoto, Y., and Crain, S. (1991). Children's interpretation of plural pronouns: collective vs. distributive interpretation," in *Proceedings of the Annual Boston University Conference on Language Development* (Cascadilla Press).

Noveck, I. A. (2001). When children are more logical than adults: Experimental investigations of scalar implicature. *Cognition*. 78, 165–188. doi: 10.1016/S0010-0277(00)00114-1

Padilla-Reyes, R. E. (2018). *Connections among scales, plurality, and intensionality in Spanish*. (Doctoral dissertation). The Ohio State University, Ohio.

Pagliarini, E., Fiorin, G., and Dotlačil, J. (2012). "The acquisition of distributivity in pluralities, in *Proceedings of the 36st Annual Boston University Conference on Language Development* (Cascadilla Press), 2, 387–399.

Politzer-Ahles, S., and Fiorentino, R. (2013). The realization of scalar inferences: context sensitivity without processing cost. *PLoS ONE* 8:e63943. doi: 10.1371/journal.pone.0063943

Politzer-Ahles, S., and Husband, E. M. (2018). Eye movement evidence for context-sensitive derivation of scalar inferences. *Collabra Psychol.* 4, 1–13. doi: 10.1525/collabra.100

Pouscoulous, N., Noveck, I. A., Politzer, G., and Bastide, A. (2007). A developmental investigation of processing costs in implicature production. *Lang. Acqu*. 14, 347–375. doi: 10.1080/10489220701600457

Romoli, J., and Schwarz, F. (2015). "An experimental comparison between presuppositions and indirect scalar implicatures," in *Experimental Perspectives on Presuppositions*, ed F. Schwarz (Cham: Springer), 215–240. doi: 10.1007/978-3-319-07980-6_10

Rouweler, L., and Hollebrandse, B. (2015). Distributive, collective and "everything" in between: Interpretation of universal quantifiers in child and adult language. *Lingu. Netherlands*. 32, 130–141. doi: 10.1075/avt.32.10rou

Ryzhova, M., and Demberg, V. (2020). "Processing particularized pragmatic inferences under load," in *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society (CogSci 2020).*

Spenader, J., and Bosnic, A. (2018). "Distributivity preferences for Dutch quantifiers elk and ieder." in *Chapter in Hollebrandse,* Vol. 41, eds B. Kim, J., Perez-Leroux, A. T., and Schulz, P. *T.O.M. and Grammar: Thoughts on Mind and Grammar: A Festschrift in honor of Tom Roeper* (Amherst, MA: Graduate Linguistics Student Association), 159–168.

Sun, C. (2018). *Scalar implicature: gricean reasoning and local enrichment*. (Doctoral dissertation). UCL (University College London)), London.

Syrett, K., and Musolino, J. (2013). Collectivity, distributivity, and the interpretation of plural numerical expressions in child and adult language. *Lang. Acqu*. 20, 259–291. doi: 10.1080/10489223.2013.828060

Tomlinson, J., Bott, L., and Bailey, T. (2011). "Understanding literal meanings before pragmatic inference: Mouse-trajectories of scalar implicatures," in *4th Biennial Conference of Experimental Pragmatics,* 2–4.

Tunstall, S. (1998). *The interpretation of quantifiers: Semantics and processing* (Doctoral dissertation). University of Massachusetts, Amherst.

van der Ziel, M. E. (2012). *The acquisition of scope interpretation in dative constructions: explaining children's non-target like performance* (Doctoral dissertation). Utrecht University, Netherland.

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. *Topics Cogn. Sci*. 5, 564–580. doi: 10.1111/tops.12029

van Tiel, B., Pankratz, E., and Sun, C. (2019). Scales and scalarity: Processing scalar inferences. *J. Mem. Lang.* 105, 93–107. doi: 10.1016/j.jml.2018.12.002

van Tiel, B., van Miltenburg, E., Zevakhina, N., and Geurts, B. (2016). Scalar diversity. *J. Sem*. 33, 137–175.

Wieling, M., Montemagni, S., Nerbonne, J., and Baayen, R. H. (2014). Lexical differences between Tuscan dialects and standard Italian: accounting for geographic and sociodemographic variation using generalized additive mixed modeling. *Language* 90, 669–692. doi: 10.1353/lan.2014.0064

Wieling, M., Nerbonne, J., and Baayen, R. H. (2011). Quantitative social dialectology: explaining linguistic variation geographically and socially. *PLoS ONE* 6:e23613. doi: 10.1371/journal.pone.0023613

Conflict of Interest: 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 © 2021 de Koster, Hendriks and Spenader. 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) and the copyright owner(s) 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 June 2021
doi: 10.3389/fpsyg.2021.628460

The Double Framing Effect of Emotive Metaphors in Argumentation

Francesca Ervas[1]*, *Maria Grazia Rossi*[2], *Amitash Ojha*[3] and *Bipin Indurkhya*[4]

[1] Department of Education, Psychology, Philosophy, University of Cagliari, Cagliari, Italy, [2] Instituto de Filosofia da Nova (IFILNOVA), Facultade de Ciências Sociais e Humanas, Universidade Nova de Lisboa, Lisbon, Portugal, [3] Department of Humanities and Social Sciences, Indian Institute of Technology Jammu, Jammu, India, [4] Department of Cognitive Science, Jagiellonian University, Kraków, Poland

In argumentation, metaphors are often considered as ambiguous or deceptive uses of language leading to fallacies of reasoning. However, they can also provide useful insights into creative argumentation, leading to genuinely new knowledge. Metaphors entail a framing effect that implicitly provides a specific perspective to interpret the world, guiding reasoning and evaluation of arguments. In the same vein, emotions could be in sharp contrast with proper reasoning, but they can also be cognitive processes of affective framing, influencing our reasoning and behavior in different meaningful ways. Thus, a double (metaphorical and affective) framing effect might influence argumentation in the case of emotive metaphors, such as "Poverty is a disease" or "Your boss is a dictator," where specific "emotive words" (disease, dictator) are used as vehicles. We present and discuss the results of two experimental studies designed to explore the role of emotive metaphors in argumentation. The studies investigated whether and to what extent the detection of a fallacious argument is influenced by the presence of a conventional vs. novel emotive metaphor. Participants evaluated a series of verbal arguments containing either "non-emotive" or "emotive" (positive or negative) metaphors as middle terms that "bridge" the premises of the argument. The results show that the *affective coherence* of the metaphor's vehicle and topic plays a crucial role in participants' reasoning style, leading to global heuristic vs. local analytical interpretive processes in the interplay of the metaphorical and the affective framing effects.

Keywords: metaphor, emotions, framing, equivocation fallacy, affective coherence, reasoning, belief in the conclusion, meaning ambiguity

OPEN ACCESS

Edited by:
Valentina Cuccio,
University of Messina, Italy

Reviewed by:
Francesca M. M. Citron,
Lancaster University, United Kingdom
Christelle Declercq,
Université de Reims
Champagne-Ardenne, France

***Correspondence:**
Francesca Ervas
ervas@unica.it

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

Received: *12 November 2020*
Accepted: *05 May 2021*
Published: *14 June 2021*

Citation:
Ervas F, Rossi MG, Ojha A and Indurkhya B (2021) The Double Framing Effect of Emotive Metaphors in Argumentation. Front. Psychol. 12:628460. doi: 10.3389/fpsyg.2021.628460

INTRODUCTION

Previous research in argumentation theory showed that reasoning errors, far from just leading to argumentation fallacies, might shed light on how we reason and what influences our evaluation of arguments (Tversky and Kahneman, 1981; Van Eemeren, 1992; Ariely, 2009; Walton, 2010). Reasoning errors have a psychological dimension (Macagno and Walton, 2010; Walton, 2010; Godden, 2015), as they are arguments that seem to be sound without being so in terms of norms and standards (Van Eemeren and Grootendorst, 2004; Tindale, 2006; Walton, 2006). Their psychological dimension is tightly connected to their linguistic dimension, as also the linguistic formulation of arguments may lead to fallacies of reasoning (Oswald et al., 2018, 2020; Hinton, 2019; Schumann et al., 2020). Fallacies of reasoning might reveal how we make sense of arguments, especially when they are formulated in natural languages, where ambiguous, polysemous, and non-literal use of words is widespread (Ervas et al., 2018). Far from being patently irrational

when committing a fallacy, people might just be drawn by the intuitive search for alternative reasons, as in the case of metaphors used in argumentation (Ervas et al., 2018). This argumentation style fits more with the "natural logic" invoked in a pragmatic perspective, where the speaker's meaning is grasped as the conclusion of an inference that makes sense of an apparent meaning violation (Grice, 1989).

Metaphors rarely come "alone," lacking the affective dimension, as they often communicate an emotional meaning (Fainsilber and Ortony, 1987; Schnall, 2005), whether positive as in "My partner is a rose" or negative as in "My job is a jail," which also contributes to the evaluation of arguments (Macagno and Rossi, 2021). The emotive language used in argumentation might have crucial implications when accepting the conclusion of an argument (Macagno and Walton, 2014; Pollaroli et al., 2019). Here, the term "argumentation" is used in a wider sense, covering all the reasons in support of the conclusions that a speaker wishes the addressee to draw from some premises. Ordinary evaluations might use other sources of reasoning, departing from normative standards and/or independent from the argument itself, i.e., actual premises and their connection to the conclusion, as already shown in the case of arguments featuring metaphors (Ervas et al., 2015, 2018). The paper aims to explore how emotions and figurative language interact in the evaluation of arguments where "emotive metaphors" connect the premises to a conclusion.

In the Western philosophical tradition, the notion of argumentation has often denied both emotions and figurative language the status of reasoning resources. Argumentation has been defined as a critical use of reason in judgment, often in contrast to emotions (Oaksford et al., 1996; Blanchette et al., 2018): rational justification seems to be the unique relevant source of knowledge at a normative level, while emotions are subjective feelings conveying only perspectives. Emotions safeguard compelling and fleeting interests, and can be "recalcitrant" to reason and arguments (Greenspan, 1988, 1992, 2004; Stocker and Hegeman, 1996; DeLancey, 2002, 2007; D'Arms and Jacobson, 2003). Emotions are often processed in an automatic, unconscious, and obliged way, while the critical use of reason is supposed to be conscious and controlled. Previous experimental studies have challenged this view (Blanchette and Caparos, 2013; Blanchette, 2014), and proposed that conscious thought does not always lead to a better performance than unconscious thought when complex decisions have to be made [but see Rossi (2013) and Rossi (2014)]. Due to a wider capacity to deal with multiple information, Dijksterhuis (2004, p. 593) claimed that the "unconscious also actively thinks," by associating and integrating the various alternatives in memory [see also Dijksterhuis et al. (2006) and Dijksterhuis et al. (2009)]. Other studies have investigated how positive and negative emotions differently influence both content and style of thought, playing an important role in the regulation of the global-local information processing (Fredrickson and Branigan, 2005; Clore and Huntsinger, 2007).

Nonetheless, figurative language has often been considered so rich in suggestion as to be dangerous in argumentation (Beardsley, 1957). Metaphors have been described as the "exemplars of the improper" (Maasen, 2000, p. 199), leading to fallacies of reasoning in argumentation. They have often been counted as semantic anomalies, deviations from the language properly used in argumentation (Hoffman, 1980; Tourangeau and Sternberg, 1982), or as ornamental devices inessential to argumentation. That metaphor is just a deviant or an ornamental use of language has been largely questioned both in philosophy by Max Black (1954, 1962) and in cognitive linguistics by Lakoff and Johnson (1980). In understanding an abstract concept (the *target*) in terms of a concrete concept (the *source*), metaphor implicitly provides a frame to think of and to reason about the target, selecting some relevant properties of the source and neglecting others (Entman, 1993; Burgers et al., 2016). Still, in the conceptual metaphor theory (Lakoff and Johnson, 1980), studies focusing on metaphorical reasoning [see Thibodeau et al. (2019) for a review] widely acknowledge that the metaphorical framing effect can often work covertly and affect reasoning and evaluation of arguments [Thibodeau and Boroditsky, 2011; but see Steen et al. (2014) for criticism].

From this perspective, metaphor plays an alternative role with respect to reasoning, conceived as the critical and deliberate use of rationality. The framing effect is interpreted as a cognitive bias, influencing how people make decisions or express their evaluations based on how an issue or an argument is presented, rather than following proper logical or normative rules. Something similar can be said about affective framing: that it is another strategy to exploit emotions to frame information and manipulate both reasoning and decision-making. Emotions can be used to present the linguistic formulation of an argument with a specific (positive or negative) valence, which can be considered as a special "semantic primitive" determining the intended (positive or negative) meaning (Barrett et al., 2007). Maiese (2014, p. 524) proposed the term "affective framing" to express the idea that emotions are "a spontaneous, non-inferential, and pre-reflective way of discriminating, filtering, and selecting information that allows us to reduce the overwhelming clutter of information" [see also DeLancey (2002), Solomon (2003), and Prinz (2004)]. From this perspective, emotions can strongly influence the evaluation of arguments (e.g., Schwarz and Clore, 1983; Bless et al., 1996; Schwarz and Skurnik, 2003). Scholars have argued for a hidden and overwhelming force of emotions overcoming normative rules in various types of social reasoning (Marcus et al., 2000; Marcus, 2002; MacKuen et al., 2010; Angie et al., 2011). In moral reasoning, scenarios based on strong emotional reactions have been used to insist on the "moral doumbfounding" effect of emotions: for instance, experiencing a strong disgust reaction after reading a story of "consensual incest" brings many participants to remain stubbornly committed to a judgment of moral inappropriateness despite the fact that they are unable to propose adequate arguments to justify their initial emotional intuition [see also Haidt et al. (2000), Haidt (2001, p. 814), and Haidt (2007)].

Previous research has highlighted the evaluative connotations entailed by the framing effect present in metaphors, such as "Poverty is a disease" or "Your boss is a dictator," where specific "emotive words" (disease/dictator) are intentionally used (Stevenson, 1944; Macagno and Walton, 2014). These examples

illustrate how metaphors might guide our reasoning, not only by framing arguments with vividness and forceful figurative images, but also by entailing the communication of emotional attitudes and value judgments (Semino, 2008; Burgers et al., 2016). In this paper, we investigate the double framing effect of emotive metaphors in arguments, and check whether and to what extent the presence of an "emotive metaphor" influences the detection of fallacies.

The Metaphorical Framing Effect on Argumentation

Scholars have shown that metaphors can be useful in argumentation to introduce a standpoint or to underpin it (Wagemans, 2016, 2019; van Poppel, 2018, 2020), while in the context of science, metaphors have been used to stimulate creative thinking (Blackburn, 1984; Hofstadter, 1995; Indurkhya, 2010). Recent studies have reconsidered traditional approaches to metaphor as a reasoning device (Black, 1962; Hesse, 1963, 1965; Perelman and Olbrechts-Tyteca, 1969; Indurkhya, 2007), claiming that metaphor itself might be considered as an "implicit argument" where the addressee is led along a chain of inferences from the source to the target to draw some conclusion (Santibáñez, 2010; Macagno and Zavatta, 2014; Oswald and Rihs, 2014; Svačinova, 2014). Other studies (Ervas et al., 2018; Ervas, 2019; Cavazzana and Bolognesi, 2020) claimed that metaphors, as implicit arguments, can be considered as *enthymemes*, having a syllogistic form of reasoning with implicit premises. Specifically, the syllogism would have the metaphor as the first premise and the relevant property or properties to attribute to the target as a second premise. Here, the role of the middle term connecting the premises to the conclusion is played by the *vehicle*, the linguistic term that refers to the source concept of the metaphor, which also provides a frame for interpreting the target concept.

When considering the metaphorical framing effect in reasoning, and its influence on the evaluation of arguments, most empirical studies focused on conventional metaphors (Thibodeau and Boroditsky, 2011, 2015), which are quite frequent and already lexicalized for a given linguistic community. On the contrary, novel metaphors might well be used as relevant moves in argumentation (van Poppel, 2018, 2020; Macagno, 2020), which intentionally frame a discourse to interfere with, or even lead, the reasoning process. When argumentation intentionally exploits a metaphor, the robust notion of truth needs to be dropped: sentences featuring a metaphor are literally "patently" false, because of their conventional meaning. Still, from a pragmatic perspective, we find "an alternative truth" in the Gricean natural logic, interpreting the speaker's meaning and making sense of the sentence according to the context (Grice, 1975, 1989; Clark, 1994; Wilson and Sperber, 2002). The Gricean cooperative principle assumes that, beyond being informative, sentences are true. However, sentences featuring conventional and novel metaphors are processed in very different ways when it comes to their truth evaluations. Empirical research (Glucksberg, 2001, 2003; Giora, 2003) has shown that most sentences with conventional metaphors are perceived as true though they are literally false, while most sentences with novel metaphors are processed as false. Response times also suggest a metaphorical interference effect in the truth evaluation of a sentence (Glucksberg et al., 1982). In particular, novel metaphors with unfamiliar meanings cannot be inhibited or ignored, which explains why it takes less time, compared to conventional metaphors, to judge whether sentences featuring novel metaphors are false. Consequently, the process of truth evaluation of the sentence in which the metaphor occurs also influences the evaluation of the whole argument having the metaphorical sentence as a premise.

For conventional metaphors, the relevant properties attributed to the target come from a "system of associated commonplaces" (Black, 1954), which are usually assumed to be true or just believed as belonging to the source concept. The relevant properties are often stereotypically believed to belong to the source concept (Ervas, 2017; Borelli and Cacciari, 2019), possibly leading to fallacies of reasoning (Fischer, 2014, 2015). When they are applied to the target, the preservation of truth in the conclusion is never guaranteed. The speaker's beliefs about the source concept can thus influence the conclusion of the argument featuring a metaphor. People might accept the conclusion of an argument just because they believe in it, rather than because it logically follows from the premises (Ball et al., 2006; Correia, 2011; Ball and Thompson, 2018), thereby influencing the evaluation of the whole argument (Ervas et al., 2018).

The Affective Framing Effect on Argumentation

As for metaphors, having an embodied (re)framing effect does not automatically mean that emotions can be viewed just as a covert and subconscious force that makes reasoning derails into fallacies. Emotions can have both a bodily and a cognitive-evaluative dimension, as they are the means by which personally relevant environmental information is made available and meaningful to the experiences of a subject (Maiese, 2015). Therefore, precisely for their framing effect, emotions can be considered as an important source of knowledge (Damasio, 1994). As cognitive processes used to represent the positive and negative valence of objects, people and/or actions in the world, they might play a crucial role in reasoning because of their strong evaluative dimension (Caruana and Cuccio, 2017). Although emotions are not explicitly intentional (Damasio, 1994; LeDoux, 1996), they can guide our behavior and be useful predictors of actions: negatively valenced stimuli represent potential threats demanding an immediate response (Rozin and Royzman, 2001; Citron et al., 2014). Previous research showed that positively valenced contexts reduces or even eliminates possible framing effects in decision-making (Cassotti et al., 2012).

Some studies have shown that when emotions are conveyed by verbal stimuli, they strongly affect reading times (Kissler et al., 2006; Citron, 2012). Specifically, emotionally-valenced terms are processed faster and more accurately than neutral terms (Larsen et al., 2006; Kousta et al., 2009). Behavioral ratings of emotionally-valenced stimuli show that both highly positively valenced and negatively valenced stimuli are more

arousing than neutral stimuli (Bradley and Lang, 1999) and negatively valenced stimuli are usually rated as more arousing than positively valenced stimuli (Citron et al., 2013). When emotions are more difficult to translate into plain language, we might resort to non-literal and/or figurative language, where meanings can afford one with the liberty to implicitly convey emotions without being overtly committed to the literal value of the words (Gibbs et al., 2002; Ervas, 2020). Kövecses (2000, 2005), who dedicated his work to the different perspectives entailed by metaphors expressing emotions, refers to this use of language as the figurative descriptive function of emotive terms.

Previous research investigated how frames affect individuals' beliefs on a variety of issues, focusing on the place of emotions in the framing process (Petty and Cacioppo, 1986a; Scherer et al., 2014). Being spontaneous action tendencies, emotions can have a direct behavioral effect, but can also indirectly affect judgment, by generating other emotions, or by changing the preference ordering in motivations (Elster, 1999). Previous studies also focused on different cognitive processes explaining how framing effects operate, such as *accessibility change*, making information more salient, *belief importance change*, altering the weight of information, or *belief content change*, adding new information potentially changing the conclusion (Slothuus, 2008), but also affective factors might have a role in mediating framing effects (Gross and D'Ambrosio, 2004; Gross, 2008). The valence of the terms in a message modulates the subjective state of feeling pleasure or displeasure in response to the message (Lecheler et al., 2013). Frames themselves often rely upon emotional appeals, represented by "emotion-laden" terms, which could be properly designed to elicit positive vs. negative emotional reactions. This is also the case of emotive metaphors, whose framing effect does not merely depend on the properties selected from the source term, the *vehicle*, but also on the affective (positive vs. negative) valence of the target term, the *topic*. Both the metaphorical and the affective framing play a role in the interpretation of the target, influencing people's evaluation of the whole sentence featuring the metaphor. However, it is not clear how the metaphorical framing and the affective framing interact in the evaluation of an argument featuring an emotive metaphor.

Current Research

In argumentation theory, the fallacies of ambiguity are based on some equivocation of meaning, possibly caused by different literal meanings of the same word (Walton, 1996; Tindale, 2006). In linguistics, lexical ambiguity includes both homonymy (referring to words with two completely different literal meanings) and polysemy (referring to words with two partially overlapping literal meanings). The vehicle of a conventional metaphor has a literal and a lexicalised non-literal meaning and can be considered more similar to polysemy (Carston, 1997, 2002). Novel metaphors cannot be considered as cases of lexical ambiguity, because of the completely new and creative non-literal meaning of their vehicles (Ervas, 2015). However, fallacies of ambiguity can also be caused by metaphors, whose vehicle can have both a literal and a non-literal meaning. Previous research (Ervas et al., 2015, 2018) investigated the role of different types of meaning ambiguity as a possible source of fallacious reasoning in argumentation, ranging from literal (homonymy and polysemy) to metaphorical (conventional and novel) words. Ervas et al. (2018) showed that people commit an ambiguity fallacy, especially when evaluating syllogisms with a conventional metaphor as the middle term, i.e., the term that connects the premises of an argument, and with a plausible conclusion. The authors suggested that, when arguments do not present a patently false conclusion, the participants could accept the conclusion just because it is believed to be true on the basis of *a priori* beliefs, and not because it logically follows from the premises. The *belief in the conclusion bias* is well-known to influence the overall evaluation of arguments (Evans et al., 1983; Oakhill et al., 1989; Oakhill and Garnham, 1993; Ball et al., 2006; Correia, 2011; Ball and Thompson, 2018), possibly leading to a reinterpretation of the premises in a creative search for alternative reasons to hold the conclusion (Oakhill et al., 1989; Oakhill and Garnham, 1993; Ball et al., 2006). For conventional metaphors, which go unnoticed to the participants in reading the premises, the meaning of the middle term could be revitalized to justify the conclusion (Ervas et al., 2018).

An example of a standard *equivocation fallacy* (or *quaternio terminorum*) featuring a metaphor as middle term is the following:

P1: B.B. King is a *myth*
P2: A *myth* is a traditional story
C: B.B. King is a traditional story

In the first premise (P1), "myth" is the middle term having a metaphorical meaning, i.e., famous, outstanding person, while in the second premise (P2), "myth" is the middle term having a literal meaning, i.e., traditional story. Because of the meaning shift of the term "myth," the argument assumes the structure of a *quaternio terminorum* (Barth, 1974; Macagno and Walton, 2009), i.e., a fallacious argument based on the ambiguity of its middle term (Hamblin, 1970; Woods and Walton, 1989; Copi et al., 2014). If the middle term assumes a different meaning in the two premises, then the syllogism contains four terms, rather than three, which causes the fallacy. We called "metaphoric fallacy" a *quaternio terminorum* based on an ambiguity connected to the metaphorical premise of the argument (Walton, 1996; Lightbody and Berman, 2010; Fischer, 2014).

As far as we know, previous empirical research on argument evaluation did not include "emotive metaphors" in the premises and did not focus on their double framing effect on the acceptance of the conclusion. The double framing effect of metaphors might have a strong influence on people's beliefs involved in the argument's evaluation process, possibly leading to equivocation fallacies. The double framing effect of metaphors might depend both on the type (conventional vs. novel) and on the emotional meaning (positive vs. negative valence) of the metaphor featured in the first premise of the argument. This raises the following research questions:

Q1. Under what conditions does the double framing effect of emotive metaphors mostly influence the evaluation of the argument, leading to a fallacy of equivocation?

Q2. Under what conditions does the metaphorical framing influence the evaluation of an argument? Are participants more prone to commit a fallacy of equivocation in the case of conventional metaphors or in the case of novel metaphors?
Q3. Under what conditions does the affective framing influence the evaluation of the argument? Are participants more prone to commit a fallacy of equivocation in the case of negatively valenced metaphors or in the case of positively valenced metaphors?

Two empirical studies were designed to address these questions and to investigate the double framing effect of emotive metaphors in arguments' evaluation. The studies were approved by the Ethics Committee of the Department of Education, Psychology, Philosophy, University of Cagliari (n. 25, 10/07/2018). We tested participants' evaluation of arguments having the standard syllogistic form, with two premises and a plausible conclusion. An example of syllogism with an emotive metaphor is given below:

P1: Freedom is a smile
P2: A smile is an expression of joy
C: Freedom is an expression of joy

The middle terms, such as *smile*, bridge the premises and have their metaphorical meaning in the first premise, diverging from the literal meaning in the second premise, thus leading to a fallacious - but still meaningful - conclusion. The middle terms in the first premise might be either the vehicle of a *conventional metaphor* (CM), or the vehicle of a *novel metaphor* (NM), as in the example provided above. The middle terms might be positively valenced (+), as in the example, negatively valenced (−), or non-emotive (0), based on their emotional meaning. We also devised two sets of first premises having the valence of the metaphor vehicle respectively *coherent* with the valence of the topic (Experiment 1) and *incoherent* with the valence of the topic (Experiment 2), to check whether the double framing effect depends on the emotional meaning of the vehicle or on the emotional meaning of the overall premise featuring the metaphor.

Hypotheses and Expectations

Based on previous research, we advanced the following hypotheses:

H1: Emotive metaphors, i.e., metaphors based on an "emotive word" as a vehicle, entail a double framing effect in argumentation, systematically leading one to commit an equivocation fallacy.

As only the metaphor vehicle is the middle term that bridges the premises of the syllogism, we expected the participants to provide more inaccurate responses to syllogisms with emotive metaphors than to syllogisms with non-emotive metaphors, independently from the affective coherence of the vehicle and the topic.

H2: The metaphorical framing effect on the evaluation of arguments is stronger in the case of conventional rather than novel metaphors.

We therefore expected the participants to commit more equivocation fallacies and provide less accurate responses for arguments that contain conventional metaphors compared to novel metaphors, because in the former case participants are not aware of the metaphorical framing effect, while in the latter case they are aware of the "deviant," creative use of language in the premises leading to the conclusion.

H3: The affective framing effect is stronger especially in the case of negatively valenced metaphors when compared to non-emotive metaphors.

We therefore expected a lower accuracy in the case of arguments with negatively valenced metaphors compared to non-emotive metaphors, as negative stimuli represent potential threats demanding an immediate rather than a deliberative response.

We also aimed to explore why participants could accept a fallacious syllogism as sound, when the middle term is an emotive metaphor and the argument conclusion is plausible. Based on previous research (Ervas et al., 2018), we were interested to check whether the reasons could be:

1) *Understandability*: the overall syllogism with the emotive metaphor is simply not understood by the participants, leading to an inaccurate evaluation of the argument;
2) *Convincingness*: the participants feel convinced by the argument, thus thinking it is also sound;
3) *Emotional appeal*: the participants is emotionally engaged or persuaded by the argument;
4) *Logical relation*: the participants think to have found a logical connection between the premises and the conclusion of the argument, when it is not the case;
5) *Ambiguity*: the participants think that in the argument there is no word used in two different meanings, when it is instead the case;
6) *Belief in the conclusion*: the participants believe in what is stated in the conclusion, independently of the content of the premises;
7) *Real-world experience*: the participants are used to hearing similar arguments in their everyday experience and thus uncritically accept them.

We expected that the participants' reasons to accept the fallacious arguments as sound differ from the reasons to reject them as fallacious. For instance, finding an ambiguity could be a reason to reject an argument as fallacious, while the emotional appeal of an argument could be a reason to accept a fallacious argument as sound. We also expected some differences on the participant's evaluation according to the syllogism type: for instance, syllogisms with non-emotive metaphors would be less emotionally appealing than syllogisms with emotive metaphors, or the participants' belief in the conclusion would be higher in the case of syllogisms with conventional metaphors compared to novel metaphors (Ervas et al., 2018).

Method

For both the experiments on syllogisms with affectively coherent and incoherent metaphors, we presented the participants a series of fallacious syllogisms with either an emotive or a non-emotive

metaphor as middle term and with a plausible conclusion, asking whether the conclusion of the syllogisms follows from the premises ("Yes"/"No" answer). Literal syllogisms were planned as fillers to check participants' basic ability to distinguish between clearly strong (SL) and weak (WL) literal arguments, without any explicit instruction, and to understand whether participants are prone to accept fallacious literal arguments with plausible conclusions (PL). We planned to explore why participants answered "Yes" vs. "No," asking the participants to rate the syllogisms on a 1–5 Likert scale (1 = least likely; 5 = most likely), based on the following measures:

1. *Understandability*: Do you understand the argument?
2. *Convincingness*: Is the argument convincing in any way?
3. *Emotional appeal*: Is the argument emotionally appealing?
4. *Logical relation*: Is the conclusion logically related to the premises?
5. *Ambiguity*: Is the ambiguity at any level influencing?
6. *Belief in the conclusion*: Do you believe in C (independent of P1 and P2)?
7. *Real-world experience*: Do you have any experience of similar arguments?

Experimental Design

Two experiments were designed to test both the metaphorical and the affective framing effects, and their interaction effect on the evaluation of the metaphoric fallacy. For this, we planned a 3×2 experimental design, having 3 "affective framing" conditions (non-emotive vs. emotive, i.e., positively valenced vs. negatively valenced metaphors) × 2 "metaphorical framing" conditions (conventional vs. novel metaphors). The experiments 1 and 2 were designed to test respectively affectively coherent vs. affectively incoherent metaphors as first premises of the syllogisms.

Data Analyses

We planned the following coding and data analyses for both the experiments. In the case of fallacious syllogisms with metaphors, we calculated the scores for accuracy in the following way: 0 for the incorrect answers "yes," 1 for the correct answers "no." Indeed, participants answering "yes" think that the conclusion follows from the premises, thus accepting the fallacious arguments as sound, even though the middle term is used with different meanings in the premises. For the same reasons, in both the case of clearly weak literal arguments and weak literal arguments with plausible conclusions, we attributed 0 for the incorrect answers "yes," 1 for the correct answers "no." On the contrary, in the case of strong literal arguments with middle terms used with the same meaning in both the premises, we assigned 1 to the correct answers "yes" and 0 for the incorrect answers "no."

We checked the accuracy of responses to the literal fillers, to ensure that the participants implicitly understood how to distinguish between strong and weak syllogisms, even in case of plausible conclusions. A Chi-squared test determined whether a significant difference existed between the participants' answers and the correct answers in the case of literal fillers. A series of paired *t*-tests were used to see whether weak literal (WL) arguments, especially when having plausible conclusions (PL), were more difficult to detect compared to strong literal (SL) arguments. A Chi-squared test was also used to check whether a significant difference existed between the participants' answers and the correct answers in the case of arguments with metaphors, to see whether participants systematically failed to detect the fallacy in each condition. We calculated the effect sizes, reporting the Phi coefficient (φ) (small effect size: $\varphi = 0.1$; medium effect size: $\varphi = 0.3$; large effect size: $\varphi = 0.5$).

Finally, we performed two statistical analyses for both Experiments 1 and 2:

1. An analysis of variance (ANOVA) for accuracy to assess the main effects of the two factors, metaphorical and affective middle term type, and the interaction of the two factors on the evaluation of the arguments. We calculated the effect sizes, reporting the Eta squared coefficient (η^2) (small effect size: $\eta^2 = 0.0099$; medium effect size: $\eta^2 = 0.0588$; large effect size: $\eta^2 = 0.1379$);
2. An exploratory analysis on the seven measures for both committing the fallacy (accepting the conclusion as following from the premises, incorrect answers "yes") and detecting the fallacy (discarding the conclusion as not following from the premises, correct answers "no"). A multiple linear regression was planned to explore the impact of the seven measures (Understandability, Convincingness, Emotional appeal, Logical relation, Ambiguity, Belief in the conclusion, and Real-world experience) on participants' evaluation of the different types of syllogisms, assuming that they could have different reasons to accept the fallacious syllogisms as sound ("yes" answer) and to discard the fallacious syllogisms as actually fallacious ("no" answer). We planned to create separate linear models for 3 "affective framing" conditions (non-emotive/positive/negative) × 2 "metaphorical framing" conditions (conventional/novel) for "yes" and "no" responses. A multiple linear regression was calculated to predict the dichotomous dependent variable ("yes"/"no") based on the seven predictors (understandability, convincingness, emotional appeal, logical relation, ambiguity, belief in the conclusion, and real world experience) for novel and conventional metaphors, by entering all the predictors simultaneously.

Rating and Pilot Studies

To provide the materials for both the experiments, we pre-tested the (1) *vehicles* of the metaphors, (2) metaphors in the first premises of the syllogisms, and (3) separate premises and conclusions of the syllogisms in a series of rating studies ($N = 257$ participants).

Metaphors' Vehicles

We selected a set of terms ($N = 206$ nouns, GRADIT; De Mauro, 2000) that could be used to form the first premises of the arguments. All the terms were preselected according to their number of letters (CM: M = 6.92, SD = 1.24; NM: M = 6.92, SD = 0.67) and frequency (both CM and NM vehicles belonging

TABLE 1 | Examples of arguments in English for each middle term condition (Experiment 1).

	Non-emotive metaphors (0)	Positively valenced metaphors (+)	Negatively valenced metaphors (-)
CM	[P1] That girl is a *gem*. (*gemma*) [P2] A *gem* is a precious stone. [C] That girl is a precious stone.	[P1] Peace is a *thaw*. (*schiarita*) [P2] A thaw is the defreezing of ice. [C] Peace is the defreezing of ice.	[P1] An insult is a *scar*. (*sfregio*) [P2] A *scar* is a wound. [C] An insult is a wound.
NM	[P1] The rooster is a *pharaoh*. (*faraone*) [P2] A *pharaoh* is a leader. [C] The rooster is a leader.	[P1] Freedom is a *smile*. (*risata*) [P2] A *smile* is an expression of joy. [C] Freedom is an expression of joy.	[P1] A betrayal is a *scald*. (*ustione*) [P2] A *scald* is a serious burn. [C] A betrayal is a serious burn.

to the "common terms" frequency category in the GRADIT, De Mauro, 2000). From this set, we selected the *metaphors' vehicles*, which were the *same* in both Experiment 1 and 2, constituting the middle terms of the syllogisms. We selected the *metaphors' vehicles*, based on the results of a rating study on their familiarity and emotional (positively and negatively valenced) meaning, on a 1–5 Likert Scale (1 = very negative/very unfamiliar, 5 = very positive/very familiar). Terms with insufficient familiarity (M_{familiar} <2) were excluded. We used three sets of *metaphors' vehicles* to form the metaphors: (1) terms with definite emotional meanings (M_{positive} <4; M_{negative} >2) were used as vehicles of non-emotive metaphors; (2) terms with definite positive emotional meanings (M_{positive} >4) were used as vehicles of positively valenced metaphors, and (3) terms with definite negative emotional meanings (M_{negative} <2) were used as vehicles of negatively valenced metaphors. Terms already having a lexicalized metaphorical meaning in the GRADIT were used as vehicles of *conventional metaphors*. Unambiguous terms were used as vehicles of *novel metaphors*, ensuring that they had no already lexicalized figurative meanings in the GRADIT.

Metaphors

From the preselected set of terms, we also selected the *metaphors' topics*, based on their ratings for familiarity and emotional (positively and negatively valenced) meaning, on a 1–5 Likert Scale (1 = very negative/unfamiliar, 5 = very positive/familiar). We devised three sets of *metaphors' topics* to form the metaphors: (1) non-emotive topics with definite emotional meanings (M_{positive} <4; M_{negative} >2); (2) positively valenced topics with definite emotional meanings (M_{positive} >4); (3) negatively valenced topics with definite emotional meanings (M_{negative} <2). A set of conventional and novel *non-emotive metaphors*, the same for both the experiments, was generated with non-emotive metaphors' vehicles and topics for the first premises of the syllogisms with non-emotive metaphors. Two sets of *emotive metaphors* constituted the first premises of the syllogisms with emotive metaphors: (1) a set of first premises where the emotional meaning of the metaphors' vehicles was coherent with the emotional meaning of the metaphors' topics (Experiment 1); (2) a set of first premises where the emotional meaning of the metaphors' vehicles was incoherent with the emotional meaning of the metaphors' topics (Experiment 2).

In two separate rating studies for Experiments 1 and 2, the sets of metaphors were tested along some major psycholinguistic variables (Bambini et al., 2014): emotional (positively and negatively valenced) meaning, familiarity, meaningfulness (i.e., confidence in metaphor interpretation), and comprehension difficulty using a 1–5 Likert scale (1 = very negative/unfamiliar/meaningless/easy, 5 = very positive/familiar/meaningful/difficult). Metaphors with insufficient meaningfulness ($M_{\text{meaningful}}$ <2) and metaphors too difficult to understand ($M_{\text{difficult}}$ >4) were excluded. We deemed metaphors with no definite emotional meanings (M_{positive} <4; M_{negative} >2) as non-emotive metaphors, and the metaphors with definite emotional meanings (M_{positive} >4; M_{negative} <2) as positively valenced and negatively valenced emotive metaphors (see **Table A1** in **Appendix** for M and SD for each measure).

Premises and Conclusions

The second premises of the syllogisms, the *same* in both materials of Experiment 1 and 2, made explicit the literal meaning of the middle term provided in the GRADIT, which differs from the metaphorical meaning of the middle term in the first premises. We pre-tested the separate premises of all the arguments to ensure that the participants actually attributed different meanings to the middle terms in the first and the second premises of the syllogisms. Besides the topics of emotive metaphors in the first premises, the materials of the experiments differed for their conclusions, which connected the metaphor's topic to the last term of the second premises. We also separately tested all the premises and conclusions of the syllogisms to ensure that participants perceived them as true or at least plausible, to avoid false premises leading to an "ex falso quodlibet."

Pilot Study

A pilot study on syllogisms with emotionally coherent premises showed that participants (N = 13, nine women, four men) accepted more fallacious arguments as sound when having emotive metaphors rather than non-emotive metaphors as middle terms.

Experiment 1

The goal of the experiment was to test the evaluation of metaphoric fallacies, having a syllogistic form and an emotive metaphor (with coherently-valenced vehicle and topic) in the first premise. We aimed to understand whether and why participants were prone to accept a *quaternio terminorum* with a plausible conclusion as sound, especially in the case of emotive metaphors. We also aimed to explore how different factors (understandability, convincingness, emotional appeal, logical relation, ambiguity, belief in the conclusion, and real-world experience) contribute to the participants' evaluation of

TABLE 2 | Mean (M) and standard deviation (SD) values of correct answers for each middle term condition (Experiment 1).

	Positively valenced		Negatively valenced		Non-emotive	
	M	*SD*	*M*	*SD*	*M*	*SD*
CM	0.16	0.21	0.09	0.17	0.15	0.25
NM	0.23	0.29	0.21	0.27	0.31	0.30
Literal	**Strong arguments**		**Weak arguments**		**Arguments with plausible conclusion**	
	0.97	0.13	0.90	0.24	0.79	0.26

arguments with a plausible conclusion, comparing emotive, and non-emotive metaphorical middle terms conditions.

Participants

The participants (93 adults, 50 women, 43 men) were undergraduate students in Communication Science at the University of Cagliari, had Italian as their first language, and normal/corrected vision. Since we aimed to check for the participants' intuitive answers on different measures concerning the acceptability of argument conclusions, we excluded participants ($N = 2$) who had advanced training in logic and/or argumentation theory, resulting in 91 participants (48 women, 43 men, $M_{age} = 23.58$ years, $SD_{age} = 7.41$ years).

Materials

Participants were presented with a set of $N = 36$ arguments in Italian (see **Table A2** in **Appendix**), having the structure of syllogisms with plausible conclusions. The set of arguments contained $N = 8$ non-emotive metaphors, four conventional (*CM0*) and four novel (*NM0*); $N = 8$ positively valenced metaphors, four conventional (*CM+*) and four novel (*NM+*); $N = 8$ negatively valenced metaphors, four conventional (*CM−*) and four novel (*NM-*), in their first premise. **Table 1** presents an example of argument translated into English for each middle term condition.

The set of arguments also included 12 literal arguments as fillers: four clearly weak literal arguments, four clearly strong literal arguments, and four literal arguments with plausible conclusions (see **Table A3** in **Appendix** for the literal arguments in Italian).

Procedure

The data was collected through an online form. After the participants signed the informed consent, the form gathered information about gender, age, language, and education. Participants were then asked to read the instructions and complete two practice trials to familiarize themselves with the task. The syllogisms were then randomly presented, followed by some questions. After each argument, the following question appeared: "Does the conclusion follow from the premises?," asking the participants to answer "Yes" or "No." Participants were then asked to rate the arguments on a 1–5 Likert scale (1 = least likely, 5 = most likely), answering the questions for the measures: (1) Understandability; (2) Convincingness; (3) Emotional appeal; (4) Logical relation; (5) Ambiguity; (6) Belief in the conclusion; (7) Real-world experience. The experiment lasted 30 min ca.

Results

All the data collected are available at the following **OSF address** https://osf.io/jzpva/?view_only=8edc3b523cbb4afba4bd71978d847a48. Mean (M) and standard deviation (SD) for accuracy are presented in **Table 2**. First of all, we checked whether the participants correctly answered in the case of literal arguments, to understand whether they carefully performed the task and whether they have a basic ability in detecting clearly strong and clearly weak arguments. The participants performed almost at ceiling for all literal arguments. The Chi-squared test showed that the participants provided significantly more correct than incorrect answers to literal arguments ($p < 0.001$; LS: $\chi^2 = 134.00$, $\varphi = 0.86$; LW: $\chi^2 = 102.34$, $\varphi = 0.75$; LP: $\chi^2 = 57.36$, $\varphi = 0.56$). A series of paired *t*-tests suggested that fallacious literal arguments with plausible conclusions were more difficult to detect when compared to both clearly strong ($p < 0.001$) and clearly weak arguments ($p < 0.01$).

The Chi-squared test showed that participants provided more incorrect than correct answers to arguments with metaphors ($p < 0.001$; CM0: $\chi^2 = 76.43$, $\varphi = 0.65$; CM+: $\chi^2 = 69.24$, $\varphi = 0.62$; CM−: $\chi^2 = 99.98$, $\varphi = 0.74$; NM0: $\chi^2 = 28.45$, $\varphi = 0.40$; NM+: $\chi^2 = 51.65$, $\varphi = 0.53$; NM-: $\chi^2 = 51.86$, $\varphi = 0.53$). A two-way ANOVA was performed for accuracy to assess the main effects of the metaphorical and affective middle term type and the interaction of the two factors on the evaluation of the fallacious arguments. The results of the main effects are reported in **Table 3**.

Overall, the results showed a significant main effect of the metaphorical type [$F_{(1,89)} = 29.30$; $p < 0.001$; $\eta^2 = 0.05$] and the affective type [$F_{(2,88)} = 4.63$; $p = 0.01$; $\eta^2 = 0.016$], but no significant interaction of the metaphorical type and the affective type [$F_{(2,86)} = 1.65$; $p = 0.192$; $\eta^2 = 0.006$] on participant's evaluation of the arguments. The significant main effect of the metaphorical type is due to the lower number of correct answers in the case of arguments with conventional metaphors when compared to arguments with novel metaphors. A *post-hoc* Tukey's test, corrected for multiple comparisons, was performed to determine the statistical significance of the difference between specific affective conditions (see **Table 4** for all the results). The *post-hoc* analysis revealed that the significant main effect of the affective type is due to the lower number of correct answers in the case of arguments with negatively valenced metaphors when compared to arguments with non-emotive metaphors [$t_{(89)} = 0.32$; $p < 0.01$].

We analyzed the answers provided by the participants on different possible factors influencing the arguments' evaluation (see M and SD values for each measure in **Table 5**). A multiple linear regression analysis was performed on the data: separate linear models were created for 3 "affective framing" conditions × 2 "metaphorical framing" conditions for both "yes" and "no" responses. **Table 6** presents the results of the linear regression for all middle term conditions.

TABLE 3 | Main effects of metaphorical and affective middle term type for accuracy in Experiments 1 and 2.

	Accuracy (Experiment 1)						Accuracy (Experiment 2)					
	SS	**df**	***F***	***P***	η^2	***f***	**SS**	**df**	***F***	***p***	η^2	***f***
C (Metaphorical)	1.9	1	29.30	<0.001	0.05	0.23	0.0	1	0.01	0.91	0.00	0.08
C (Affective)	0.6	2	4.63	0.01	0.016	0.13	5.6	2	2.11	0.12	0.007	0.01
C (Metaphorical): C(Affective)	0.2	2	1.65	0.192	0.006	0.08	20.9	2	7.89	<0.001	0.027	0.17

SS, Sum of squares; df, degrees of freedom; η^2, Eta squared (small: $\eta^2 = 0.01$; medium: $\eta^2 = 0.059$; large: $\eta^2 = 0.138$); f, Cohen's effect size (small: f = 0.1, medium: f = 0.25, large: f = 0.4).

TABLE 4 | *t*/*p*-values for accuracy, comparing middle term affective conditions (Experiment 1).

	Comparisons	***t***	***p***
C (Affective)	neutral/negative	0.32	<0.01
	positive/negative	0.18	0.19
	positive/neutral	−0.13	0.40

A multiple linear regression was calculated to predict "yes" and "no" based on understandability, convincingness, emotional appeal, logical relation, ambiguity, belief in the conclusion and real world experience for novel and conventional metaphors. When "yes" was predicted it was found that seven predictors in positive ($\beta = 0.06$, $p < 0.05$), negative ($\beta = 0.66$, $p < 0.05$), and non-emotive ($\beta = 0.12$, $p < 0.05$) metaphorical middle term conditions were significant predictors in the case of conventional metaphors. The overall model fit for positive metaphor was $R^2 = 0.64$, for negative metaphor was $R^2 = 0.33$ and non-emotive metaphors was $R^2 = 0.65$. When "yes" was predicted it was found that six predictors in positive ($\beta = 0.09$, $p < 0.05$), negative ($\beta = 0.24$, $p < 0.05$), and non-emotive ($\beta = 0.14$, $p < 0.05$) metaphorical middle term conditions were significant predictors in the case of novel metaphors. The overall model fit for positive metaphor was $R^2 = 0.05$, for negative metaphor was $R^2 = 0.47$ and non-emotive metaphor was $R^2 = 0.55$. When "no" was predicted it was found that seven predictors in positive ($\beta = -0.02$, $p < 0.05$), negative ($\beta = 0.22$, $p < 0.05$), and non-emotive ($\beta = 0.07$, $p < 0.05$) metaphorical middle term conditions were significant predictors in the case of conventional metaphors. The overall model fit for positive metaphor was $R^2 = 0.14$, for negative metaphor was $R^2 = 0.31$ and non-emotive metaphors was $R^2 = 0.27$. When "no" was predicted it was found that six predictors in positive ($\beta = 0.26$, $p < 0.05$), negative ($\beta = -0.02$, $p < 0.05$), and non-emotive ($\beta = 0.17$, $p < 0.05$) metaphorical middle term conditions were significant predictors in the case of novel metaphors. The overall model fit for positive metaphor was $R^2 = 0.27$, for negative metaphor was $R^2 = 0.16$ and non-emotive metaphor was $R^2 = 0.07$.

The results showed that understandability was a significant predictor for both accepting and discarding the conclusion of the argument as following from the premises with both conventional and novel metaphors. When the conclusion was perceived to be following from the premises, for any argument with emotive metaphors, both convincingness and emotional appeal were significant predictors for committing the equivocation fallacy [Convincingness: CM+ ($t_{(89)} = 2.11$; $p < 0.05$), CM− ($t_{(89)} = 1.79$; $p < 0.05$), NM+ ($t_{(89)} = 2.11$; $p < 0.01$), and NM- ($t_{(89)} = 2.97$; $p < 0.01$); Emotional appeal: CM+ ($t_{(89)} = 2.27$; $p < 0.01$), CM− ($t_{(89)} = 2.35$; $p < 0.01$), NM+ ($t_{(89)} = 2.67$; $p < 0.05$), and NM- ($t_{(89)} = 3.14$; $p < 0.001$)]. Emotional appeal was a significant predictor also in the case of novel non-emotive metaphors ($t_{(89)} = 2.19$; $p < 0.01$). Interestingly, in the case of negatively valenced metaphors, also the perception of having found a logical relation in the argument and the belief in the conclusion, independent from the premises, were significant predictors for accepting the plausible conclusion [Logical relation: CM− ($t_{(89)} = 2.19$; $p < 0.01$), NM - ($t_{(89)} = 2.54$; $p < 0.01$); Belief in the conclusion: CM− ($t_{(89)} = 2.81$; $p < 0.001$), NM- ($t_{(89)} = 3.44$; $p < 0.001$)]. The perception of a logical relation was a significant predictor in the case of CM+ arguments ($t_{(89)} = 2.14$; $p < 0.01$), but not in the case of CM0 arguments. In the case of arguments with positively valenced metaphors, the belief in the conclusion was not a significant predictor for committing the equivocation fallacy, while it was significant in the case of NM0 arguments ($t_{(89)} = 2.46$; $p < 0.001$).

When the conclusion was seen not to be following from the premises, ambiguity was a significant predictor for all the arguments featuring novel metaphors [NM0 ($t_{(89)} = 2.17$; $p < 0.01$), NM+ ($t_{(89)} = 2.89$; $p < 0.05$), and NM- ($t_{(89)} = 2.14$; $p < 0.05$)], independent of their valence. In the case of negatively valenced metaphors, both ambiguity and convincingness were significant predictors for detecting the equivocation fallacy [Ambiguity: CM− ($t_{(89)} = 3.92$; $p < 0.05$), NM- ($t_{(89)} = 2.14$; $p < 0.05$); Convincingness: CM− ($t_{(89)} = 2.79$; $p < 0.05$), NM- ($t_{(89)} = 2.47$; $p < 0.05$)]. In the case of positively valenced metaphors, emotional appeal was a significant predictor for detecting the equivocation fallacy [Emotional appeal: CM+ ($t_{(89)} = 3.14$; $p < 0.01$), NM+ ($t_{(89)} = 2.19$; $p < 0.05$)]. In the case of non-emotive metaphors, having found a logical relation between premises and conclusion and real-world experience of similar arguments were significant predictors for detecting the fallacy [Logical relation: CM0 ($t_{(89)} = 3.41$; $p < 0.001$), NM0 ($t_{(89)} = 3.46$; $p < 0.05$); Real world experience: CM0 ($t_{(89)} = 3.44$; $p < 0.05$), NM0 ($t_{(89)} = 3.47$; $p < 0.05$)].

Discussion

The results of the Chi-squared test confirmed that participants mostly fail in the evaluation of syllogisms with metaphors and plausible conclusions (Ervas et al., 2018). The significant

TABLE 5 | Mean (M) and standard deviation (SD) values of each predictor for each middle term condition (Experiment 1).

Predictor	CM+		CM-		CM0		NM+		NM-		NM0	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Understandability	3.23	0.66	3.57	0.66	3.21	0.67	2.92	0.80	3.00	0.67	2.51	0.67
Convincingness	3.01	0.65	3.38	0.70	2.99	0.77	2.63	0.85	2.64	0.72	2.27	0.84
Emotional appeal	2.33	0.82	2.50	0.80	2.34	0.78	2.49	0.99	2.30	0.79	1.49	0.55
Logical relation	3.11	0.75	3.45	0.71	3.24	0.81	2.89	0.84	2.92	0.78	2.55	0.76
Ambiguity	2.27	0.75	1.89	0.63	2.25	0.80	2.46	0.91	2.52	0.78	2.84	0.89
Belief in the conclusion	3.18	0.67	3.48	0.76	3.11	0.77	2.81	0.92	2.94	0.86	2.56	0.77
Real world experience	2.85	0.82	3.20	0.86	2.93	0.89	2.75	0.99	2.84	0.85	1.96	0.83

Positively valenced (CM+), negatively valenced (CM-), non-emotive (CM0) conventional metaphor. Positively valenced (NM+), negatively valenced (NM-), non-emotive (NM0) novel metaphor.

TABLE 6 | *t*/*p*-values for each predictor in the evaluation of arguments, comparing the middle term conditions (Experiment 1).

		Conventional metaphor						Novel metaphor					
Answer	**Predictors**	**Positive**		**Negative**		**Non-emotive**		**Positive**		**Negative**		**Non-emotive**	
YES		$\beta = 0.06$, $R^2 = 0.64$, $F^2 = 0.22$		$\beta = 0.66$, $R^2 = 0.33$, $F^2 = 0.37$		$\beta = 0.12$, $R^2 = 0.65$, $F^2 = 0.11$		$\beta = 0.09$, $R^2 = 0.05$, $F^2 = 0.28$		$\beta = 0.24$, $R^2 = 0.47$, $F^2 = 0.32$		$\beta = 0.14$, $R^2 = 0.55$, $F^2 = 0.34$	
		β	*t*	*β*	*t*	*β*	*t*	*β*	*t*	*β*	*t*	*β*	*t*
	Understandability	0.51	2.17**	0.22	3.43**	0.33	3.21*	0.41	1.09	0.44	3.14**	0.33	3.15*
	Convincingness	0.32	2.11*	0.27	1.79*	0.08	1.16	0.22	2.11**	0.35	2.97**	0.11	2.62
	Emotional appeal	0.22	2.27**	0.34	2.35**	0.11	1.72	0.44	2.67*	0.57	3.14***	0.43	2.19**
	Logical relation	0.41	2.14**	0.33	2.19**	0.19	1.53	0.32	2.11	0.46	2.54**	0.44	2.22*
	Ambiguity	0.17	2.7	0.13	2.66	0.21	2.72	0.08	2.17	0.52	2.19**	0.17	1.76
	Belief in the conclusion	0.21	2.31	0.45	2.81***	0.12	1.65	0.04	2.12	0.71	3.44***	0.61	2.46**
	Real world experience	0.32	2.19*	0.32	2.72	0.07	2.61	0.06	2.97	0.12	1.93	0.15	2.71
NO		$\beta = -0.02$, $R^2 = 0.14$, $F^2 = 0.18$		$\beta = 0.22$, $R^2 = 0.31$, $F^2 = 0.22$		$\beta = 0.07$, $R^2 = 0.27$, $F^2 = 0.21$		$\beta = 0.26$, $R^2 = 0.27$, $F^2 = 0.21$		$\beta = -0.02$, $R^2 = 0.16$, $F^2 = 0.12$		$\beta = 0.17$, $R^2 = 0.07$, $F^2 = 0.11$	
		β	*t*	*β*	*t*	*β*	*t*	*β*	*t*	*β*	*t*	*β*	*t*
	Understandability	0.19	2.1	0.56	3.41**	0.56	3.17***	0.32	3.16**	0.42	1.96**	0.51	2.45**
	Convincingness	0.11	2.71	0.43	2.79*	0.06	1.19	0.11	1.8	0.36	2.47*	0.10	2.71
	Emotional appeal	0.36	3.14**	0.21	1.16	0.15	3.32	0.25	2.19*	0.10	2.96	0.32	2.27*
	Logical relation	0.17	2.76	0.11	1.88	0.66	3.41***	0.22	1.14	0.22	3.27	0.46	3.46*
	Ambiguity	0.11	1.16	0.34	3.92*	0.12	3.11	0.27	2.89*	0.40	2.14*	0.50	2.17**
	Belief in the conclusion	0.09	3.5*	0.20	2.18	0.14	2.12	0.09	1.18	0.13	2.96	0.10	2.24
	Real world experience	0.05	2.18	0.21	1.67	0.38	3.44*	0.11	2.92	0.06	2.74	0.39	3.47*

$^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

main effect of the metaphor type of the middle term on the evaluation of the arguments suggests that participants are more prone to commit the metaphoric fallacy when a conventional metaphor occurs in the first premise. This effect is independent of whether conventional metaphors are based on an emotive vehicle or not, and it can be explained by the metaphorical interference effect (MIE) (Glucksberg et al., 1982) in the truth evaluation of the first premise. When reading the premise with a conventional metaphor, we can assign it intuitive truth conditions, even though it is literally false. In a contextualist pragmatic perspective (Sperber and Wilson, 1986; Carston, 2002; Recanati, 2004, 2010), we understand the premise when we know the concrete circumstances of truth, which can depart from their literal ones. Processes of lexical modulation are supposed to adjust the literal meaning of the sentence and to provide an "adjusted" meaning, better fitting the context. This is also the

case of conventional metaphors, whose vehicle encodes a source concept that is contextually modulated to generate an *ad hoc* concept in the proposition the speaker intends to communicate (Carston, 1997, 2002), corresponding to the intuitive truth-conditions she assigned. Diverging from the predictions of classical argumentation, we commit the metaphoric fallacy because we systematically and unawarely reject the "literal" truth conditions in the pragmatic process of conventional metaphor understanding, thus compromising the evaluation of the strength of the overall argument (Ervas et al., 2015).

Another route to metaphor understanding is followed when understanding novel metaphors (Carston, 2010, 2018). Especially in the case of unfamiliar and novel metaphors, the literal meaning of the vehicle would linger in the interpretation process of the premise, possibly requiring more contextual information to be intuitively perceived as true (Indurkhya, 2006, 2016; Carston, 2010; Ervas, 2019). However, the sentential context of the first premise, as well as the narrow syllogistic context of argumentation, would make novel metaphors' interpretation more easily recognized as false when compared to conventional metaphors. In a sense, novel (rather than conventional) metaphors are processed in a Gricean style in the first premise of the syllogism, as they are recognized from the very beginning as "patently false." As there is a strict link between the truth conditions of the premises and the evaluation of the strength of the whole argument, it would be easier for participants to detect the metaphoric fallacy in the case of novel metaphors. Also, the suppression mechanism of the literal meaning should be more difficult in novel rather than in conventional metaphors interpretation (Gernsbacher and Faust, 1991; Gernsbacher et al., 2001; Rubio Fernandez, 2007). This would be crucial for the detection of a fallacy based on the equivocation between the metaphorical and the literal meaning of the middle term. Participants recognized ambiguity as a reason for rejecting the argument as unsound in all the cases of arguments with novel metaphors, independently of their emotive valence.

The significant main effect of the affective type of the middle term on the evaluation of the arguments and the *post-hoc* analysis' results suggest that participants are more prone to commit the metaphoric fallacy when a negatively valenced rather than a non-emotive metaphor occurs in the first premise. This effect is independent of whether the metaphors are conventional or novel and confirms previous studies on the effect of emotions on deductive reasoning: even though the validity of a conclusion should not depend on the emotional valence of the premises, it has been shown that negatively valenced contents are associated with decreased logicality compared to neutral contents (Lefford, 1946; Blanchette and Richards, 2004; Blanchette, 2006; Blanchette and Leese, 2011). Beyond confirming that the affective valence of the content influences normatively correct answers when reasoning about syllogisms, the results seem to confirm also the idea of a general negative effect of emotions on cognitive performance [see e.g., Lieberman et al. (2005)]. The answers provided by the participants to the seven questions - following the main question on the argument strength - let us better understand *why* participants were more prone to accept the metaphoric fallacy as sound in the case of emotive metaphors. When participants accepted the metaphoric fallacy as sound, its *emotional appeal* and *convincingness* had a crucial role, while the logical structure and real-life experience played a major role in the case of non-emotive metaphors, i.e., when participants did detect the metaphoric fallacy. However, emotional appeal is also a significant predictor for *detecting* the metaphoric fallacy in the case of positively valenced metaphor: this suggests that the positive valence of the metaphorical middle term might increase participants' logicality, thus questioning the idea that emotional content always undermines cognitive performance in reasoning tasks.

The absence of a significant interaction between the metaphor type and the affective type of the middle term suggests that the metaphorical framing and the affective framing might act independently. Nonetheless, when the affective values of the vehicle and the topic are coherent, both effects are in place in the evaluation of arguments featuring the metaphor. On the one hand, especially in the case of conventional metaphor, the metaphorical framing could have been covert (Thibodeau and Boroditsky, 2011, 2015) and thus more influential in guiding the participants to make evaluations consistent with the metaphorical heuristic rather than with logicality (Robins and Mayer, 2000). On the other hand, especially in the case of negatively valenced metaphors, the affective framing might have led participants to derail from the route to logicality, embracing heuristics when reasoning (Eliades et al., 2013). Previous research has argued in favor of the thesis that "emotions promote a form of reasoning which is less analytical and more heuristic-based" (Blanchette et al., 2018, p. 61), especially when the arguments are susceptible to the biasing effects of *conclusion believability* (Eliades et al., 2012). The results of the experiment confirm that, especially in the case of negatively valenced metaphors, participants' belief in the conclusion significantly leads them to accept the metaphoric fallacy as a sound argument, even thinking of having found a logical relation between premises and conclusion. However, controversial results on the influence of negatively valenced emotional content in reasoning can be found within the literature (Hofmann et al., 2009; MacKuen et al., 2010): it might be claimed that not necessarily the influence of negatively valenced stimuli leads participants to make more "errors." Negatively valenced emotional content might have led the participants to find alternative reasons, regardless of their logical validity or of the argument strength, thus reinterpreting the premises of the arguments with metaphors to make sense of the believed conclusion. In this perspective, a global process of sense-making would have precedence over the analytic process in the evaluation of the "metaphoric fallacy" and lead the (re)interpretation of the overall argument as metaphorical.

Experiment 2

The goal of the experiment was to test the evaluation of metaphoric fallacies, having a syllogistic form and - in their first premise - an emotive metaphor, whose vehicle has an emotional meaning *incoherent* with the emotional meaning of the topic. We aimed to understand whether the emotional meaning of the overall sentential metaphorical context of the first premise or the

TABLE 7 | Examples of arguments in English for each middle term condition (Experiment 2).

	Non-emotive metaphors (0)	Positively valenced metaphors (+)	Negatively valenced metaphors (-)
CM	[P1] That girl is a *gem*. (*gemma*) [P2] A *gem* is a precious stone. [C] That girl is a precious stone.	[P1] A separation is a *thaw*. (*schiarita*) [P2] A thaw is the defreezing of ice. [C] A separation is the defreezing of ice.	[P1] Bravery is a *scar*. (*sfregio*) [P2] A *scar* is a wound. [C] Bravery is a wound.
NM	[P1] The rooster is a *pharaoh*. (*faraone*) [P2] A *pharaoh* is a leader. [C] The rooster is a leader.	[P1] Insolence is a *smile*. (*risata*) [P2] A *smile* is an expression of joy. [C] Insolence is an expression of joy.	[P1] Religion is a *scald*. (*ustione*) [P2] A *scald* is a serious burn. [C] Religion is a serious burn.

specific emotional meaning of the vehicle as middle term leads the participants to accept a fallacious argument as sound.

Participants

The participants (99 adults, 69 women, 33 men) were undergraduate students in Communication Science recruited at the University of Cagliari, had Italian as their first language, and normal/corrected vision. As in the first experiment, we checked for the participants' intuitive answers on different measures concerning the acceptability of argument conclusions, and excluded participants ($N = 3$) who had advanced training in logic and/or argumentation theory, resulting in 96 participants (63 women, 33 men, $M_{age} = 23.90$ years, $SD_{age} = 8.39$ years).

Materials

Participants were presented with a set of $N = 36$ arguments in Italian (**Table A2** in **Appendix**), including the 12 literal arguments-fillers presented in the first experiment (**Table A3** in **Appendix**). The arguments had the structure of syllogisms with plausible conclusions and contained $N = 8$ non-emotive metaphors, 4 CM0 and 4 NM0; $N = 8$ positively valenced metaphors, 4 CM+ and 4 NM+; $N = 8$ negatively valenced metaphors, 4 CM− and 4 NM-, in their first premise. **Table 7** presents an example argument for each middle term condition.

The set of syllogisms with non-emotive metaphors was the same as used in the first study, and the set of syllogisms with emotive metaphors had the same second premises of the syllogisms as in the first study. The emotive metaphors had the same metaphor vehicles as middle terms of the syllogisms, but different metaphor topics when compared to the emotive metaphors used in Experiment 1. In Experiment 2, the emotive metaphors in the first premises of the syllogisms were emotionally incoherent. Consequently, the conclusions differed from the conclusions of the syllogisms in Experiment 1, because they connect different metaphor topics to the last term of each second premise. Still, the evaluation of the syllogisms depends on the same middle terms, which are the metaphor vehicles.

Procedure

We collected the data through an online form, where all the participants signed their informed consent and provided information about their gender, age, language, and education. We followed the same procedure used in Experiment 1, asking participants to read the instructions, complete the practice block and to evaluate the randomly presented arguments. For each argument, the participants were asked to evaluate whether the conclusion of the argument followed from the premises ("Yes" or "No" answers) and to rate the arguments on a 1–5 Likert scale (1 = least likely, 5 = most likely), based on the answers to the measures: (1) Understandability; (2) Convincingness; (3) Emotional appeal; (4) Logical relation; (5) Ambiguity; (6) Belief in the conclusion; (7) Real-world experience. Participants employed ca. 30 min to complete the experiment.

TABLE 8 | Mean (M) and standard deviation (SD) values of correct answers for each middle term condition (Experiment 2).

	Positively valenced		Negatively valenced		Non-emotive	
	M	*SD*	*M*	*SD*	*M*	*SD*
CM	0.23	0.30	0.30	0.34	0.15	0.26
NM	0.28	0.32	0.17	0.23	0.24	0.27
Literal	**Strong arguments**		**Weak arguments**		**Arguments with plausible conclusion**	
	0.97	0.10	0.89	0.21	0.80	0.24

Results

All the data collected are available at the following **OSF address** (https://osf.io/jzpva/?view_only=8edc3b523cbb4afba4bd71978d847a48). Mean and standard deviation for accuracy are presented in **Table 8**. We first checked whether the participants correctly answered the fillers, i.e., the literal arguments, to understand whether they actually have a basic ability in detecting clearly strong vs. weak arguments. As in the first experiment, the participants performed almost at ceiling in evaluating literal arguments. The Chi-squared test suggested that they provided significantly more correct than incorrect answers to literal arguments ($p < 0.001$; LS: $\chi^2 = 140.29$, $\varphi = 0.85$; LW: $\chi^2 = 95.41$, $\varphi = 0.70$; LP: $\chi^2 = 62.00$, $\phi = 0.57$). We performed paired *t*-tests between literal arguments with plausible conclusions and clearly strong vs. weak literal arguments: the results suggested that fallacious arguments with plausible conclusions were more difficult to detect when compared to both clearly strong ($p < 0.001$) and clearly weak arguments ($p < 0.01$). The Chi-squared test also suggested that participants provided more incorrect than correct answers to arguments with metaphors ($p < 0.001$; CM0:

TABLE 9 | *t*/*p*-values for accuracy, comparing middle term conditions (Experiment 2).

	Comparisons	*t*	*p*
C(Metaphorical): C (Affective)	CM0/CM- middle terms	−0.60	0.004
	CM+/CM- middle terms	−0.29	0.49
	NM-/CM- middle terms	−0.52	0.02
	NM0/CM- middle terms	−0.25	0.66
	NM+/CM- middle terms	−0.093	0.99
	CM+/CM0 middle terms	0.31	0.41
	NM-/CM0 middle terms	0.08	0.99
	NM0/CM0 middle terms	0.35	0.27
	NM+/CM0 middle terms	0.51	0.02
	NM-/CM+ middle terms	−0.22	0.73
	NM0/CM+ middle terms	0.04	0.99
	NM+/CM+ middle terms	0.19	0.84
	NM0/NM- middle terms	0.27	0.57
	NM+/NM- middle terms	0.42	0.10
	NM+/NM0 middle terms	0.15	0.93

Positively valenced (CM+), negatively valenced (CM-), non-emotive (CM0) conventional metaphor. Positively valenced (NM+), negatively valenced (NM-), non-emotive (NM0) novel metaphor.

$\chi^2 = 81.91$, $\varphi = 0.65$; CM+: $\chi^2 = 52.76$, $\varphi = 0.52$; CM−: $\chi^2 = 40.48$, $\varphi = 0.46$; NM0: $\chi^2 = 44.50$, $\varphi = 0.48$; NM+: $\chi^2 = 47.35$, $\varphi = 0.50$; NM-: $\chi^2 = 67.50$, $\varphi = 0.59$).

A two-way ANOVA was performed for accuracy, to evaluate the main effects of the metaphorical and affective middle term type and the effect of their interaction on the evaluation of the fallacious arguments. The results of the main effects are reported in **Table 3**, comparing them to the results of the first experiment. Differently from the first experiment, the ANOVA results showed no significant main effects of the metaphorical [$F_{(1,94)} = 0.01$; $p = 0.91$; $\eta^2 = 0.00$] and affective [$F_{(2,93)} = 2.11$; $p = 0.12$; $\eta^2 = 0.007$] middle term type, but instead a significant interaction effect between them [$F_{(2,91)} = 7.89$; $p < 0.001$; $\eta^2 = 0.027$]. A Tukey's test, corrected for multiple comparisons, was performed to assess the statistical significance of the interaction effect (see **Table 9** for all the results). The *post-hoc* analysis revealed a higher level of accuracy especially in the evaluation of CM− arguments, when compared to both CM0 arguments [$t_{(94)} = 0.60$; $p = 0.004$] and NM- arguments [$t_{(94)} = 0.52$; $p = 0.02$]. Furthermore, it showed a significant difference between the level of accuracy in the evaluation of NM+ arguments when compared to CM0 arguments, which received less correct answers [$t_{(94)} = 0.51$; $p = 0.02$].

A multiple linear regression was calculated to predict "yes" and "no" based on understandability, convincingness, emotional appeal, logical relation, ambiguity, belief in the conclusion and real world experience for novel and conventional metaphors (see M and SD values for each measure in **Table 10**). **Table 11** presents the results of the linear regression for all middle term conditions. When "yes" was predicted it was found that four predictors in positive ($\beta = 0.16$, $p < 0.05$), negative ($\beta = 0.44$, $p < 0.05$), and non-emotive ($\beta = 0.37$, $p < 0.05$) metaphorical middle term conditions were significant predictors in the case of conventional metaphors. The overall model fit for positive metaphor was $R^2 = 0.24$, for negative metaphor was $R^2 = 0.21$ and non-emotive metaphors was $R^2 = 0.25$. When "yes" was predicted it was found that seven predictors in positive ($\beta = 0.11$, $p < 0.05$), negative ($\beta = 0.27$, $p < 0.05$), and non-emotive ($\beta = 0.27$, $p < 0.05$) metaphorical middle term conditions were significant predictors in the case of novel metaphors. The overall model fit for positive metaphor was $R^2 = 0.07$, for negative metaphor was $R^2 = 0.24$ and non-emotive metaphors was $R^2 = 0.35$. When "no" was predicted it was found that four predictors in positive ($\beta = 0.19$, $p < 0.001$), negative ($\beta = 0.12$, $p < 0.01$), and non-emotive ($\beta = 0.13$, $p < 0.05$) metaphorical middle term conditions were significant predictors in the case of conventional metaphors. The overall model fit for positive metaphor was $R^2 = 0.04$, for negative metaphor was $R^2 = 0.41$ and non-emotive metaphors was $R^2 = 0.17$. When "no" was predicted it was found that seven predictors in positive ($\beta = 0.16$, $p < 0.05$), negative ($\beta = -0.07$, $p < 0.05$), and non-emotive ($\beta = 0.07$, $p < 0.05$) metaphorical middle term conditions were significant predictors in the case of novel metaphors. The overall model fit for positive metaphor was $R^2 = 0.13$, for negative metaphor was $R^2 = 0.11$ and non-emotive metaphor was $R^2 = 0.19$.

As in the first experiment, the results of the linear regression showed that understandability was a significant predictor for both accepting and discarding the conclusion of the argument as following from the premises in both conventional and novel metaphors. When the conclusion was seen to be following from the premises, for any argument with conventional emotive metaphor, both convincingness and emotional appeal were significant predictors for committing the equivocation fallacy [Convincingness: CM+ ($t_{(94)} = 3.2$; $p < 0.01$), CM− ($t_{(94)} = 3.4$; $p < 0.05$); Emotional appeal: CM+ ($t_{(94)} = 2.7$; $p < 0.05$), CM− ($t_{(94)} = 1.96$; $p < 0.05$)]. Convincingness was also a significant predictor for committing the fallacy in the case of CM0 arguments ($t_{(94)} = 3.71$; $p < 0.05$). Emotional appeal was also a significant predictor for accepting the fallacious argument in the case of NM- arguments ($t_{(94)} = 2.01$; $p < 0.01$), as well as the perception of having found a logical relation between premises and conclusion ($t_{(94)} = 3.77$; $p < 0.05$) and some ambiguity in the argument ($t_{(94)} = 1.9$; $p < 0.01$). In the case of NM0 arguments, having found a logical relation between premises and conclusion ($t_{(94)} = 3.79$; $p < 0.01$) and real-world experience of similar arguments ($t_{(94)} = 2.7$; $p < 0.05$) were significant predictors for committing the fallacy. Interestingly, the belief in the conclusion was a significant predictor in the case of arguments with negatively valenced metaphors [CM− ($t_{(94)} = 1.91$; $p < 0.001$), NM- ($t_{(94)} = 2.7$; $p < 0.05$)].

When the conclusion was detected as not following from the premises, convincingness was a significant predictor in the case of arguments with negatively valenced metaphors [CM− ($t_{(94)} = 2.4$; $p < 0.01$), NM- ($t_{(94)} = 2.71$; $p < 0.05$)]. Logical relation was significant predictor for detecting the metaphoric fallacy in the case of CM+ ($t_{(94)} = 2.22$; $p < 0.001$), CM0 ($t_{(94)} = 2.41$; $p < 0.01$) and NM0 ($t_{(94)} = 2.4$; $p < 0.01$) arguments. Ambiguity and emotional appeal were instead significant predictor for detecting the metaphoric fallacy in the case of positively valenced novel metaphors [Ambiguity:

TABLE 10 | Mean (M) and standard deviation (SD) values of each predictor for each middle term condition (Experiment 2).

Predictor	CM+		CM-		CM0		NM+		NM-		NM0	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Understandability	2.91	0.77	2.52	0.78	3.48	0.76	2.78	0.74	3.08	0.78	2.77	0.81
Convincingness	2.42	0.64	2.04	0.68	3.14	0.72	2.25	0.61	2.76	0.62	2.27	0.76
Emotional appeal	1.92	0.75	1.51	0.57	2.29	0.79	1.95	0.72	1.85	0.70	1.49	0.47
Logical relation	2.86	0.70	2.53	0.76	3.23	0.72	2.61	0.69	2.98	0.71	2.69	0.74
Ambiguity	2.35	0.62	2.60	0.71	2.13	0.70	2.40	0.72	2.20	0.61	2.51	0.76
Belief in the conclusion	2.41	0.70	1.93	0.64	3.23	0.73	2.20	0.68	2.64	0.71	2.51	0.77
Real world experience	2.33	0.89	1.90	0.80	3.04	0.84	2.19	0.79	2.58	0.88	2.07	0.87

Positively valenced (CM+), negatively valenced (CM-), non-emotive (CM0) conventional metaphor. Positively valenced (NM+), negatively valenced (NM-), non-emotive (NM0) novel metaphor.

TABLE 11 | *t*/*p*-values for each predictor in the evaluation of arguments, comparing the middle term conditions (Experiment 2).

Answer	Predictors	Conventional metaphor						Novel metaphor					
		Positive		Negative		Non-emotive		Positive		Negative		Non-emotive	
YES		$\beta = 0.16$, $R^2 = 0.24$, $F^2 = 0.10$		$\beta = 0.44$, $R^2 = 0.21$, $F^2 = 0.33$		$\beta = 0.37$, $R^2 = 0.25$, $F^2 = 0.28$		$\beta = 0.11$, $R^2 = 0.07$, $F^2 = 0.23$		$\beta = 0.27$, $R^2 = 0.24$, $F^2 = 0.11$		$\beta = 0.27$, $R^2 = 0.35$, $F^2 = 0.33$	
		β	*t*	*β*	*t*	*β*	*t*	*β*	*t*	*β*	*t*	*β*	*t*
	Understandability	0.33	2.91*	0.34	2.29**	0.29	2.51*	0.55	3.15***	0.47	2.66**	0.57	2.97*
	Convincingness	0.29	3.2**	0.22	3.4*	0.33	3.71*	0.43	2.72**	0.55	3.61*	0.30	3.1
	Emotional appeal	0.22	2.7*	0.09	1.96*	0.21	2.66	0.20	2.37*	0.48	2.01**	0.08	1.6
	Logical relation	0.20	1.47	0.21	1.15	0.07	2.42	0.11	1.18	0.44	3.77*	0.31	3.79**
	Ambiguity	0.32	2.51	0.17	2.9	0.12	2.01	0.09	2.91	0.53	1.9**	0.04	1.46
	Belief in the conclusion	0.17	2.33	0.67	1.91***	0.11	3.07	0.07	2.12	0.40	2.7*	0.05	2.1
	Real world experience	0.11	1.56	0.22	1.41	0.20	2.4	0.13	2.11	0.21	1.9	0.24	2.7*
NO		$\beta = 0.19$, $R^2 = 0.04$, $F^2 = 0.34$		$\beta = 0.12$, $R^2 = 0.41$, $F^2 = 0.21$		$\beta = 0.13$, $R^2 = 0.17$, $F^2 = 0.11$		$\beta = 0.16$, $R^2 = 0.13$, $F^2 = 0.12$		$\beta = -0.07$, $R^2 = 0.11$, $F^2 = 0.17$		$\beta = 0.07$, $R^2 = 0.19$, $F^2 = 0.21$	
		β	*t*	*β*	*t*	*β*	*T*	*β*	*t*	*β*	*t*	*β*	*t*
	Understandability	0.24	1.4	0.36	2.74**	0.33	3.47***	0.76	3.04**	0.30	2.01**	0.41	2.97**
	Convincingness	0.09	2.72	0.29	2.4**	0.14	2.07	0.10	1.14	0.21	2.71*	0.11	3.41
	Emotional appeal	0.19	2.47	0.31	2.8	0.19	2.11	0.36	2.9*	0.06	1.13	0.15	3.07
	Logical relation	0.57	2.22***	0.09	1.61	0.41	2.41**	0.09	1.8	0.03	1.71	0.28	2.4**
	Ambiguity	0.22	1.91	0.11	2.11	0.22	3.53	0.28	2.17*	0.17	1.72	0.09	2.02
	Belief in the conclusion	0.19	1.71	0.27	2.34	0.05	1.17	0.07	2.44	0.29	2.7*	0.30	1.26
	Real world experience	0.07	1.01	0.07	2.02	0.37	1.21*	0.10	2.21	0.47	3.4	0.28	2.48*

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

$t_{(94)} = 2.17$; $p < 0.05$; Emotional appeal: $t_{(94)} = 2.9$; $p < 0.05$]. Real world experience of similar arguments was a significant predictor in the case of arguments with non-emotive metaphors [CM0: $t_{(94)} = 1.21$; $p < 0.05$, NM0: $t_{(94)} = 2.48$; $p < 0.05$] and negatively valenced novel metaphors ($t_{(94)} = 3.4$; $p < 0.01$).

Discussion

The results confirm that participants mostly fail in the evaluation of the metaphoric fallacy. However, they suggest that the affective valence of the overall metaphor, rather than the vehicle alone as middle term, influences the detection of the metaphoric fallacy. When the first premise is affectively incoherent and more difficult to make sense of, neither the metaphorical nor the affective framing significantly influence participants' evaluation. This suggests that the initial incoherence of the arguments make participants abandon the global heuristic process at work in both metaphorical and affective framing in arguments' assessment. As suggested in argumentation theory by Walton (1996), the metaphoric fallacy is a special kind of equivocation fallacy where a global process of metaphorical interpretation is in place,

differing from the mere process of disambiguation required by literal middle terms in the detection of the equivocation fallacy. A global interpretation would involve an evaluation of the overall first premise featuring the metaphor, including both the topic and the vehicle. Global aspects prevailed when participants committed the fallacy: the emotional appeal and the convincingness of the argument in the case of emotive metaphors and, as in the first experiment, the belief in the conclusion when assessing syllogisms with negatively valenced metaphors.

When it is more difficult to make sense of the *affective incoherence* of the metaphor in the first premise, the metaphorical and the affective framing interact, *increasing participants' logicality* not only in the case of novel positively valenced metaphors (compared to conventional non-emotive metaphors), but also in the case of conventional negatively valenced metaphors (compared to conventional non-emotive and novel negatively valenced metaphors). More correct answers are provided in those conditions, suggesting a more nuanced picture, where emotive metaphors do not have the usual deleterious role in reasoning. In the case of NM+ arguments, participants detected the ambiguity of the fallacy, and in the case of CM−, participants checked for the convincingness of the argument itself, without being affected by the believability of the conclusion (as in the case of NM- arguments). This suggests that the initial incoherence of the arguments makes participants more vigilant in specific and local cases where the type of metaphor and its affective valence did play a role in avoiding the negative consequences associated with endorsing unsound arguments. Thus, a more analytical style of reasoning is embraced by the participants: in the case of novel positively valenced metaphors, the evident distance between the literal and the metaphorical meaning is more easily detected; in the case of conventional negatively valenced metaphors, the intention to convince the participants make them more doubting and less likely to endorse the believable but unsound conclusion. This more systematic information processing style probably required the allocation of extra attention and more cognitive resources (Forgas, 1995), requiring to assess both the metaphorical and affective effects.

General Discussion

Both metaphors and emotions have been said to entail a negative effect on reasoning. The studies highlight how the picture is more complicated: not always the metaphorical framing interacts with the affective framing leading to a decreased logicality and not always a decreased logicality is justified by a deleterious use of the heuristics entailed by the metaphorical and the affective framing in reasoning. This could be explained by dual-processing models [see e.g., Kahneman (2003), Evans (2008), and Evans and Frankish (2009)], envisaging two distinct inferential (automatic vs. controlled) processes: notoriously, System 1 and System 2. The first includes quick associative, emotional and heuristic processes working in a parallel, effortless and unconscious way; the second includes slow rule-governed, neutral and content-blind processes working in a serial, effortful and often conscious way (Ervas et al., 2015). Metaphors can be seen as heuristic, System 1-type processes that never guarantee the preservation of truth in deductive reasoning tasks up to System 2 (Fischer, 2014; Keefer and Landau, 2016), though intuitively leading to genuinely new knowledge *via* a creative argumentation style (Schn, 1993; Leung et al., 2012). In this perspective, as pointed out by Haack (1994, p. 4), metaphors can be seen as "cognitively vital" and illuminating, but "can also be feeble; can be exploited to the purpose of persuading by emotional appeal rather than rational argument." As argued in recent psychological literature (Citron and Goldberg, 2014, p. 9), metaphors are "more emotionally engaging than literal expressions," especially when they are grounded in perceptual and sensorimotor representations (Indurkhya, 1992, 1994; Gibbs, 2006), recalling people's experiential knowledge rather than more abstract and analytical constructs. In the same vein, emotions have been said to elicit System 1-type processes, depleting available cognitive resources at the expense of System 2, thus affecting analytical reasoning and decreasing logicality to a greater extent (Channon and Baker, 1994; Kensinger and Corkin, 2003; De Neys, 2012).

However, the relationship between System 1 and System 2 processes is largely understood in terms of competition or conflict:embodiment plays a role only in heuristics and/or emotional processes, while "proper" reasoning processes still deserve a superior function of control and/or revision. From a philosophical point of view, dual-processing models have been criticized as proposing an anachronistic view of mind, having a disembodied normative System 2 working in opposition to an embodied System 1 (Marraffa, 2014). More recent "integrated" dual-processing models offer an alternative view of reasoning, based on the coordination rather than competition between different inferential processes [see e.g., Moshman (2004), Mercier and Sperber (2011), Carruthers (2011), Fletcher and Carruthers (2012), Baumard and Boyer (2013), and Rossi (2014)]. From a psychological point of view, it has been pointed out that emotional heuristic processes are not always in contrast with normatively-correct or analytical reasoning processes. A number of studies showed that positive emotional content improves performance on conditional and/or syllogistic reasoning (Isen et al., 1987; Melton, 1995). Especially when this content is *relevant* to the participants' prior real-life experience and/or their current emotional state (Blanchette, 2006; Johnson-Laird et al., 2006; Blanchette et al., 2007; Blanchette and Campbell, 2012; Caparos and Blanchette, 2016), there is a reduction of the classic belief bias effect on arguments' evaluation in the case of negative emotional content (Goel and Vartanian, 2011). Rather than contraposing emotional and reasoning processes, Blanchette et al. concluded that, depending on the *emotional relevance* to the evaluator, the same emotional content might have either an "incidental affect" (non-relevant) or a more "integral affect" (relevant) on the reasoning task, respectively enhancing the heuristic processes of System 1 and the analytical processes of System 2 [Blanchette and Richards, 2010; Huntsinger, 2013; Caparos and Blanchette, 2015; but see Jung et al. (2014) for different results].

The present study shows that another element, *affective coherence*, needs to be considered to have a more comprehensive (but also more nuanced) view of the interaction between emotional content and metaphorical (heuristic) processes and

rule-governed (normative) processes. The comparison between the main results of the first and the second experiments (**Table 3**) suggests that affective coherence is crucial to understand the main effects of the metaphor type and the affective type of the middle term on the evaluation of the arguments, and their interaction. When the first premises are *affectively coherent* both the metaphorical and the affective framing significantly influence the argument evaluation. However, when the first premises are *affectively incoherent*, neither the metaphorical nor the affective framing significantly influence the argument evaluation. The results suggest that the *affective coherence* of the vehicle and the topic is responsible for the emergence of both the framing effects in the syllogisms' evaluation. Different from relevance, affective coherence concerns the emotional content of the metaphor's vehicle/topic, rather than the relationship with the evaluators' emotional state. The results suggest that affective coherence is more basic and primitive than relevance, determining the range of possible metaphorical and affective framing effects and influencing the overall participants' attempt to make sense of the argument. Affective coherence has an impact on the early stages of semantic sentence processing, automatically influencing sense-making (Schauenburg et al., 2019). Other studies showed that affective coherence also leads to better recall of the content (Richards and Gross, 1999), serving "as evidence of the appropriateness of affective concepts that come to mind" (Centerbar et al., 2008, p. 560). The condition of affective coherence induces "affective certainty" (Tamir et al., 2002), which "allow participants to devote themselves fully to the task at hand" (Clore and Schnall, 2008, p. 3).

The results of the experiments also suggest that, for reasoning with emotive metaphors, affective coherence promotes more holistic, *global and heuristic processes* when both metaphorical framing and affective framing contribute to creatively make sense of the overall argument, diverging to strict normative rules of logic in the metaphoric fallacy evaluation. However, affective incoherence makes global processes of sense-making more difficult to be carried out, neutralizing both general metaphorical and affective framing effects and improving logicality in *local and analytical processes*, where the framing effects interact to get rid of ambiguity or the believability of the conclusion, usually deviating normative reasoning. The direct emotional impact comes first (Zajonc, 1980) and determines the attitudes toward what is coming next (Petty and Cacioppo, 1986b; Petty and Briñol, 2015) and the reasoning style of the evaluator: affective coherent representations make participants more certain and prone to read/listen to what may be coming next, while affective incoherent representations make participants more dubious and vigilant toward possible rule violations. Affective incoherence vs. coherence draw us to play two different games in reasoning with emotive metaphors, which could be called respectively the "doubting game" and the "believing game" (Elbow, 1998). The doubting game questions previous beliefs and preconceptions, especially when self-evident, as well as "systems of commonplaces" associated with metaphors, relying only on literal truth and necessary consequences. The believing game accepts believable conclusions to figure out new or revitalized meanings in the premises and make sense of the whole argument metaphorically (Elbow, 1998). By inhibiting the doubting game operating in rule-governed deduction, the believing game might lead to knowing something new or to see some previously unveiled alternatives. The games are not exclusive even though they cannot be played simultaneously, as some affective situations may be more appropriate than others to play the believing vs. the doubting game.

Depending on the affective coherence vs. incoherence between the concepts involved in metaphor understanding, the interplay between metaphorical and affective framing could be consistent with both the possibility that emotional stimuli activated concepts or images that required additional processing resources in favor of the heuristic processes of the System 1 (at the expenses of the System 2), and the possibility that the concepts activated by emotional stimuli overlap with the representations necessary for the resolution of the reasoning problem, thus favoring System 2 processes. Specifically, we can now answer the questions leading us till here:

Q1. Under what conditions does the double framing effect of emotive metaphors mostly influence the evaluation of the argument, leading to a fallacy of equivocation?

R1. In the condition of affective incoherence rather than affective coherence, the metaphorical and the affective framing mostly interact, influencing the evaluation of the argument, and increasing the ability to identify the fallacy of equivocation. Under affective uncertainty, participants are more prone to questioning the premises of the arguments and the believability of the conclusion, improving System 2-type analytical and local processes at the expense of System 1-type global processes. The double framing effect of emotive metaphors is most significant in reasoning with conventional negatively valenced metaphors and novel positively valenced metaphors, though for different reasons. For conventional negatively valenced metaphors, the arguments' level of convincingness led participants to discard them as fallacious. For positively valenced metaphors, arguments' emotional appeal and the ambiguity between the metaphorical and literal meanings of the premises led participants to identify the equivocation fallacy.

Q2. Under what conditions does the metaphorical framing influence the evaluation of an argument? Are participants more prone to commit a fallacy of equivocation in the case of conventional metaphors or in the case of novel metaphors?

R2. In the condition of affective coherence rather than affective incoherence, the metaphorical framing influences the overall evaluation of an argument, leading to making sense of the premises in light of the believed conclusion *via* a global, System 1-type heuristic process. Especially in the case of conventional metaphors, which are not consciously processed as metaphors and whose intuitive truth conditions are more difficult to question, participants

are more prone to disattend System 2-type reasoning, thus falling into the equivocation fallacy.

Q3. Under what conditions does the affective framing influence the evaluation of the argument? Are participants more prone to commit a fallacy of equivocation in the case of negatively valenced metaphors or in the case of positively valenced metaphors?

R3. In the condition of affective coherence rather than affective incoherence, the affective framing influences the overall evaluation of an argument *via* System 1-type processes, i.e., leveraging on the convincingness and emotional appeal of the arguments featuring emotive metaphors. Especially in the case of negatively valenced metaphors, participants are more prone to committing the equivocation fallacy, accepting the believable conclusion and finding a premise/conclusion connection alternative to the corrected System 2-type logical connection.

CONCLUSIONS

Far from just leading to fallacies of reasoning, both metaphors and emotions entail a framing effect that makes us see things under a new perspective, influencing our decisions and evaluations in many, and sometimes unexpected, ways. Rather than always contrasting normative reasoning, the double framing effect can promote creative heuristic processes and increase logicality. Emotive metaphors do entail a double framing effect in argumentation, coming from their being metaphors and having an emotive (evaluative) connotation, as hypothesized (H1), but they do not lead to equivocation fallacies as expected. Depending on the contextual affective coherence and incoherence of the first premise, respectively, the double framing effect led participants to globally interpret the argument as metaphoric, or locally detect the ambiguity which leads to the equivocation fallacy. The metaphorical framing effect is stronger for conventional metaphors and the affective framing effect for negatively valenced metaphors, as hypothesized (respectively in H2 and H3), but they are both significant only when there is a more basic affective coherence between the vehicle and the topic of the metaphor, which also makes it possible to make more sense of the argument as a whole.

The experiments were limited in only considering coherence with respect to the affective valence of metaphors. Further research should consider also the affective coherence with respect to arousal and intensity (Clore and Schnall, 2005), which could provide information about other dimensions of the affective as well as metaphorical framing effect in reasoning with emotive metaphors. More ecologically valid studies could be designed to include the relevance of emotive metaphors to participants' emotional states and/or specific relevant properties of the metaphor that could be crucially extended in the second premises, thus further promoting a holistic interpretation of the metaphoric fallacy. Further research is also needed to understand whether, as Walton (1996) suggested, emotive metaphors are not as responsible for ambiguity fallacies as for other clarity fallacies. Finally, the double framing effect of emotive metaphors might be further investigated in fallacious arguments which explicitly appeal to emotions, to check whether conventional negative metaphors and novel positive metaphors still help to detect persuasive but fallacious arguments. This might also help us to better understand when people play the believing game or the doubting game (Elbow, 1998) in reasoning with emotive metaphors.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: https://osf.io/jzpva/?view_only=8edc3b523cbb4afba4bd71978d847a48.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the ethics committee of the Department of Education, Psychology, Philosophy, University of Cagliari. The participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

FE, MR, AO, and BI: design. FE: construction of the materials and manuscript writing. FE, MR, and AO: data collection. FE and AO: data analysis (respectively analysis of variance and regression analysis). All authors provided feedback on the draft and approved the final version of the manuscript.

FUNDING

FE gratefully acknowledges the support of the University of Cagliari (Premialità 2019, n. 246/19C, 28.11.2019 and RICCAR_FIR_2020), the Sardinia Regional Government (research grant RAS FSC 2017, n. F76C18001040002), and Fondazione Banco di Sardegna (research grant n. F72F20000420007). MR gratefully acknowledges the support of the Fundação para a Ciência e a Tecnologia (research grants n. SFRH/BPD/115073/2016 and PTDC/FER-FIL/28278/2017).

ACKNOWLEDGMENTS

We thank Valentina Favrin, Sonia Lisco, Francesco Paoli, Lucrezia Pellizzon, Antonio Pierro as well as Céline Boisserie-Lacroix, Jérôme Dokic, Stéphane Lemaire, Christine Tappolet, and all the participants to the Workshop Perspectives on Affective Evaluation (Institut Jean Nicod, Ecole Normale Supérieure, March 25, 2019) for all the helpful comments and insights on this work.

REFERENCES

Angie, A. D., Connelly, S., Waples, E. P., and Kligyte, V. (2011). The influence of discrete emotions on judgement and decision-making: a meta-analytic review. *Cogn. Emot.* 25, 1393–1422. doi: 10.1080/02699931.2010.550751

Ariely, D. (2009). *Predictably Irrational*. New York, NY: Harper Collins Publishers.

Ball, L. G., and Thompson, V. A. (2018). "Belief bias and reasoning," in *International Handbook of Thinking and Reasoning,* eds L. G. Ball, and V.A. Thompson (New York, NY: Routledge), 16–36.

Ball, L. J., Phillips, P., Wade, C. N., and Quayle, J. D. (2006). Effects of belief and logic on syllogistic reasoning: eye-movement evidence for selective processing models. *Exp. Psychol.* 53, 77–86. doi: 10.1027/1618-3169.53.1.77

Bambini, V., Resta, D., and Grimaldi, M. (2014). A dataset of metaphors from the Italian literature: exploring psycholinguistic variables and the role of context. *PLoS ONE* 9:e105634. doi: 10.1371/journal.pone.0105634

Barrett, L. F., Mesquita, B., Ochsner, K. N., and Gross, J. J. (2007). The experience of emotion. *Annu. Rev. Psychol.* 58, 373–403. doi: 10.1146/annurev.psych.58.110405.085709

Barth, E. M. (1974). "Argument by analogy," in *The Logic of the Articles in Traditional Philosophy* (Springer: Dordrecht), 291–336. doi: 10.1007/978-94-010-9866-3

Baumard, N., and Boyer, P. (2013). Religious beliefs as reflective elaborations on intuitions: a modified dual- process model. *Curr. Direct. Psychol. Sci.* 22, 295–300. doi: 10.1177/0963721413478610

Beardsley, M. C. (1957). *Thinking Straight*. Englewood Cliffs, NJ: Prentice-Hall.

Black, M. (1954). Metaphor. *Proc. Aristotelian Soc.* 55, 273–294. doi: 10.1093/aristotelian/55.1.273

Black, M. (1962). *Models and Metaphors*. Ithaca, NY: Cornell University Press. doi: 10.7591/9781501741326

Blackburn, S. (1984). *Spreading the Word: Groundings in the Philosophy of Language*. Oxford: Clarendon Press.

Blanchette, I. (2006). The effect of emotion on interpretation and logic in a conditional reasoning task. *Memory Cogn.* 34, 1112–1125. doi: 10.3758/BF03193257

Blanchette, I. (2014). (eds.). *Emotion and Reasoning*. London; New York, NY: Psychology Press. doi: 10.4324/9781315888538

Blanchette, I., and Campbell, M. (2012). Reasoning about highly emotional topics: syllogistic reasoning in a group of war veterans. *J. Cogn. Psychol.* 24, 157–164. doi: 10.1080/20445911.2011.603693

Blanchette, I., and Caparos, S. (2013). When emotions improve reasoning: the possible roles of relevance and utility. *Thinking Reason.* 19, 399–413. doi: 10.1080/13546783.2013.791642

Blanchette, I., Caparos, S., and Trémolière, B. (2018). "Emotion and reasoning," in *International Handbook of Thinking and Reasoning*, eds L. G. Ball and V. A. Thompson (New York, NY: Routledge), 57–70.

Blanchette, I., and Leese, J. (2011). Physiological arousal and logicality: the effect of emotion on conditional reasoning. *Exp. Psychol.* 58, 235–246. doi: 10.1027/1618-3169/a000090

Blanchette, I., and Richards, A. (2004). Reasoning about emotional and neutral materials. Is logic affected by emotion? *Psychol. Sci.* 15, 745–752. doi: 10.1111/j.0956-7976.2004.00751.x

Blanchette, I., and Richards, A. (2010). The influence of affects on higher level cognition: a review of research on interpretation, judgement, decision making and reasoning. *Cogn. Emotion* 24, 561–595. doi: 10.1080/02699930903132496

Blanchette, I., Richards, A., Melnyk, L., and Lavda, A. (2007). Reasoning about emotional contents following shocking terrorist attacks: a tale of three cities. *J. Exp. Psychol. Appl.* 13, 47–56. doi: 10.1037/1076-898X.13.1.47

Bless, H., Clore, G. L., Schwarz, N., Golisano, V., Rabe, C., and Wölk, M. (1996). Mood and the use of scripts: does being in a happy mood really lead to mindlessness? *J. Personal. Soc. Psychol.* 71, 665–679. doi: 10.1037/0022-3514.71.4.665

Borelli, E., and Cacciari, C. (2019). The comprehension of metaphorical descriptions conveying gender stereotypes. An exploratory study. *Front. Psychol.* 10:2615. doi: 10.3389/fpsyg.2019.02615

Bradley, M. M., and Lang, P. J. (1999). *Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings (30, 1, 25–36). Technical Report C-1, the Center for Research in Psychophysiology*. Gainesville, FL: University of Florida.

Burgers, C., Konijn, E. A., and Steen, G. J. (2016). Figurative framing: shaping public discourse through metaphor, hyperbole, and irony. *Commun. Theory* 26, 410–430. doi: 10.1111/comt.12096

Caparos, S., and Blanchette, I. (2015). Affect et pense logique: comment les motions influencent notre raisonnement. *Revue Qubecoise de Psychologie*, 36, 57–70.

Caparos, S., and Blanchette, I. (2016). Independent effects of relevance and arousal on deductive reasoning. *Cogn. Emotion* 31, 1012–1022. doi: 10.1080/02699931.2016.1179173

Carruthers, P. (2011). *The Opacity of Mind: An Integrative Theory of Self-Knowledge*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780199596195.001.0001

Carston, R. (1997). Enrichment and loosening: complementary processes in deriving the proposition expressed? *Linguistische Berichte* 8, 103–127. doi: 10.1007/978-3-663-11116-0_7

Carston, R. (2002). *Thoughts and Utterances: The Pragmatics of Explicit Communication*. Oxford: Blackwell. doi: 10.1002/9780470754603

Carston, R. (2010). Metaphor: *ad hoc* concepts, literal meaning and mental images. *Proc. Aristotelian Soc.* 110, 295–321. doi: 10.1111/j.1467-9264.2010.00288.x

Carston, R. (2018). Figurative language, mental imagery and pragmatics. *Metaphor. Symbol* 33, 198–217. doi: 10.1080/10926488.2018.1481257

Caruana, F., and Cuccio, V. (2017). Overcoming the acting/reasoning dualism in intelligent behavior. *Phenom. Cogn. Sci.* 16, 709–713. doi: 10.1007/s11097-016-9471-1

Cassotti, M., Habib, M., Poirel, N., Aïte, A., Houdé, O., and Moutier, S. (2012). Positive emotional context eliminates the framing effect in decision-making. *Emotion* 12:926. doi: 10.1037/a0026788

Cavazzana, A., and Bolognesi, M. (2020). Uncanny resemblance: words, pictures, and conceptual representations in the field of metaphor. *Cogn. Linguist. Stud.* 7, 31–57. doi: 10.1075/cogls.00048.cav

Centerbar, D. B., Schnall, S., Clore, G. L., and Garvin, E. D. (2008). Affective incoherence: when affective concepts and embodied reactions clash. *J. Personal. Soc. Psychol.* 94, 560–578. doi: 10.1037/0022-3514.94.4.560

Channon, S., and Baker, J. (1994). Reasoning strategies in depression: effects of depressed mood on a syllogism task. *Personal. Individ. Differ.* 17, 707–711. doi: 10.1016/0191-8869(94)90148-1

Citron, F. M. M. (2012). Neural correlates of written emotion word processing: a review of recent electrophysiological and hemodynamic neuroimaging studies. *Brain Lang.* 122, 211–226. doi: 10.1016/j.bandl.2011.12.007

Citron, F. M. M., and Goldberg, A. E. (2014). Metaphorical sentences are more emotionally engaging than their literal counterparts. *J. Cogn. Neurosci.* 26, 2585–2595. doi: 10.1162/jocn_a_00654

Citron, F. M. M., Gray, M. A., Critchley, H. D., Weekes, B. S., and Ferstl, E. C. (2014). Emotional valence and arousal affect reading in an interactive way: neuroimaging evidence for an approach-withdrawal framework. *Neuropsychologia* 56, 79–89. doi: 10.1016/j.neuropsychologia.2014.01.002

Citron, F. M. M., Weekes, B. S., and Ferstl, E. C. (2013). Effects of valence and arousal on written word recognition: time course and ERP correlates. *Neurosci. Lett.* 533, 90–95. doi: 10.1016/j.neulet.2012.10.054

Clark, S. R. L. (1994). The possible truth of metaphor. *Int. J. Philos. Stud.* 2, 19–30. doi: 10.1080/09672559408570781

Clore, G. L., and Huntsinger, J. R. (2007). How emotions inform judgment and regulate thought. *Trends Cogn. Sci.* 11, 393–399. doi: 10.1016/j.tics.2007.08.005

Clore, G. L., and Schnall, S. (2005). "The influence of affect on attitude," in *The Handbook of Attitudes,* eds D. Albarracin, B. T. Johnson, and M. P. Zanna (Mahwah, NJ: Erlbaum), 437–489.

Clore, G. L., and Schnall, S. (2008). "Affective coherence: affect as embodied evidence in attitude, advertising and art," in *Embodied Grounding: Social, Cognitive, Affective, and Neuroscientific Approaches,* eds G. R. Semin and E. R. Smith (Cambridge: Cambridge University Press), 211–236. doi: 10.1017/CBO9780511805837.010

Copi, I. M., Cohen, C., and McMahon, K. (2014). *Introduction to Logic*. Harlow: Pearson.

Correia, V. (2011). Biases and fallacies: the role of motivated irrationality in fallacious reasoning. *Cogency J. Reason. Argument.* 3, 107–126.

Damasio, A. (1994). *Descartes' Error: Emotion, Reason, and the Human Brain*. New York, NY: G. P. Putnam's Sons.

D'Arms, J., and Jacobson, D. (2003). The significance of recalcitrant emotions; or anti-quasi judgmentalism. *Philos. Trans. R Soc.* 52, 127–146. doi: 10.1017/CBO9780511550270.009

De Mauro, T. (2000). *GRADIT: Grande Dizionario Italiano Dell'uso*. Torino: UTET.

De Neys, W. (2012). Bias and conflict: a case for logical intuitions. *Perspectiv. Psychol. Sci.* 7, 28–38. doi: 10.1177/1745691611429354

DeLancey, C. (2002). *Passionate Engines. What Emotions reveals about Mind and Artificial Intelligence*. New York: Oxford University Press. doi: 10.1093/0195142713.001.0001

DeLancey, C. (2007). "Emotion and Cognition: a new map of the terrain," in *Cartographies of the Mind. Philosophy and Psychology in Intersection,* eds M. Marraffa, M. De Caro, and F. Ferretti (Berlin: Springer), 93–103. doi: 10.1007/1-4020-5444-0_7

Dijksterhuis, A. (2004). Think different: the merits of unconscious thought in preference development and decision making. *J. Pers. Soc. Psychol.* 87, 586–598. doi: 10.1037/0022-3514.87.5.586

Dijksterhuis, A., Bos, M. W., Nordgren, L. F., and van Baaren, R. B. (2006). On making the right choice: the deliberation-without-attention effect. *Science* 311, 1005–1007. doi: 10.1126/science.1121629

Dijksterhuis, A., Bos, M. W., Van Der Leij, A., and Van Baaren, R. B. (2009). Predicting soccer matches after unconscious and conscious thought as a function of expertise. *Psychol. Sci.* 20, 1381–1387. doi: 10.1111/j.1467-9280.2009.02451.x

Elbow, P. (1998). *The Doubting Game and the Believing Game. An Analysis of the Intellectual Enterprise. In Writing Without Teachers*. New York, NY: Oxford University Press.

Eliades, M., Mansell, W., and Blanchette, I. (2013). The effect of emotion on statistical reasoning: findings from a base rates task. *J. Cogn. Psychol.* 25, 277–282. doi: 10.1080/20445911.2012.761632

Eliades, M., Mansell, W., Stewart, A. J., and Blanchette, I. (2012). An investigation of belief-bias and logicality in reasoning with emotional contents. *Thinking Reason.* 18, 461–479. doi: 10.1080/13546783.2012.713317

Elster, J. (1999). *Alchemies of the Mind: Rationality and the Emotions*. New York, NY: Cambridge University Press. doi: 10.1017/CBO9781139173308

Entman, R. M. (1993). Framing: toward clarification of a fractured paradigm. *J. Commun.* 43, 51–58. doi: 10.1111/j.1460-2466.1993.tb01304.x

Ervas, F. (2015). (Becoming) experts in meaning ambiguities. *Humana. Mente. J. Philos. Stud.* 8, 225–243.

Ervas, F. (2017). Another metaphor is possible. Challenging social stereotypes in figurative language comprehension. *Reti Saperi Linguaggi* 4, 79–96. doi: 10.12832/87357

Ervas, F. (2019). Metaphor, ignorance and the sentiment of (ir) rationality. *Synthese* 19, 1–25. doi: 10.1007/s11229-019-02489-y

Ervas, F. (2020). "How nice does it sound? An argumentative approach to the affective aspects of irony," in *Producing Figurative Expression: Theoretical, Experimental and Practical Perspectives,* eds J. Barnden, A. Gargett (Amsterdam: John Benjamins), 175–210. doi: 10.1075/ftl.10.07erv

Ervas, F., Gola, E., and Rossi, M. G. (2015). Metaphors and emotions as framing strategies in argumentation. *Proc. EuroAsianPacific Joint Conf. Cogn. Sci.* 1419, 645–650.

Ervas, F., Gola, E., and Rossi, M. G. (2018). "Argumentation as a bridge between metaphor and reasoning," in *Argumentation and Language. Linguistic, Cognitive and Discursive Explorations,* eds S. Oswald, T. Herman, and J. Jacquin (Berlin: Springer), 153–170. doi: 10.1007/978-3-319-73972-4_7

Evans, J. (2008). Dual-processing accounts of reasoning, judgment and social cognition. *Ann. Rev. Psychol.* 59, 255–278. doi: 10.1146/annurev.psych.59.103006.093629

Evans, J., Barston, J. L., and Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Mem. Cogn.* 11, 295–306. doi: 10.3758/BF03196976

Evans, J., and Frankish, K. (2009). *In Two Minds. Dual Processes and Beyond*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780199230167.001.0001

Fainsilber, L., and Ortony, A. (1987). Metaphorical uses of language in the expression of emotions. *Metaphor Symb. Act.* 2, 239–250. doi: 10.1207/s15327868ms0204_2

Fischer, E. (2014). Philosophical intuitions, heuristics, and metaphors. *Synthese* 191, 569–606. doi: 10.1007/s11229-013-0292-2

Fischer, E. (2015). Mind the metaphor! A systematic fallacy in analogical reasoning. *Analysis* 75, 67–77. doi: 10.1093/analys/anu124

Fletcher, L., and Carruthers, P. (2012). Metacognition and reasoning. *Philos. Trans. R Soc. Lond. B Biol. Sci.* 367, 1366–1378. doi: 10.1098/rstb.2011.0413

Forgas, J. P. (1995). Mood and judgment: the affect infusion model. *Psychol. Bulletin* 117, 39–66. doi: 10.1037/0033-2909.117.1.39

Fredrickson, B. L., and Branigan, C. (2005). Positive emotions broaden the scope of attention and thought-action repertoires. *Cogn. Emotion* 19, 313–332. doi: 10.1080/02699930441000238

Gernsbacher, M. A., and Faust, M. (1991). "The role of suppression in sentence comprehension," in *Understanding Word and Sentence,* ed G. B. Simpson (Amsterdam: Elsevier), 97–128. doi: 10.1016/S0166-4115(08)61531-9

Gernsbacher, M. A., Keysar, B., Robertson, R. R. W., and Werner, N. K. (2001). The role of suppression and enhancement in understanding metaphors. *J. Mem. Lang.* 45, 433–450. doi: 10.1006/jmla.2000.2782

Gibbs, R. W., Leggitt, J. S., and Turner, E. A. (2002). "What's special about figurative language in emotional communication," in *The Verbal Communication of Emotions: Interdisciplinary Perspectives,* ed S. R. Fussell (Mahwah, NJ: Lawrence Erlbaum Associates), 125–149.

Gibbs, R. W. Jr. (2006). Metaphor interpretation as embodied simulation. *Mind Language* 21, 434–458. doi: 10.1111/j.1468-0017.2006.00285.x

Giora, R. (2003). *On Our Mind: Salience, Context and Figurative Language*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780195136166.001.0001

Glucksberg, S. (2001). *Understanding Figurative Language: From Metaphors to Idioms*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780195111095.001.0001

Glucksberg, S. (2003). The psycholinguistics of metaphor. *Trends Cogn. Sci.* 7, 92–96. doi: 10.1016/S1364-6613(02)00040-2

Glucksberg, S., Gildea, P., and Bookin, H. B. (1982). On understanding nonliteral speech: can people ignore metaphors? *J. Verbal Learning Verbal Behav.* 21, 85–98. doi: 10.1016/S0022-5371(82)90467-4

Godden, D. (2015). Argumentation, rationality, and psychology of reasoning. *Inf. Logic* 35, 135–166. doi: 10.22329/il.v35i2.4124

Goel, V., and Vartanian, O. (2011). Negative emotions can attenuate the influence of beliefs on logical reasoning. *Cogn. Emotion* 25, 121–131. doi: 10.1080/02699931003593942

Greenspan, P. S. (1988). *Emotions and Reasons: An Inquiry Into Emotional Justification*. New York, NY: Routledge, Chapman and Hall.

Greenspan, P. S. (1992). Subjective guilt and responsibility. *Mind* 101, 287–303. doi: 10.1093/mind/101.402.287

Greenspan, P. S. (2004). "Emotions, rationality, and mind-body," in *Thinking About Feeling: Contemporary Philosophers on Emotions,* ed R. C. Solomon (New York, NY: Oxford University Press), 125–134.

Grice, P. (1975). "Logic and conversation," in *Syntax and Semantics 3: Speech Acts,* eds P. Cole and J. Morgan (New York, NY: Academic Press), 41–58. doi: 10.1163/9789004368811_003

Grice, P. (1989). *Studies in the Ways of Words*. Cambridge, MA: Harvard University Press.

Gross, J. J. (2008). "Emotion regulation," in *Handbook of Emotions*, eds M. Lewis, J. M. Haviland-Jones, and L. F. Barrett (New York, NY: The Guilford Press), 497–512.

Gross, K., and D'Ambrosio, L. (2004). Framing emotional response. *Polit. Psychol.* 25, 1–29. doi: 10.1111/j.1467-9221.2004.00354.x

Haack, S. (1994). "'Dry truth and real knowledge"': epistemologies of metaphor and metaphors of epistemology," in *Aspects of Metaphor. Synthese Library, vol 238*, ed J. Hintikka (Dordrecht: Springer), 1–22.

Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychol. Rev.* 108, 814–834. doi: 10.1037/0033-295X.108.4.814

Haidt, J. (2007). The new synthesis in moral psychology. *Science* 316, 998–1002. doi: 10.1126/science.1137651

Haidt, J., Bjorklund, F., and Murphy, S. (2000). Moral dumbfounding: when intuition finds no reason. *Lund Psychol. Rep.* 1, 191–221.

Hamblin, C. L. (1970). *Fallacies*. London: Methuen.

Hesse, M. (1963). *Models and Analogies in Science*. London: Sheed and Ward.

Hesse, M. (1965). Aristotle's logic of analogy. *Philos. Q.* 15, 328–340. doi: 10.2307/2218258

Hinton, M. (2019). Language and argument: a review of the field. *Res. Lang.* 17, 93–103. doi: 10.2478/rela-2019-0007

Hoffman, R. R. (1980). "Metaphor in science," in *The Psycholinguistics of Figurative Language,* eds R. P. Honeck and R. R. Hoffman (Hillsdale, NJ: Erlbaum), 393–423. doi: 10.4324/9780429432866-16

Hofmann, S. G., Heering, S., Sawyer, A. T., and Asnaani, A. (2009). How to handle anxiety: the effects of reappraisal, acceptance, and suppression strategies on anxious arousal. *Behav. Res. Therapy* 47, 389–394. doi: 10.1016/j.brat.2009.02.010

Hofstadter, D. (1995). *Fluid Concepts and Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought*. New York, NY: Basic Books.

Huntsinger, J. R. (2013). Incidental experiences of affective coherence and incoherence influence persuasion. *Pers. Soc. Psychol. Bull.* 39, 792–802. doi: 10.1177/0146167213482588

Indurkhya, B. (1992). *Metaphor and Cognition.* Dordrecht: Kluwer Academic Publishers. doi: 10.1007/978-94-017-2252-0

Indurkhya, B. (1994). "Metaphor as change of representation: an interaction theory of cognition and metaphor," in *Aspects of Metaphor. Synthese Library*, ed J. Hintikka (Dordrecht: Springer), 95–150.

Indurkhya, B. (2006). Emergent representations, interaction theory and the cognitive force of metaphor. *New Ideas Psychol.* 24, 133–162. doi: 10.1016/j.newideapsych.2006.07.004

Indurkhya, B. (2007). Rationality and reasoning with metaphors. *New Ideas Psychol.* 25, 16–36. doi: 10.1016/j.newideapsych.2006.10.006

Indurkhya, B. (2010). "On the role of metaphor in creative cognition," in *Proceedings of the International Conference on Computational Creativity: ICCC-X,* eds D. Ventura, A. Pease, R. Perez y Perez, G. Ritchie, and T. Veale (Department of Informatics Engineering, University of Coimbra: Coimbra), 51–59.

Indurkhya, B. (2016). "Towards a model of metaphorical understanding," in *Metaphor and Communication,* eds E. Gola and F. Ervas (Amsterdam: John Benjamins), 123–146. doi: 10.1075/milcc.5.07ind

Isen, A. M., Daubman, K. A., and Nowicki, G. P. (1987). Positive affect facilitates creative problem solving. *J. Pers. Soc. Psychol.* 52, 1122–1131. doi: 10.1037/0022-3514.52.6.1122

Johnson-Laird, P. N., Mancini, F., and Gangemi, A. (2006). A hyper-emotion theory of psychological illnesses. *Psychol. Rev.* 113, 822–841. doi: 10.1037/0033-295X.113.4.822

Jung, N., Wranke, C., Hamburger, K., and Knauff, M. (2014). How emotions affect logical reasoning: evidence from experiments with mood-manipulated participants, spider phobics, and people with exam anxiety. *Front. Psychol.* 5:570. doi: 10.3389/fpsyg.2014.00570

Kahneman, D. (2003). A perspective on judgment and choice: mapping bounded rationality. *Am. Psychol.* 58, 697–720. doi: 10.1037/0003-066X.58.9.697

Keefer, L. A., and Landau, M. J. (2016). Metaphor and analogy in everyday problem solving. *Wiley Interdiscip. Rev. Cogn. Sci.* 7, 394–405. doi: 10.1002/wcs.1407

Kensinger, E. A., and Corkin, S. (2003). Effect of negative emotional content on working memory and long-term memory. *Emotion* 3, 378–393. doi: 10.1037/1528-3542.3.4.378

Kissler, J., Assadollahi, R., and Herbert, C. (2006). Emotional and semantic networks in visual word processing: insights from ERP studies. *Prog. Brain Res.* 156, 147–183. doi: 10.1016/S0079-6123(06)56008-X

Kousta, S.-T., Vinson, D. P., and Vigliocco, G. (2009). Emotion words, regardless of polarity, have a processing advantage over neutral words. *Cognition* 112, 473–481. doi: 10.1016/j.cognition.2009.06.007

Kövecses, Z. (2000). *Metaphor and Emotions*. Cambridge: Cambridge University Press.

Kövecses, Z. (2005). *Metaphor in Culture: Universality and Variation*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511614408

Lakoff, G., and Johnson, M. (1980). *Metaphors We Live By*. Chicago, IL: University of Chicago Press.

Larsen, R. J., Mercer, K. A., and Balota, D. A. (2006). Lexical characteristics of words used in emotional Stroop experiments. *Emotion* 6, 62–72. doi: 10.1037/1528-3542.6.1.62

Lecheler, S., Schuck, A. R. T., and de Vreese, C. H. (2013). Dealing with feelings: positive and negative discrete emotions as mediators of news framing effects. *Communications* 38, 189–209. doi: 10.1515/commun-2013-0011

LeDoux, J. E. (1996). *The Emotional Brain. The Mysterious Underpinnings of Emotional Life*. New York, NY: Simon and Schuster.

Lefford, A. (1946). The influence of emotional subject matter on logical reasoning. *J. General Psychol.* 34, 127–151. doi: 10.1080/00221309.1946.10544530

Leung, A. K., Kim, S., Polman, E., Ong, L., Qiu, L., Goncalo, J. A., et al. (2012). Embodied metaphors and creative "acts". *Psychol. Sci.* 23, 502–509. doi: 10.1177/0956797611429801

Lieberman, H. R., Bathalon, G. P., Falco, C. M., Morgan, C. A., Niro, P. J., and Tharion,W.J. (2005).The fog of war: decrements in cognitive performance and mood associated with combat-like stress. *Aviat. Space Environ. Med.* 76, 7–14.

Lightbody, B. A., and Berman, M. P. (2010). The metaphoric fallacy to a deductive inference. *Informal Logic* 30, 185–193. doi: 10.22329/il.v30i2.1192

Maasen, S. (2000). "Metaphors in the social sciences: making use and making sense of them," in *Metaphor and Analogy in the Sciences,* ed F. Hallyn (Dordrecht: Springer), 99–244.

Macagno, F. (2020). How can metaphors communicate arguments? *Intercult. Pragmat.* 17, 335–363. doi: 10.1515/ip-2020-3004

Macagno, F., and Rossi, M. G. (2021). Inferential patterns of emotive meaning. *Perspect. Pragmatics Philos. Psychol.* 28, 83–110. doi: 10.1007/978-3-030-56696-8_5

Macagno, F., and Walton, D. (2009). Argument from analogy in law, the classical tradition, and recent theories. *Philos. Rhetor.* 42, 154–182. doi: 10.1353/par.0.0034

Macagno, F., and Walton, D. (2010). The argumentative uses of emotive language. *Rev. Iberoam. Argument.* 1, 1–33.

Macagno, F., and Walton, D. (2014). *Emotive Language in Argumentation.* Cambridge: Cambridge University Press. doi: 10.1017/CBO9781139565776

Macagno, F., and Zavatta, B. (2014). Reconstructing metaphorical meaning. *Argumentation* 28, 453–488. doi: 10.1007/s10503-014-9329-z

MacKuen, M., Wolak, J., Keele, L., and Marcus, G. E. (2010). Civic engagements: resolute partisanship or reflective deliberation. *Am. J. Pol. Sci.* 54, 440–458. doi: 10.1111/j.1540-5907.2010.00440.x

Maiese, M. (2014). How can emotions be both cognitive and bodily? *Phenomenol. Cogn. Sci.* 13, 513–531. doi: 10.1007/s11097-014-9373-z

Maiese, M. (2015). Thought insertion as a disownership symptom. *Phenomenol. Cogn. Sci.* 14, 911–927. doi: 10.1007/s11097-014-9387-6

Marcus, G. E. (2002). *The Sentimental Citizen. Emotion in Democratic Politics*. Pennsylvania, PA: The Pennsylvania State University Press.

Marcus, G. E., Neuman, W. R., and Mackuen, M. (2000). *Affective Intelligence and Political Judgment*. Chicago/London: The University of Chicago Press.

Marraffa, M. (2014). Emozioni e razionalit: oltre il modello ≪antagonistico≫. *Sistemi Intelligenti* 1, 149–160. doi: 10.1422/77223

Melton, R. J. (1995). The role of positive affect in syllogism performance. *Pers. Soc. Psychol. Bull.* 21, 788–794. doi: 10.1177/0146167295218001

Mercier, H., and Sperber, D. (2011). Why do Humans reason? Arguments for an argumentative theory. *Behav. Brian Sci.* 34, 57–111. doi: 10.1017/S0140525X10000968

Moshman, D. (2004). From inference to reasoning: the construction of rationality. *Thinking Reason.* 10, 221–239. doi: 10.1080/13546780442000024

Oakhill, J., and Garnham, A. (1993). On theories of belief bias in syllogistic reasoning. *Cognition* 46, 87–92. doi: 10.1016/0010-0277(93)90023-O

Oakhill, J., Johnson-Laird, P. N., and Garnham, A. (1989). Believability and syllogistic reasoning. *Cognition* 31, 117–140. doi: 10.1016/0010-0277(89)90020-6

Oaksford, M., Morris, F., Grainger, B., and Williams, J. M. G. (1996). Mood, reasoning, and central executive processes. *J. Exp. Psychol. Learn. Mem. Cogn.* 22, 476–492. doi: 10.1037/0278-7393.22.2.476

Oswald, S., Greco, S., Miecznikowski, J., Pollaroli, C., and Rocci, A. (2020). Argumentation and meaning. Semantic and pragmatic reflexions. *J. Argument. Context* 9:5. doi: 10.1075/jaic.00005.osw

Oswald, S., Herman, T., and Jacquin, J. (2018). *Argumentation and Language – Linguistic, Cognitive and Discursive Explorations*. Cham: Springer. doi: 10.1007/978-3-319-73972-4

Oswald, S., and Rihs, A. (2014). Metaphor as argument: rhetorical and epistemic advantages of extended metaphors. *Argumentation* 28, 133–159. doi: 10.1007/s10503-013-9304-0

Perelman, C., and Olbrechts-Tyteca, L. (1969). *The New Rhetoric: A Treatise on Argumentation*. Notre Dame: University of Notre Dame Press.

Petty, R., and Briñol, P. (2015). Emotion and persuasion: cognitive and meta-cognitive processes impact attitudes. *Cogn. Emotion* 29, 1–26, doi: 10.1080/02699931.2014.967183

Petty, R., and Cacioppo, J. (1986a). The elaboration likelihood model of persuasion. *Adv. Exp. Soc. Psychol.* 19, 123–205. doi: 10.1016/S0065-2601(08)60214-2

Petty, R., and Cacioppo, J. (1986b). *Communication and Persuasion: Central and Peripheral Routes to Attitude Change*. New York, NY: Springer-Verlag.

Pollaroli, C., Greco, S., Oswald, S., Miecznikowski, J., and Rocci, A. (2019). Rhetoric and Language: emotions and style in argumentative discourse. *Informal Logic* 39:6047. doi: 10.22329/il.v39i4.6047

Prinz, J. J. (2004). *Gut Reactions: A Perceptual Theory of Emotion*. New York, NY: Oxford University Press.

Recanati, F. (2004). *Literal Meaning*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511615382

Recanati, F. (2010). *Truth-Conditional Pragmatics*. *Oxford: Oxford University Press*. doi: 10.1093/acprof:oso/9780199226993.001.0001

Richards, J. M., and Gross, J. L. (1999). Composure at any cost? The cognitive consequences of emotion suppression. *Personal. Soc. Psychol. Bullet.* 8, 1033–1044. doi: 10.1177/01461672992511010

Robins, S., and Mayer, R. E. (2000). The metaphor framing effect: metaphorical reasoning about text-based dilemmas. *Discourse Processes* 30, 57–86. doi: 10.1207/S15326950dp3001_03

Rossi, M. G. (2013). *Il Giudizio del Sentimento. Emozioni, giudizi morali, natura umana*. Rome: Editori Riuniti.

Rossi, M. G. (2014). Emozioni e deliberazione razionale. *Sistemi Intelligenti* 26, 161–170. doi: 10.1422/77224

Rozin, P., and Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personal. Soc. Psychol. Rev.* 5, 296–320. doi: 10.1207/S15327957PSPR0504_2

Rubio Fernandez, P. (2007). Suppression in metaphor interpretation: differences between meaning selection and meaning construction. *J. Semant.* 24, 345–371. doi: 10.1093/jos/ffm006

Santibáñez, C. (2010). Metaphors and argumentation: the case of Chilean parliamentarian media participation. *J. Pragmat.* 42, 973–989. doi: 10.1016/j.pragma.2009.08.019

Schauenburg, G., Conrad, M., von Scheve, C., Barber, H. A., Ambrasat, J., Aryani, A., et al. (2019). Making sense of social interaction: emotional coherence drives semantic integration as assessed by event-related potentials. *Neuropsychologia* 4, 1–13. doi: 10.1016/j.neuropsychologia.2019.01.002

Scherer, A. M., Scherer, L. D., and Fagerlin, A. (2014). Getting ahead of illness: using metaphors to influence medical decision making. *Med. Decis. Mak.* 35, 37–45. doi: 10.1177/0272989X14522547

Schn, D. A. (1993). "Generative metaphor: a perspective on problem-setting in social policy," in *Metaphor and Thought,* ed A. Orthony (Cambridge: Cambridge University Press), 137–163. doi: 10.1017/CBO9781139173865.011

Schnall, S. (2005). The pragmatics of emotion language. *Psychol. Inq.* 16, 28–31.

Schumann, J., Zufferey, S., and Oswald, S. (2020). The linguistic formulation of fallacies matters: the case of causal connectives. *Argumentation* 2020, 1–28. doi: 10.1007/s10503-020-09540-0

Schwarz, N., and Clore, G. L. (1983). Mood, misattribution, and judgments of well-being: informative and directive functions of affective states. *J. Personal. Soc. Psychol.* 45, 513–523. doi: 10.1037/0022-3514.45.3.513

Schwarz, N., and Skurnik, I. (2003). "Feeling and thinking: implications for problem solving," in *The Psychology of Problem Solving,* eds J. E. Davidson and R. J. Sternberg (Cambridge: Cambridge University Press), 263–290. doi: 10.1017/CBO9780511615771.010

Semino, E. (2008). *Metaphor in discourse*. Cambridge: Cambridge University Press.

Slothuus, R. (2008). More than weighting cognitive importance: a dual-process model of issue framing effects. *Polit. Psychol.* 29, 1–28. doi: 10.1111/j.1467-9221.2007.00610.x

Solomon, R. C. (Ed.). (2003). *What Is an Emotion?* Oxford: Oxford University Press.

Sperber, D., and Wilson, D. (1986). *Relevance: Communication and Cognition*. Oxford: Blackwell.

Steen, G. J., Reijnierse, W. G., and Burgers, C. (2014). When do natural language metaphors influence reasoning? A follow-up study to Thibodeau and Boroditsky (2013). *PLoS ONE* 9:e113536. doi: 10.1371/journal.pone.0113536

Stevenson, C. (1944). *Ethics and Language*. New Haven, CT: Yale University Press.

Stocker, M., and Hegeman, E. (1996). *Valuing Emotions*. Cambridge: Cambridge University Press.

Svačinova, I. (2014). Reconstruction of metaphors in argumentation: a case study of Lincoln's metaphor of "swapping horses while crossing a stream". *Cogency* 6, 67–89.

Tamir, M., Robinson, M. D., and Clore, G. L. (2002). The epistemic benefits of trait-consistent mood states: an analysis of extraversion and mood. *J. Personal. Soc. Psychol.* 83, 663–677. doi: 10.1037/0022-3514.83.3.663

Thibodeau, P. H., and Boroditsky, L. (2011). Metaphors we think with: the role of metaphor in reasoning. *PLoS ONE* 6:e16782. doi: 10.1371/journal.pone.0016782

Thibodeau, P. H., and Boroditsky, L. (2015). Measuring effects of metaphor in a dynamic opinion landscape. *PLoS ONE* 10, 1–22. doi: 10.1371/journal.pone.0133939

Thibodeau, P. H., Matlock, T., and Flusberg, S. J. (2019). The role of metaphor in communication and thought. *Lang. Linguist. Compass* 13:e12327. doi: 10.1111/lnc3.12327

Tindale, C. W. (2006). *Fallacies and Argument Appraisal*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511806544

Tourangeau, R., and Sternberg, R. J. (1982). Understanding and appreciating metaphors. *Cognition* 11, 203–244. doi: 10.1016/0010-0277(82)90016-6

Tversky, A., and Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science* 211, 453–458. doi: 10.1126/science.7455683

Van Eemeren, F. H. (1992). *Argumentation, Communication and Fallacies*. Hillsdale: Erlbaum.

Van Eemeren, F. H., and Grootendorst, R. (2004). *A Systematic Theory of Argumentation. The Pragma-Dialectical Approach*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511616389

van Poppel, L. (2018). "Argumentative functions of metaphors: how can metaphors trigger resistance?," in *Argument and Inference. Proceedings of the 2nd European Conference on Argumentation volume II,* eds S. Oswald and D. Maillat. Fribourg.

van Poppel, L. (2020). The study of metaphor in argumentation theory. *Argumentation* 20:9523. doi: 10.1007/s10503-020-09523-1

Wagemans, J. H. M. (2016). Analyzing metaphor in argumentative discourse. *RIFL* 102, 79–94. doi: 10.4396/20161207

Wagemans, J. H. M. (2019). Four basic argument forms. *Res. Lang.* 17, 57–69. doi: 10.2478/rela-2019-0005

Walton, D. N. (1996). *Fallacies Arising from Ambiguity*. Dordrecht: Springer. doi: 10.1007/978-94-015-8632-0

Walton, D. N. (2006). *Fundamentals of Critical Argumentation*. Cambridge, MA: Cambridge University Press.

Walton, D. N. (2010). Why fallacies appear to be better arguments than they are. *Inf. Log.* 30, 159–184. doi: 10.22329/il.v30i2.2868

Wilson, D., and Sperber, D. (2002). Truthfulness and relevance. *Mind* 111, 583–632. doi: 10.1093/mind/111.443.583

Woods, J., and Walton, D. N. (1989). *Fallacies: Selected Papers, 1972–1982. Dordrecht: Foris*. doi: 10.1515/9783110816082

Zajonc, R. B. (1980). Feeling and thinking: preferences need no inferences. *Am. Psychol.* 35, 151–175. doi: 10.1037/0003-066X.35.2.151

Conflict of Interest: 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 © 2021 Ervas, Rossi, Ojha and Indurkhya. 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) and the copyright owner(s) 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 | Mean (M) and standard deviation (SD) for each measure of the metaphors in Experiment 1 and Experiment 2.

	Emotional meaning		Familiarity		Meaningfulness		Comprehension difficulty	
	M	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**
Rating study (Experiment 1) *Metaphors*								
CM+	4.36	0.48	3.48	1.04	4.43	0.52	1.47	0.47
NM+	4.29	0.49	2.71	0.73	4.08	0.61	1.75	0.63
CM-	1.97	0.80	3.33	1.00	4.30	0.78	1.71	0.78
NM-	1.89	0.71	2.48	0.97	3.56	0.77	2.46	0.88
CM0	3.01	0.74	3.69	0.81	4.17	0.61	1.68	0.58
NM0	2.99	0.59	2.20	0.96	3.29	0.74	2.65	0.86
Rating study (Experiment 2) *Metaphors*								
CM+	4.10	0.51	3.11	0.81	3.96	0.49	1.88	0.55
NM+	4.09	0.52	2.02	0.73	3.82	0.58	2.15	0.53
CM-	1.98	0.45	3.02	0.86	3.29	0.70	2.64	0.75
NM-	1.92	0.55	2.03	0.77	3.17	0.55	2.76	0.67
CM0	3.05	0.68	3.67	0.93	4.01	0.41	1.89	0.44
NM0	2.93	0.51	2.13	0.88	3.29	0.50	2.48	0.60

Positively valenced (CM+), negatively valenced (CM-), non-emotive (CM0) conventional metaphor. Positively valenced (NM+), negatively valenced (NM-), non-emotive (NM0) novel metaphor.

TABLE A2 | Table of materials in Italian for each middle term condition in Experiment 1 and Experiment 2.

	Non-emotive metaphors (0)	Positive-valenced metaphors (+)	Negative-valenced metaphors (–)
		Experiment 1	
CM	[P1] Il corpo è un tempio. [P2] Un tempio è sacro. [C] Il corpo è sacro.	[P1] Un caffè è un ristoro. [P2] Un ristoro ridà benessere. [C] Un caffè ridà benessere.	[P1] Un insulto è uno sfregio. [P2] Uno sfregio è una ferita. [C] Un insulto è una ferita.
	[P1] Una risata è un farmaco. [P2] Un farmaco serve alla salute. [C] Una risata serve alla salute.	[P1] Lo sport è la vitalità. [P2] La vitalità è energia. [C] Lo sport è energia.	[P1] Una separazione è un trauma. [P2] Un trauma è un evento violento. [C] Una separazione è un evento violento.
	[P1] La vita è un capitolo. [P2] Un capitolo è una parte di un libro. [C] La vita è una parte di un libro.	[P1] La tolleranza è un antidoto. [P2] Un antidoto contrasta i veleni. [C] La tolleranza contrasta i veleni.	[P1] Una sconfitta è un tonfo. [P2] Un tonfo è una caduta a terra. [C] Una sconfitta è una caduta a terra.
	[P1] Quella ragazza è una gemma. [P2] Una gemma è una pietra preziosa. [C] Quella ragazza è una pietra preziosa.	[P1] La pace è una schiarita. [P2] Una schiarita è il ritorno del sereno. [C] La pace è il ritorno del sereno.	[P1] Il traffico è un ingorgo. [P2] Un ingorgo è un intasamento. [C] Il traffico è un intasamento.
NM	[P1] Il gallo è un faraone. [P2] Un faraone è un capo. [C] Il gallo è un capo.	[P1] La libertà è una risata. [P2] Una risata è un'espressione di gioia. [C] La libertà è un'espressione di gioia.	[P1] Lo smog è un accusato. [P2] Un accusato è un possibile colpevole. [C] Lo smog è un possibile colpevole.
	[P1] Un grido è un megafono. [P2] Un megafono emette suoni. [C] Un grido emette suoni.	[P1] Il tempo è un alleato. [P2] Un alleato è un amico. [C] Il tempo è un amico.	[P1] Il circo è una slealtà. [P2] Una slealtà è un'attività disonesta. [C] Il circo è un'attività disonesta.
	[P1] Il cuore è un'anfora. [P2] Un'anfora trasporta liquidi. [C] Il cuore trasporta liquidi.	[P1] La serenità è un vaccino. [P2] Un vaccino rende immuni. [C] La serenità rende immuni.	[P1] La politica è un tugurio. [P2] Un tugurio è in pessime condizioni. [C] La politica è in pessime condizioni.
	[P1] Il lavoro è un badile. [P2] Un badile serve nei campi. [C] Il lavoro serve nei campi.	[P1] La felicità è un brillio. [P2] Un brillio è una luminosità intensa. [C] La felicità è una luminosità intensa.	[P1] Un tradimento è un'ustione. [P2] Un'ustione è una grave scottatura. [C] Un tradimento è una grave scottatura.
		Experiment 2	
CM	[P1] Il corpo è un tempio. [P2] Un tempio è sacro. [C] Il corpo è sacro.	[P1] Una separazione è un ristoro. [P2] Un ristoro ridà benessere. [C] Una separazione ridà benessere.	[P1] La finitura è uno sfregio. [P2] Uno sfregio è una ferita. [C] La finitura è una ferita.
	[P1] Una risata è un farmaco. [P2] Un farmaco serve alla salute. [C] Una risata serve alla salute.	[P1] La disubbidienza è la vitalità. [P2] La vitalità è energia. [C] La disubbidienza è energia.	[P1] L'eroismo è un trauma. [P2] Un trauma è un evento violento. [C] L'eroismo è un evento violento.
	[P1] La vita è un capitolo. [P2] Un capitolo è una parte di un libro. [C] La vita è una parte di un libro.	[P1] L'indifferenza è un antidoto. [P2] Un antidoto contrasta i veleni. [C] L'indifferenza contrasta i veleni.	[P1] L'ammirazione è un tonfo. [P2] Un tonfo è una caduta a terra. [C] L'ammirazione è una caduta a terra.
	[P1] Quella ragazza è una gemma. [P2] Una gemma è una pietra preziosa. [C] Quella ragazza è una pietra preziosa.	[P1] Un distacco è una schiarita. [P2] Una schiarita è il ritorno del sereno. [C] Un distacco è il ritorno del sereno.	[P1] Lo spettacolo è un ingorgo. [P2] Un ingorgo è un intasamento. [C] Lo spettacolo è un intasamento.
NM	[P1] Il gallo è un faraone. [P2] Un faraone è un capo. [C] Il gallo è un capo.	[P1] L'insolenza è una risata. [P2] Una risata è un'espressione di gioia. [C] L'insolenza è un'espressione di gioia.	[P1] Il bagno è un accusato. [P2] Un accusato è un possibile colpevole. [C] Il bagno è un possibile colpevole.

(Continued)

TABLE A2 | Continued

	Non-emotive metaphors (0)	Positive-valenced metaphors (+)	Negative-valenced metaphors (–)
	[P1] Un grido è un megafono. [P2] Un megafono emette suoni. [C] Un grido emette suoni.	[P1] Il silenzio è un alleato. [P2] Un alleato è un amico. [C] Il silenzio è un amico.	[P1] Il casinò è una slealtà. [P2] Una slealtà è un'attività disonesta. [C] Il casinò è un'attività disonesta.
	[P1] Il cuore è un'anfora. [P2] Un'anfora trasporta liquidi. [C] Il cuore trasporta liquidi.	[P1] Il dolore è un vaccino. [P2] Un vaccino rende immuni. [C] Il dolore rende immuni.	[P1] L'ateneo è un tugurio. [P2] Un tugurio è in pessime condizioni. [C] L'ateneo è in pessime condizioni.
	[P1] Il lavoro è un badile. [P2] Un badile serve nei campi. [C] Il lavoro serve nei campi.	[P1] Una battaglia è un brillio. [P2] Un brillio è una luminosità intensa. [C] Una battaglia è una luminosità intensa.	[P1] La religione è un'ustione. [P2] Un'ustione è una grave scottatura. [C] La religione è una grave scottatura.

TABLE A3 | Table of literal arguments (fillers in Experiment 1 and Experiment 2).

	Strong arguments	Weak arguments	Arguments with plausible conclusion
Literal	[P1] Il cane rincorre il gatto. [P2] Il gatto è un felino. [C] Il cane rincorre un felino.	[P1] La sarta rammenda il vestito. [P2] Il vestito è un indumento. [C] La sarta è un indumento.	[P1] Chiara ha una zia. [P2] Una zia è la moglie dello zio. [C] Chiara è la moglie dello zio.
	[P1] Brad Pitt è una persona. [P2] Una persona è un essere umano. [C] Brad Pitt è un essere umano.	[P1] Il lunedì inizia la settimana. [P2] La settimana ha sette giorni. [C] Il lunedì ha sette giorni.	[P1] Antonio richiama un soldato. [P2] Un soldato è un militare. [C] Antonio è un militare.
	[P1] Il Louvre è un museo. [P2] Un museo espone oggetti. [C] Il Louvre espone oggetti.	[P1] La nonna prepara la cena. [P2] La cena è un pasto. [C] La nonna è un pasto.	[P1] La nipote visita la zia. [P2] La zia ha sessant'anni. [C] La nipote ha sessant'anni.
	[P1] La mucca produce il latte. [P2] Il latte è un alimento. [C] La mucca produce un alimento.	[P1] Melania mangia la mela. [P2] La mela è un frutto. [C] Melania è un frutto.	[P1] Carla possiede una bottiglia. [P2] Una bottiglia ha un collo. [C] Carla ha un collo.

frontiers in Psychology

ORIGINAL RESEARCH
published: 19 July 2021
doi: 10.3389/fpsyg.2021.679491

The Cost of the Epistemic Step: Investigating Scalar Implicatures in Full and Partial Information Contexts

Maria Spychalska[1]*, *Ludmila Reimer*[2], *Petra B. Schumacher*[1] and *Markus Werning*[2]*

[1] Institute for German Language and Literature I, Linguistics, University of Cologne, Cologne, Germany, [2] Institute of Philosophy II, Ruhr University of Bochum, Bochum, Germany

We present the first ERP experiments that test the online processing of the scalar implicature *some* ⇝ *not all* in contexts where the speaker competence assumption is violated. Participants observe game scenarios with four open cards on the table and two closed cards outside of the table, while listening to statements made by a virtual player. In the full access context, the player makes a fully informed statement by referring only to the open cards, as *cards on the table*; in the partial access context, she makes a partially informed statement by referring to the whole set of cards, as *cards in the game*. If all of the open cards contain a given object X (Fullset condition), then *some cards on the table contain Xs* is inconsistent with the *not all* reading, whereas it is unknown whether *some cards in the game contain X* is consistent with this reading. If only a subset of the open cards contains X (Subset condition), then both utterances are known to be consistent with the *not all* implicature. Differential effects are observed depending on the quantifier reading adopted by the participant: For those participants who adopt the *not all* reading in the full access context, but not in the partial access context (weak pragmatic reading), a late posterior negativity effect is observed in the partial access context for the Fullset relative to the Subset condition. This effect is argued to reflect inference-driven context retrieval and monitoring processes related to epistemic reasoning involved in evaluating the competence assumption. By contrast, for participants who adopt the logical interpretation of *some* (*some and possibly all*), an N400 effect is observed in the partial access context, when comparing the Subset against the Fullset condition, which is argued to result from the competition between the two quantifying expressions *some cards on the table* and *some cards in the game* functioning in the experiment as scalar alternatives.

Keywords: scalar implicature, primary and secondary implicature, epistemic step, competence assumption, alternatives, late posterior negativity, N400

OPEN ACCESS

Edited by:
Valentina Cuccio,
University of Messina, Italy

Reviewed by:
Chris Cummins,
University of Edinburgh,
United Kingdom
Jacques Moeschler,
Universit de Genve, Switzerland
Walter Schaeken,
KU Leuven, Belgium

***Correspondence:**
Maria Spychalska
mspychal@uni-koeln.de
Markus Werning
markus.werning@rub.de

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

Received: *11 March 2021*
Accepted: *11 May 2021*
Published: *19 July 2021*

Citation:
Spychalska M, Reimer L, Schumacher PB and Werning M (2021) The Cost of the Epistemic Step: Investigating Scalar Implicatures in Full and Partial Information Contexts. Front. Psychol. 12:679491. doi: 10.3389/fpsyg.2021.679491

1. INTRODUCTION

In Gricean tradition, scalar implicatures are contents that are derived through a pragmatic mechanism that involves reasoning about the speaker's intentional and epistemic states. For instance, consider the paramount example of a scalar implicature: If a speaker says

(1) *Some of the cookies have disappeared,*

a listener is in a position to infer that

(2) *Not all of the cookies have disappeared,*

even though semantically (1) is also true if all of the cookies have disappeared. This mechanism can be described in more detail as follows: (i) the speaker chose to use the quantifier *some*; (ii) the speaker is obeying the Maxim of Quantity; (iii) if the speaker believed that a semantically stronger sentence of similar content with the alternative quantifier *all* was true, she would have said so; (iv) it follows, that she does not have a belief that the sentence with *all* is true, namely

(3) $\neg B_S(p)$,

where p is a sentence *all cookies have disappeared*. The latter can be the case either since the speaker believes p to be false or since she is not informed whether it is true. Note that (3), sometimes referred to as *primary implicature*, is not logically equivalent to sentence (2). To arrive at (2) as the speaker implied content, one needs to make an additional assumption that the speaker is informed, or at least opinionated, about the issue under discussion, i.e., she either believes the sentence with *all* to be true, or she believes it to be false ($B_S(p) \vee \neg B_S(p)$). Only given this so-called "competence assumption" (Sauerland, 2004; Schulz and Van Rooij, 2006; Geurts, 2010), one can infer that the speaker actually believes the alternative sentence with *all* to be false ($B_S(\neg p)$), equivalently, she takes (2) to be true (*secondary implicature*). This inference from the primary to the secondary implicature is sometimes referred to as "epistemic step" (Sauerland, 2004; Breheny et al., 2013).

Practically all current theories of scalar implicatures assume that they result from some sort of interplay between relevant semantic alternatives in the lexicon, e.g., *some* and *all* in the case of quantifiers. Since Horn (1972, 1989) this mechanism has been described more formally by assuming that these alternatives can be ordered according to their semantic strength to create the so-called implicational scale: ⟨*all, some*⟩. The stronger quantifier *all* semantically entails the weaker quantifier *some*, but the use of *some* gives rise to the implicature that a sentence of similar content with the stronger alternative is false.

However, it has been debated whether this mechanism is indeed pragmatic in nature and involves epistemic reasoning as described above, or whether it is rather embedded in grammar (Chierchia, 2004; Chierchia et al., 2011). In the grammatical account by Chierchia et al. (2011), it is argued that scalar implicatures arise as an effect of inserting the silent exhaustification operator (*exh*) of semantics similar to *only*, that acts directly on the scalar alternatives. Consequently, sentences such as *Some As are B* are ambiguous between two readings which correspond to two distinct logical forms, namely the literal, existential reading *There are As that are B* (paraphrased as *Some and possibly all As are B* and referred to as the *logical reading*) and the strengthened reading that corresponds to the secondary implicature: *There are As that are B but not all As are B* (paraphrased as *Only some As are B*). Under this view, the exhaustification occurs as obligatory once scalar alternatives are contextually active. The role of pragmatics is then reduced to activating the alternatives, for instance, based on the contextual access, and this process is optional.

The main difference between the more traditional, neogricean theories and the grammatical ones seems to concern the very nature of scalar implicatures, namely, the role of pragmatics in their derivation. From this perspective, the role of epistemic reasoning should be considered of most importance. This epistemic component in deriving scalar implicatures can be investigated by considering contexts where the speaker lacks access to full information that would allow her to evaluate the sentence with the stronger alternative (*partial access contexts*). Under the neogricean view, in such contexts, only the primary implicature can be inferred, namely, that the speaker does not hold a believe that a sentence with *all* is true (henceforth referred to as the *weak pragmatic reading*). Since this implicature is then trivially satisfied, no further strengthening is possible, as it would lead to contradiction. By contrast, under the grammatical view, the reading corresponding to the secondary implicature (henceforth referred to as the *strong pragmatic reading*) is available as an alternative logical form. As it does not depend on the Gricean mechanism described above, it is also available in partial access contexts. Importantly, in contexts with fully informed speakers (*full access contexts*), the weak and strong pragmatic readings cannot be differentiated, since if the speaker competence can be assumed, the *some but not all* reading is compatible with both accounts. This interpretation is then, in the literature, simply referred to as the *pragmatic reading*.

At face value it appears that partial access contexts provide a straightforward test between the neogricean and grammatical view; however, the issue is more subtle. Whereas, the former predicts that strong pragmatic readings are unavailable in such contexts, the latter has less specific predictions. The *exh* operator is considered mandatory "if the scalar alternatives are contextually active" (cf. Chierchia et al., 2011), but it is not clear what it means for the alternatives to be active. In particular, a proponent of the grammatical view may argue that in partial access scenarios the activation of scalar alternatives is inhibited, resulting in the endorsement of the logical interpretation. In this case, the grammatical view would not predict strong pragmatic readings in such contexts either. Furthermore, the reading with (only) the primary implicature is possible under the grammatical view as well, namely, either as an alternative logical form or as a pragmatic enrichment of the logical reading resulting from the neogricean mechanism (cf. Fox, 2014; Dieuleveut et al., 2019). In particular, the grammatical view does not deny that pragmatic mechanisms exist, rather it postulates that the *some, but not all* reading is available also as an alternative semantic parse of a sentence with *some*, and not just as a pragmatically enriched reading.

Thus, on the one hand, if one could show that strong pragmatic readings are available in partial access contexts, such a result would constitute good evidence for the grammatical theory; on the other hand if such readings cannot be found, it might not suffice to disprove this account. Therefore, it seems more promising to investigate not just the availability of the weak and strong readings on the behavioral level, but also the processing costs involved in deriving these readings. The main question is whether partial access contexts involve a processing cost that could be linked to the presumed

epistemic reasoning. For instance, if the derivation of the *not all* implicature involves epistemic processes of evaluating the speaker's competence, we should see a mark of these processes both in the case when the implicature is derived, as well as in the case when it has to be canceled, respectively inhibited, due to the competence assumption violation. In the current paper we test this question by comparing the processing of pragmatically ambiguous sentences with *some* in contexts with partial and full access to the quantified domain.

1.1. Implicatures in Contexts of Full Information

The majority of prior research on the processing of scalar implicatures has involved experiments where full information was available to all parties involved. In these experiments, the status of the stronger alternative could be determined based on world-knowledge or a visual scenario. It has repeatedly been observed that in contexts where the strong alternative *all* is known to be true, sentences with *some*—which are then considered *underinformative*—tend to trigger divergent truth-value judgments: Their evaluation as true is taken to indicate that *some* is interpreted logically, whereas their evaluation as false indicates the pragmatic interpretation of *some*, i.e., with the scalar implicature. Such divergent truth-value judgments have been observed both for cases where the truth-value can be determined based on world-knowledge, e.g., *Some people have lungs* (Bott and Noveck, 2004), as well as in sentence-picture verification paradigms (Spychalska et al., 2016).

In addition, it has generally been assumed that if scalar implicatures are generated as results of a complex Gricean-like reasoning process, they should involve more effortful processing relative to the literal interpretation. However, both the experimental results on the implicature processing as well as the interpretation of these results have been mixed. Although, in principle both response times and eye-tracking results suggest that there exists a processing cost related to scalar implicatures (Bott and Noveck, 2004; Huang and Snedeker, 2009; Spychalska et al., 2016), other studies showed that scalar implicatures may be processed at no cost (Grodner et al., 2010; Politzer-Ahles and Fiorentino, 2013), especially if the context sufficiently supports the pragmatic interpretation (Degen and Tanenhaus, 2014, 2015; Hartshorne et al., 2015).

To directly investigate whether the scalar implicature is processed incrementally, some studies have used event-related brain potentials (ERPs), focusing especially on the N400 component. ERPs are scalp-recorded voltage changes time-locked to trigger events, such as spoken or written words; the N400 is a negative-going shift in the ERP waveform occurring between 200 and 600ms post-stimulus onset, and usually maximal around 400 ms post-onset over the centro-parietal scalp sites (Kutas and Federmeier, 2011; Swaab et al., 2012). It is modulated by broadly-understood semantic expectancy and predictability of the stimulus, e.g., it tends to be larger for words that are semantically less appropriate or less expected in the context (Kutas and Hillyard, 1980; Kutas and Van Petten, 1994; Kutas and Federmeier, 2000). The size of the N400 is also inversely correlated with the cloze probability of the triggering word, i.e., the percentage of individuals who would continue a given sentence fragment with that word (Federmeier et al., 2007). In affirmative sentences, sentence continuations that make the sentence semantically false tend to elicit larger N400 than "true" sentence continuations (Nieuwland and Kuperberg, 2008; Nieuwland and Martin, 2012; Nieuwland, 2016; Spychalska et al., 2016). Although, the functional role of the N400 has been debated (inter alia Kutas and Federmeier, 2000; Hagoort et al., 2004; DeLong et al., 2005; Nieuwland et al., 2018), many recent accounts model the N400 in terms of probabilistically understood meaning-related predictability/expectancy of the stimulus (Lau et al., 2013; Kuperberg and Jaeger, 2015; Rabovsky et al., 2018).

The majority of ERP studies on scalar implicatures have focused on measuring the modulation of the N400 elicited by content nouns downstream from the quantifier phrase when the pragmatic interpretation of the quantifier makes those nouns more or less expected (Noveck and Posada, 2003; Nieuwland et al., 2010; Hunt et al., 2013; Spychalska et al., 2016). The hypothesis behind this approach is that if scalar implicatures are incrementally processed, they should modulate the expectation for upcoming words in the linguistic input so that words consistent with the implicature would be more expected by the processor and hence elicit smaller N400 ERPs than words that are only consistent with the semantic reading but inconsistent with the implicature. Those studies have indeed showed such an effect but only under certain conditions, for instance, for those individuals who based the truth-related evaluation of the utterance on the pragmatic interpretation (Hunt et al., 2013; Spychalska et al., 2016) or for those with specific personality traits, such as low autistic spectrum quotient (AQ) (Nieuwland et al., 2010).

For instance, in the ERP experiment by Spychalska et al. (2016) (see also Hunt et al., 2013, for a similar design), participants were asked to truth-evaluate sentences such as *Some/All pictures contain Xs* with respect to visual displays consisting of five pictures and containing two categories of objects: one of them was presented in each of the pictures, whereas the other one only in two or three out of all five pictures. An example scenario displayed five pictures, each of them containing an image of a cat on the left hand side and three of them containing an additional image of a ball on the right hand side. In such a scenario, *Some pictures contain cats* is true but inconsistent with the implicature (*Some-Infelicitous* condition), whereas *Some pictures contain balls* is true and consistent with the implicature (*Some-True* condition). For those participants who predominately evaluated Some-Infelicitous sentences as false (they were labeled as *pragmatists*), a biphasic ERP effect (an N400 followed by a P600) was observed for sentence-final nouns in this condition relative to the Some-True condition. No similar effect was found for those participants who predominately evaluated Some-Infelicitous trials as true (*logicians*). Thus, the processing patterns were fully determined by the reading of *some* as reflected in the participants truth-value judgements: If the scalar implicature was taken as part of sentence truth-conditions, then its processing was incremental as indicated by the observed N400 effect.

It is also interesting that in the study by Spychalska et al. (2016), a P600 effect was observed in addition to the N400. The P600 is a long-lasting, late positive shift in the ERP wave, maximal around 600 ms post-onset over centro-parietal sites. It occurs in response to syntactic errors or complexities (Osterhout and Holcomb, 1992; Hagoort et al., 1993; Osterhout et al., 1994), but also other linguistic irregularities (Kuperberg et al., 2003; Hoeks et al., 2004; Van Herten et al., 2005; Bornkessel-Schlesewsky and Schlesewsky, 2008), as well as some pragmatic violations (Chevallier et al., 2010; Regel et al., 2011). It has been argued to reflect combinatorial aspects of linguistic processing (Kuperberg, 2007), inferential processing (Burkhardt, 2006; Schumacher, 2014), or even semantic integration mechanisms (Brouwer et al., 2012). In the study by Spychalska et al. (2016), the observed P600 was argued to reflect truth-related reprocessing of the sentence.

1.2. Implicatures in Contexts of Partial Information

Thus, far, only a few studies have attempted to test the role of the speaker's epistemic state for the listener's interpretation of pragmatically weak statements. Goodman and Stuhlmüller (2013) run online questionnaires using scenarios where the speaker, who had either a complete or partial access to the model, made a statement using the quantifier *some*, e.g., *I have looked at 2 out of 3 letters* (partial access)/*3 out of 3 letters* (complete access). *Some of them have checks inside*. Participants were asked to estimate, by means of betting, the number of letters that had checks inside. A diminution of the pragmatic interpretation of *some* was observed in contexts where the speaker had only a partial access to the scenario: Whereas in the complete access condition, bets on 3 letters were much lower than bets on 2 letters, which shows that the implicature was derived, in the partial access condition no difference between bets on 2 and 3 letters was observed. In a self-paced reading study, Bergen and Grodner (2012) also showed a suppression of the pragmatic interpretation of *some* in situations where the speaker was assumed to have only partial knowledge. These results are generally in line with the neogricean view on scalar implicatures. However, Dieuleveut et al. (2019) provide some evidence in support of the grammatical account: Using a sentence-picture verification questionnaire, they found that *some but not all* readings may also be available in contexts where the competence assumption is violated. In this study, participants were presented with two players, who had a different view of the same set of playing cards: one of the players could see all of the cards (knowledgeable player), the other one could only see a subset of the cards (ignorant player), whereas the participants could always see all of the cards. The participants were then asked to evaluate whether a given player "can say" a given sentence. For instance, in a situation where all of the cards were hearts, the neogricean view predicts that the ignorant player, who had only a partial access to the set of cards, could say that *some of the cards are hearts*, whereas the grammatical view allows for a response that "she cannot say that," which would reflect the strong pragmatic reading. Notably, Dieuleveut et al. (2019) report up to 45% of "cannot say that" responses in such cases, supporting the grammatical account. The authors note that this result is not compatible with the traditional Gricean view, although they acknowledge that it may be reconcilable with more recent pragmatic approaches, such as the game-theoretic Rational Speech Act model (Goodman and Stuhlmüller, 2013; Bergen et al., 2016). Given that their design leaves it underspecified what kind of information the two players have about each other, specifically, what are the listener's (the other player's) beliefs about the speaker's beliefs, it could happen that the subjects engage in complex higher-order reasoning that takes into account speaker's beliefs about the listener's beliefs about the speaker's beliefs. In this case, subjects would give "cannot say that" responses not to mislead the listener about the speaker's epistemic state, more specifically, not to suggest to the other player that the speaker has full access to her cards and, thus, intends to actually communicate the scalar implicature.

Breheny et al. (2013) investigated the role of epistemic reasoning in the processing of so-called "ad hoc" (based on context-derived scales) quantity implicatures. They used a paradigm similar to the director's task known from studies on the theory of mind reasoning in *reference resolution* (Keysar et al., 2000, 2003; Nadig and Sedivy, 2002; Hanna et al., 2003; Sedivy, 2003; for a review see Noveck and Reboul, 2008). Participants were listening to a confederate speaker, who described events presented on the computer screen. The situations were observed both by the confederate speaker and the listener (i.e., subject), and consisted of an agent moving objects into two boxes, e.g., a spoon into box A, a spoon into box B, and then a fork into box A. Whereas, in the *knowledge condition* the whole sequence of actions was visible to both parties, in the *ignorance condition* the subject was aware that the speaker could not see the last action. A knowledgeable speaker could then describe the action, for instance, by saying that the agent *put a spoon into B and a spoon and a fork into A*, whereas an ignorant speaker could only say that the agent *put a spoon into B and a spoon into A*. In the latter case, the subject, who knew that nothing else was put into B (unlike into A), should have already anticipated the box B upon hearing "into" in the first subclause. Yet, subjects were aware that the speaker, who could not see the last action, could have also referred to box A in the first subclause. Using eye tracking it was shown that this information about the speaker's epistemic state modulated the listeners' processing of the speaker's utterance, namely, the listeners were delayed in anticipating the right box in the ignorance condition.

The paradigm where the listener and the speaker have a different access to the scenario has the advantage of allowing to test whether the listener takes the speaker's or egocentric perspective while interpreting the utterance. However, to distinguish between the primary and secondary implicature of *some*, it is sufficient to use a less complex paradigm, where partial access scenarios are introduced, without allowing the listener a privileged access.

Let us consider a game scenario with six cards, four of which are placed face-up, whereas two are placed face-down. All of the face-up cards are hearts but it is not known what suit the face-down cards are. Suppose now that a speaker looks at the cards and says *Some of the cards are hearts*. Under the neogricean view, it is a pragmatically felicitous statement, since

it is not known whether all of the cards are hearts, thus, the primary implicature is satisfied. Under the grammatical view, the statement is ambiguous between the two readings: the logical reading and the strong pragmatic reading, i.e., the reading with the *exh* operator. If *Some of the cards are hearts* is interpreted under its strong pragmatic reading, it is not expected to be uttered in this scenario since its truth-value cannot be determined.

Note that it makes little sense to ask whether *Some of the cards are hearts* is true in this case, since such a question would bias toward the neogricean view on SIs. In this framework the statement is simply true: The logical reading of *some* is true and so is the primary implicature. In contrast, under the grammatical view, the truth-value of this statement is only determined if it is interpreted logically. Otherwise, it is unknown.

Instead of asking whether the sentence is true or false, we should rather ask whether the sentence "can be uttered," or whether "it is appropriate for the speaker to utter it," given the information she has. It is another Gricean maxim, Maxim of Quality, that requires from the speakers to make their contribution one that is true, by not making false statements or statements for which they lack sufficient evidence. Taking the reformulation of the maxim by Gazdar (1979), the speaker is then expected to only utter statements which she knows or believes to be true. Therefore, if the listener adopts the strong pragmatic reading, she should consider *Some cards are hearts* as an "inappropriate" utterance if not all cards can be seen.

Partial access scenarios allow us to test the availability of the strong pragmatic reading; however, they do not allow us to differentiate between the primary implicature (weak pragmatic reading) and the logical reading: In such scenarios, *Some cards are hearts* is true and appropriate under both these readings. To contrast the primary implicature and the logical interpretation, full access scenarios are necessary. Suppose now that the four face-up cards are dealt to the table and the two additional face-down cards are dealt outside of the table, as the speaker's cards; however, in such a way that the speaker cannot look at them. The speaker then describes the game situation by referring either to all of the dealt cards as "cards in the game" or to the open cards only, as "cards on the table." Any statement about the "table-cards" can be truth-evaluated; however, for statements referring to the cards in the game one has to consider also the closed cards. Thus, if the speaker refers to the cards on the table, she has full access to the quantifier domain, whereas if she refers to the cards in the game, she has only partial access.

Introducing two alternative domain restrictions allows us to compare partial with full access contexts in situations where the visual scene remains unchanged. However, the choice of such an experimental design has further consequences regarding the interplay between the alternatives in the experimental context, since the two domain restrictions function as scalar alternatives as well. *Some* is a monotone increasing quantifier in its first argument, which means that if *Some As are B* is true and A is a subset of A', then *Some A's are B* is true as well. Given that, in this specific setup, *cards on the table* is a subset of *cards in the game*, *Some cards on the table are hearts* entails that *Some cards in the game are hearts*, but the inverse does not hold. Accordingly, *Some cards in the game* may give rise to the implicature that *Some cards on the table* is not applicable. This inadequacy of the Table restriction must be due to pragmatic reasons, since whenever *Some cards in the game are hearts* is *known* to be true, *Some cards on the table are hearts* must be semantically true as well (since hearts must be visible in order for the former to be known). In short, *Some cards in the game are hearts* may be seen as pragmatically more appropriate in contexts where all visible cards are hearts simply because if only a subset of visible cards are hearts, the Table restriction is more informative. In fact, if all visible cards are hearts, the Game restriction can be used without making a commitment about the scalar implicature of the bare *some*. The same is the case if only a subset of the visible cards are hearts and one uses the Table restriction. Such a global pragmatic mechanism is predicted by the hypothesis that scalar implicatures are derived not only at the level of quantifier and based on a locally-given context (a given "hand" of cards) but they also arise in relation to a global context which, in this case, is created by a competition between two alternative domain restrictions that are used in the whole experimental setting.

1.3. The Current Study

In the current experiment, we used a paradigm similar to the above-described game-scenario to test the processing of sentences with the quantifier *some* in contexts with a partial and full access to the quantified domain. Participants were exposed to game situations and asked to judge statements of a previous player, introduced as Lena, who had recorded her descriptions of these scenarios before the experiment.

The target scenarios consisted of (i) the speaker's avatar presented at the top left part of the screen; (ii) a game table with **four open cards**; and (iii) **two closed cards** placed face-down outside of the table (**Figure 1**). In the target scenarios, all visible cards showed two different object categories, one in each of the four open cards (e.g., *apples* in **Figure 1**), and one in 2 or 3 of the open cards (*hedgehogs*). The subjects were informed that the face-down cards were also closed from the speaker's perspective, so that she had not seen what was depicted on them. The speaker's utterances were presented auditorily while the scene was displayed. The experiment employed a 2 (**Context**) $\times$ 3 (**Set**) design: The speaker's utterances either referred to the cards in the game, i.e., all cards including the face-down cards (**partial access** or **Game context:** *Some cards in the game contain As*), or to the cards on the table only (**full access** or **Table context:** *Some cards on the table contain As*). The critical noun (factor **Set**) referred to (i) the object category contained by every visible card (**Fullset** condition); (ii) the object category contained by a subset of visible cards (**Subset** condition); (iii) another object category not presented on the screen (**Unprimed** condition) (**Table 1**). The ERPs were measured at the onset of the critical noun.

The participants were told that Lena had a task to describe the given game scenarios. Some of her descriptions were appropriate and some were not. The inappropriate statements could haven been mistakes or they could have been chosen deliberately. The subjects' task was to determine which statements of Lena were good, i.e., "appropriate" (German "angemessen") and which

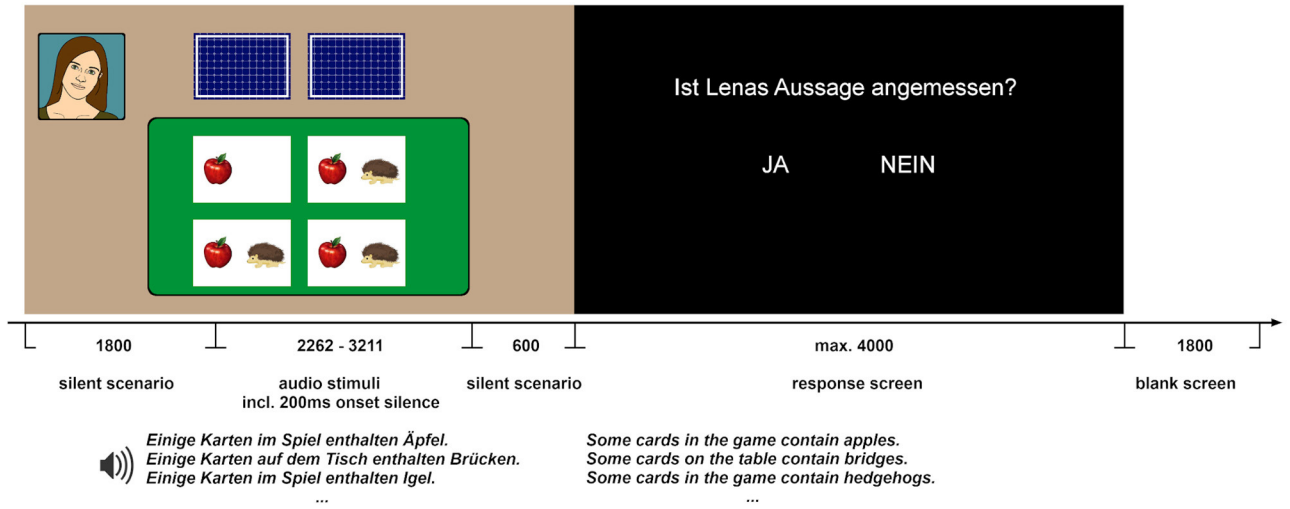

FIGURE 1 | The structure of a trial: The picture was presented 1,800 ms before the onset of the audio and until 600 ms after the audio offset. The audio files varied in length (range: 2,262–3,211 ms, the average duration of Table-sentences was 2,800 ms and the average duration of Game-sentences was 2,603 ms) and contained approximately 200 ms of an initial silence for a more natural sound onset. The onset of the critical word was between 1,857 and 2,664 ms after the onset of the audio, the average onset was 2,139 ms. Subsequently, a screen with the assignment of the buttons was presented (left vs. right hand is counterbalanced). For a given scene, the condition was dependent on the specific sentence played, as explained in **Table 1**.

TABLE 1 | Experimental conditions for the example scene in **Figure 1**.

	Set		
Context	***hedgehogs* "*Igel*"**	***apples* "Äpfel"**	***bridges* "*Brücken*"**
Table	**Subset**	**Fullset**	**Unprimed**
Some cards on the table contain...	true	true/false*	false
'Einige Karten auf dem Tisch enthalten...'	appropriate	appropriate/inappropriate*	inappropriate
Game	**Subset**	**Fullset**	**Unprimed**
Some cards in the game contain...	true	true/unknown*	unknown
'Einige Karten im Spiel enthalten...'	appropriate	appropriate/inappropriate*	inappropriate

For the given scene, the condition is determined by the sentence context (Context factor: Table vs. Game) and the critical noun (Set factor: Fullset/Subset/Unprimed). For each condition, the table provides the truth-value of the target sentence as well as the expected appropriateness judgement; (*) marks the ambiguous case, where the truth-value and the response depend on the logical/weak pragmatic/strong pragmatic interpretation.

were "inappropriate" ("unangemessen")[1]. **"Inappropriate" statements** were explained to include: (i) all statements that were visibly false, such as, e.g., *No cards on the table contain frogs* in a set-up showing frogs, as well as (ii) all statements that could not be judged true or false based on the visible information, e.g., *All cards in the game contain mice* in a setting with four cards containing mice and two additional cards face-down. Only statements that were **known to be true** were **"appropriate"**. The subjects were explicitly told that Lena had seen the same set-up of cards and had no access to the content of cards that were shown backside-up[2].

In the full access contexts, that is, if the speaker referred to *some cards on the table*, it was always possible for the subject to evaluate whether the statement was true and consistent with the implicature (**Table 1**): The speaker's utterances were unambiguously true in the Table-Subset condition and, hence, appropriate. They were false in the Table-Unprimed condition and as such should be judged as inappropriate. Finally, in the Table-Fullset condition, they were inconsistent with the implicature and expected to trigger divergent appropriateness judgments: A response "inappropriate" would indicate that *at least* the primary implicature was derived, whereas a response "appropriate" would indicate the logical interpretation.

By contrast, in the partial access context, that is, if the speaker referred to *cards in the game*, the truth-value could unambiguously be determined only in Game-Subset condition: In this case, the speaker's utterances were known to be true, independently of the reading of *some*, and should be judged as "appropriate" descriptions. In the Game-Unprimed condition,

[1]By using "appropriate/inappropriate" rather than "can/cannot say that" we avoided the reading of "is able to," making it more natural for the subjects to judge a past event (i.e., Lena's utterance that has already happened). Another benefit is that "appropriate" is used less frequently than "can" in German and, thus, it is easier to establish a certain, intended reading of "appropriate" in an experimental setting, avoiding any potential "fuzzy" readings that might occur with the use of "can".
[2]See **Supplementary Materials** for the full text of the instructions.

the truth-value was unknown, and according to the instruction, subjects should judge such descriptions as "inappropriate." However, in the Game-Fullset condition, the speaker did not know the status of the stronger alternative *all*. Thus, the utterances were known to be true under the logical interpretation, as well as under the reading with the primary implicature, which should trigger "appropriate" judgments, but they were false and hence "inappropriate" under the strong pragmatic reading.

Consequently, three different interpretation patterns were possible in the experiment. **Logicians** would give "appropriate" judgments both in the Table-Fullset and Game-Fullset condition. Participants who derive only a primary implicature (referred to as **weak pragmatists**) would respond "inappropriate" in the Table-Fullset but "appropriate" in the Game-Fullset condition. Finally, responding "inappropriate" both in the Table-Fullset and Game-Fullset conditions would indicate that a secondary implicature was derived in spite of the speaker competence assumption being violated (**strong pragmatists**). Importantly, the strong pragmatic interpretation is inconsistent with the classical Gricean account, but predicted under the grammatical theory. Dieuleveut et al. (2019) showed that such interpretations are in principle available to the speakers.

1.3.1. Predictions Regarding the Expected ERP Effects

Given that in the Unprimed conditions, unlike in the Fullset or Subset conditions, the critical nouns referred to objects that were absent in the respective scenarios and, thus, not visually primed, we expected larger N400 ERPs in the Unprimed conditions relative to the Fullset and Subset conditions, for both context types. This prediction is uncontroversial and based on the prior literature that shows a modulation of the N400 by priming (Kutas and Federmeier, 2011).

The contrast between conditions Table-Fullset and Table-Subset allows us to test the effect of the implicature violation in the full access context. From a logical perspective, this contrast corresponds to the comparison between Some-Infelicitous and Some-True conditions in Spychalska et al. (2016). Therefore, a similar ERP pattern could be expected, namely, differential effects depending on the (appropriateness) judgments given by the participants: a null effect in the case of the logical interpretation and a biphasic N400/P600 effect in the case of the pragmatic interpretation. However, one should keep in mind that the current study differs to a large extent from the prior one, including factors that potentially may affect the time-course of the implicature processing, such as (i) different modality (auditory vs. visual presentation of the linguistic stimuli) and, consequently, a different (faster) pace of the stimulus presentation, (ii) different type of judgment (appropriateness vs. truth-value) (iii) different proportion of scalar alternatives in the design (e.g., fewer trials with *all*), and finally (iv) the presence of the two different contexts (Table vs. Game) in the current design. The role of these aspects for the observed effects is discussed in more detail in the section 4.

Contrasting conditions Game-Subset and Game-Fullset allowed us to directly test the role of epistemic access for the processing of pragmatically ambiguous sentences with *some*. In the former condition, the speaker's utterances are unambiguously true and in the latter condition, their status depends on the interpretation: they are true under logical and weak pragmatic interpretation, but have an unknown truth-value under the strong pragmatic interpretation. Therefore, differential effects were expected depending on the appropriateness judgments' pattern. In principle, as the strong pragmatic reading reflects a high level of lexicalization of the implicature, the meaning-related expectancy for the critical noun should be modulated by reading and reflected in higher N400 ERPs for the condition directly inconsistent with this reading, i.e., condition Game-Fullset.

Although both the logical and weak pragmatic reading would be associated with the same "appropriate" judgments in the Game-Fullset condition, the two reading were expected to lead to differential ERP patterns. Unlike the logical reading, the weak pragmatic interpretation assumes sensitivity to the implicature, which is derived in the full access context and is either suppressed or canceled in the partial access context. Thus, for weak pragmatists, the epistemic uncertainty associated with the Game-Fullset condition was expected to lead to more effortful processing, possibly due to the evaluation of the speaker's epistemic access. Although these processes were not expected to modulate the N400 ERPs, they were expected to trigger a different type of effect. For instance, in memory research, late (and sustained) posterior negativity has been linked to (delayed) context retrieval/context monitoring/revision processes (see Mecklinger et al., 2016, for a review). By contrast, other authors link late positivity effects to inferencing and reevaluation mechanisms (cf. Burkhardt, 2006; Schumacher, 2013).

Considering the expected ERP results, one should also take into account the contrast between the two context types in the experiment. Based on the prior literature, it is known that contextually available alternatives have an effect on the observed ERP effects (e.g., Augurzky et al., 2019). Based on the scalar relation between the alternative contexts, *Some cards in the game* may be considered underinformative in the Subset condition, given that for the Subset condition a more informative expression, i.e., *some cards on the table*, is available. By contrast, in the Fullset condition, the use of the Game restriction is pragmatically more optimal than the use of the Table restriction, as it does require any commitment about the scalar implicature of the bare *some*. This interplay between the informativity of the two expressions may have an effect on the processing patters. Upon hearing *some cards in the game* the processor may expect that the object contained by all visible cards will be mentioned rather than the one contained by a subset of visible cards. Similarly, the expectation for the object contained by the subset of cards should be larger in the Table context rather than in the Game context. This may lead to larger N400 ERPs for the Game-Subset relative to the Game-Fullset as well as relative to the Table-Subset condition. Notably, this prediction goes almost directly against the expected effect that should be observed in the case of strong pragmatists. Given that this prediction is based on the assumption that the processor does not commit to the scalar

implicature of bare *some*, this effect might be most pronounced for logicians.

A direct comparison between the two context types (Table vs. Game) for a given Set condition allows to contrast the role of context. The question is whether the Game context leads to generally more effortful processing due to the epistemic uncertainty aspect.

2. EXPERIMENT 1

2.1. Materials

A list of 240 German nouns and their corresponding images was used to construct the target trials. All nouns were used in their plural form, were dissyllabic, 4 – 9 characters long (mean: 6) and of medium corpus frequency (8–17, mean: 12.82)[3]; compound nouns were excluded. All nouns denoted concrete objects that are easy to identify in a picture and well-known to an average German speaker. For each word two sentences were recorded:

(4) *Einige Karten im Spiel enthalten Xs.*
Some cards in the game contain Xs.

(5) *Einige Karten auf dem Tisch enthalten Xs.*
Some cards on the table contain Xs.

These sentence forms were selected to be as similar as possible to the stimuli used by Spychalska et al. (2016) with the only modification being the domain restriction *in the game/on the table*. The sentences were pseudo-randomly distributed across conditions per participant. To this aim, the nouns were first organized into 240 unique triples in such a way that each word was never combined twice with the same word. These triples were then randomly assigned to conditions and used to generate visual scenarios. Consequently, throughout the experiment, each object from the set of 240 targets was shown exactly twice, but always in a combination with a different object and never on the same side (left vs. right) of the picture. The combined words were controlled with respect to their frequency (maximal difference was 4; mean difference was 1). In addition, the semantic relationship between the combined words was also weakly controlled on the basis of the Latent Semantic Analysis (LSA): since no easily-accessible LSA-evaluation tool based on German corpora existed yet, the LSA values were estimated based on English translation of words using the server http://lsa.colorado.edu/. The LSA values were kept under 0.3 (mean value: 0.092) and checked for every word pair combination within a triple. To avoid co-activation of phonetically competing strings, words with the same phonetic onsets, such as *Käfer* and *Kämme* were never combined.

There were 240 target trials (40 per condition): 80 unambiguously true (appropriate), 80 inappropriate (40 visibly false and 40 of unknown truth-value), 40 where the weak pragmatic reading would lead to the "inappropriate" answer, and 40 where the strong pragmatic reading would lead to the "inappropriate" answer.

[3]The frequency *value v* of a word *w* is equal to log_2 of the quotient of the frequency of the word "der" and the frequency of the word *w* in the corpus. The values were checked with the *Wortschatz Leipzig*.

A set of 180 filler trials was created, using additional 90 words (each filler object was shown 3–4 times, whereas each filler word was used twice auditorily). Filler trials constituted an important aspect of the design. They were used to introduce variation, reduce predictability and balance the distribution of appropriateness judgments in the experiment. Since all target trials used scenes with two closed cards outside of the table, we added filler trials with scenes where no additional cards outside of the table were dealt and scenes where these two cards were shown open. To prevent subjects from developing a strategy for predicting potential objects on the closed cards, the number of visible object categories in the filler trials was varied between 1 and 3, and subjects were explicitly instructed that the cards *may contain* 1–3 object categories. In this way, when seeing only two object categories, the subjects could not assume that there is no other category on the closed cards. The filler trials used the quantifiers: *all* (*alle*), *no* (*keine*), *more than two/three* (*mehr als zwei/drei*), *fewer than three/four* (*weniger als drei/vier*), *two/three* (*zwei/drei*), *some* (*einige*). The role of the trials with *all* was to create a contrast between *some* and its stronger alternative. Since sentences with *all* and *no* can never be visibly true in partial access scenarios (they are either visibly false or cannot be truth-evaluated), these quantifiers were also used to increase the number of (i) pragmatically unambiguous trials of unknown truth-value, as well as (ii) visibly false trials. Trials with the comparative quantifier *more than (two/three)* were introduced so that sentences with *some* were not the only ones that could be visibly true in partial access scenarios. : For instance, *More than three cards in the game contain Xs* is visibly true if there are four cards containing Xs, independently of the closed cards. Sentences with *fewer than* were used as a contrast, e.g., *Fewer than three cards contain Xs* is known to be false if four cards show Xs, independently of the closed cards. We also introduced trials with *zwei* and *drei* in order to make the materials further comparable with those from Spychalska et al. (2016), where fillers with bare numerals were used as well. Finally, fillers with *some* were used in order to vary the cards configuration and the number of shown object categories also for this quantifier. For more details see **Figure 2** and **Supplementary Materials**.

The Game and Table contexts were counterbalanced for all trial filler types. The filler trials were constructed in such a way that the sentence appropriateness and truth-value varied across all quantifiers. Since the target trials included more appropriate trials, we constructed more inappropriate filler trials. In the whole experiment the ratio of appropriate vs. inappropriate trials was 56/44% for a logician, 45/55% for a weak pragmatist and 35/65% for a strong pragmatist.

2.2. Audio Recording and Preprocessing

The audio files were recorded in a professional studio at the phonetics department of the Cologne University. The speaker was instructed to read the stimuli sentences in a neutral voice without emphasizing single words. The audio recording was preprocessed with *praat*. Intervals of around 200 ms (197–203 ms) were left at the beginning of each sentence for a more natural acoustic onset, the cut-off at the end was determined by the end of the coda of the last word. The onset of the target word for every

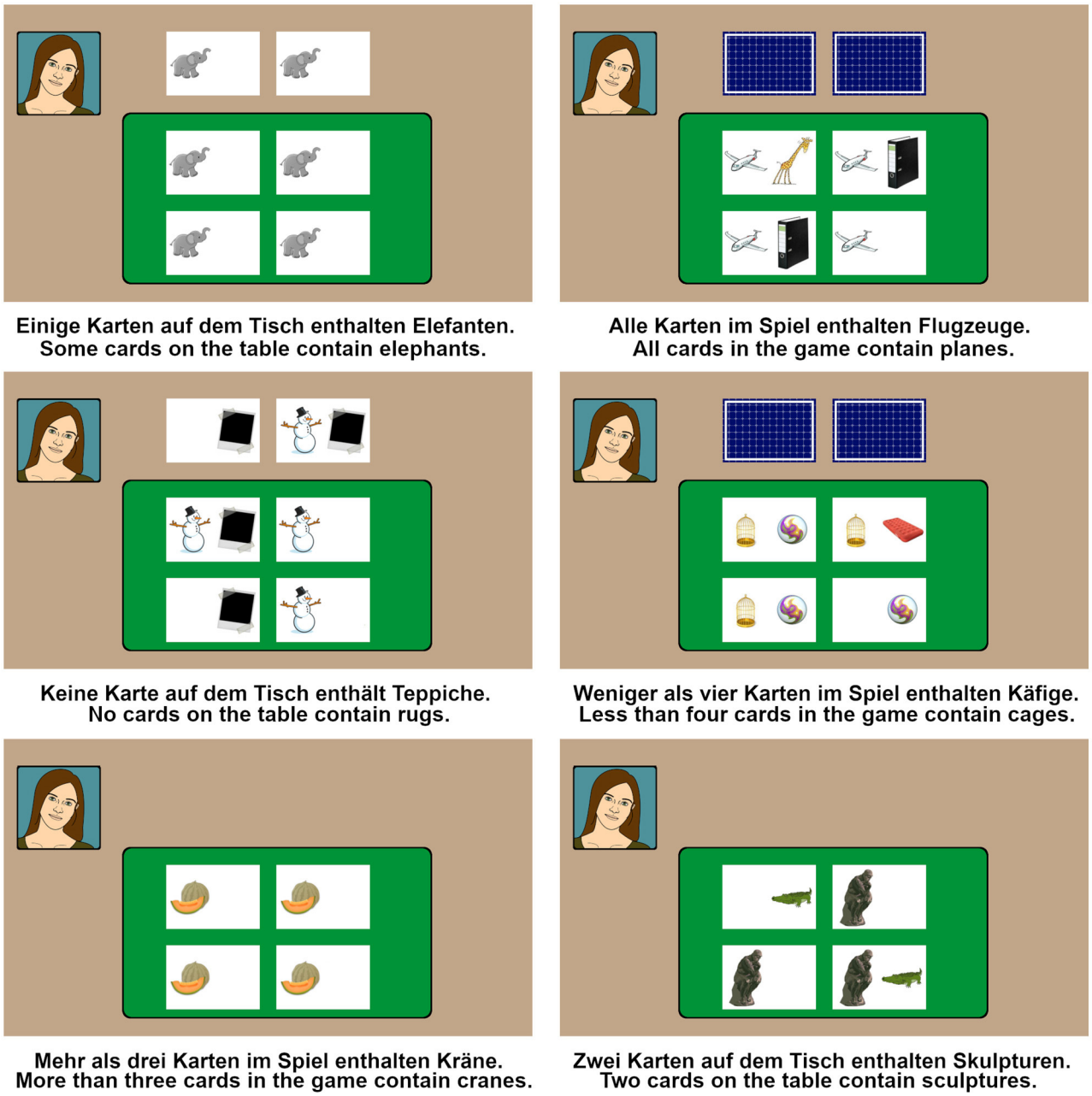

FIGURE 2 | Six example filler trials of various sorts. For each quantifier all three card configurations (closed/open/no additional cards) were used. The number of visible object categories 1–3 was counterbalanced per quantifier/context/card configuration. The two contexts (Table/Game) were cross-balanced for each filler category. Sentences with bare numerals are subject to similar pragmatic ambiguity as those with *some*, i.e., based on the so-called *exactly* vs. *at least* readings. Since all target trials with *some* always showed two additional cards face-down, we introduced more fillers with two closed cards than those with other card configurations, so that the ratio of trials with closed/open/no additional cards would be relatively balanced across quantifiers.

sentence was determined for sending the triggers into the EEG data file. After cutting the stimuli segments, every audio file was normalized for volume with the *audacity* software.

2.3. Participants

Fifty (34 women) participants were recruited for the experiment (age range: 18–42, mean: 25.14, SD: 4.86); They had at least a secondary degree (German Abitur), were German native speakers who did not learn a second language before the school age, had normal or corrected to normal vision, were right-handed, had no history of psychological or neurological problems, and were not under any medication at the time of testing. Three subjects had to be excluded since they did not complete the experiment due to an early disruption caused by technical problems with

the audio. Behavioral responses from all these 47 subjects are evaluated. From the EEG analysis, one subject was excluded due to a broken electrode channel and four additional subjects due to unexpected, inconsistent, or isolated response patterns, resulting in 43 subjects used in the statistical analysis of the ERP data.

2.4. Procedures

The experiment was conducted in the EEG laboratory at the Philosophy Department of the Ruhr-University of Bochum. Participants were seated inside an electrically isolated and acoustically attenuated cabin, in front of a shielded glass with a computer screen behind it. The USB-powered speakers were placed inside the cabin as well as the Cedrus (USB-powered) response pad with two designated buttons.

Upon arrival every subject signed a written, informed consent of participation. They were informed that their data will be stored and handled in a fully anonymous manner, and that they have the right to withdraw from the experiment at any time. They also filled in a statement concerning their vision, medication, neurological or psychiatric history, and the Edinburgh Handedness Inventory test. Additionally, they were screened using the Autistic Spectrum Quotient (AQ) Questionnaire, two parts of the Wechsler Adult Intelligence Scale (WAIS), the Matrix Reasoning test and the digit span working memory test, and an adjusted German variant of the Standardized Reading Span Memory Test developed by Van den Noort et al. (2008).

The experiment started with an instruction presented on the screen and 8 exercise trials. The exercises included trials with quantifiers *all*, *no*, *fewer than*, *more than*. Game and Table contexts, the cards configuration, the number of visible object categories as well as the sentence truth-value and appropriateness were varied (see **Supplementary Materials** for the full list of exercise trials). Feedback was provided for the exercises: Subjects were explicitly told that the sentence is not appropriate if its truth-value is unknown, e.g., in the case of *All cards in the game contain Xs*, where X is on all visible cards, but there are two additional closed cards. No feedback was provided after the exercise session ended[4].

Every trial started with a presentation of the scene (see **Figure 1**). After 1,800 ms from the onset of the scenario, the audio file was played. Triggers were sent for the onset of the trial, the onset of the audio and the onset of the sentence-final critical noun. The scenario stayed on the screen for an additional 600 ms after the offset of the audio file to allow for a more natural offset of the visual input as well as for recording undistorted ERPs up to approximately 1,000 ms post-onset the critical noun (average duration of the critical noun was 563 ms, $SD = 78.15$ ms).

2.5. EEG Recording and Data Preprocessing

The EEG was recorded with a 64-channel BrainAmp actiCAP EEG system. FCz location was used as the physical reference and AFz as the ground electrode. Four electrodes were relocated and used to measure eye-movement: FT9 and FT10 were used for horizontal movements (placed on the right and left temple), PO9 and PO10 for vertical movements (placed above and below the right eye). Impedance was kept below $5k\Omega$. The EEG was recorded with a sampling rate of 500 Hz, a 10 s low cut-off filter and a hardware anti-aliasing filter. The EEG data was processed using the software Brain Vision Analyzer 2.0. An off-line band-pass filter was applied: 0.1–30 Hz (order 4). Breaks and other periods of noisy signal were manually excluded. Automatic raw data inspection rejected all trials that had an absolute amplitude difference higher than $150\mu V$/150 ms or with activity lower than $0.5\mu V$ per 100 ms intervals. The maximal voltage step was $50\mu V$/ms. Both vertical and horizontal eye-movements were corrected by means of independent component analysis. Data was re-referenced to the average of mastoid electrodes (TP9 and TP10). Segments from 200 ms pre-target onset until 1,000 ms post-onset were extracted for every trial and condition. Baseline correction used the 200 ms interval preceding the onset of the stimulus. Segments with any remaining physical artifacts, including those with the amplitude lower than $-90\mu V$ or higher than $90\mu V$, were excluded and condition averages were calculated for each subject. The minimal number of segments that were preserved per subject and condition was 31/40.

2.6. Statistical Analysis

For the accuracy analysis we report descriptive statistics as well as non-parametric tests such as Friedmann test (for multiple conditions' comparisons) and Wilcoxon signed-rank tests for pairwise comparisons[5]. Kendall's W (Coefficient of Concordance) is used to estimate the effect size for the Friedman test.

For the analysis of reaction times (RTs), we performed a mixed ANOVA with *Context* (Table vs. Game) and *Set* (Fullset, Subset, Unprimed) as within-subject factors, and *Group* (2 levels: logicians vs. weak pragmatists) as a between-subject factor. For each subject the mean RTs in a given condition were computed after excluding missed and incorrect responses. Outliers (trials with response times longer than the mean in the given condition +2 standard deviations from this mean) were excluded before averaging[6]. Greenhouse-Geisser correction was applied whenever the result of the Mauchly's test indicated the violation of the sphericity assumption. For pairwise *post hoc* comparisons the *p*-values were Bonferroni corrected. Partial eta squared is reported for the effect size. For the sake of brevity, generally only the significant effects are reported.

For the statistical analysis of the EEG data we used the Matlab Fieldtrip package. We performed a non-parametric statistical procedure called cluster-based permutation test (Maris and Oostenveld, 2007; Oostenveld et al., 2011). For each subject the ERPs were averaged across trials in the compared conditions, in epochs of 0–1,000 ms post-onset and for all channels. The data-points (time × channel) between the sets were compared by a two-tailed dependent *t*-test. The significantly different ($\alpha =$

[4]See **Supplementary Materials**.

[5]The statistical methods were selected to correspond to those from Spychalska et al. (2016) for a better comparability of the results.

[6]No outliers shorter than the mean −2 SD were found.

0.05) data-points were then clustered according to the time-spatial adjacency. The cluster-level T-statistics were calculated by taking the sum over the *t*-values for each cluster. The cluster-level *p*-values were evaluated with a Monte Carlo simulation: For each subject the ERP averages were randomly swapped between the two conditions. The cluster-level statistics were computed again and the maximum of the cluster-level statistics was taken as the test statistics for this permutation. This procedure was repeated 10,000 times and the *p*-values of the observed cluster-level statistics were estimated as the proportion of permutations that resulted in a higher test-statistics than the observed one. This method allows one to test the significance of the effects over a broad spatio-temporal window, without having to choose any specific time-window or region, while correcting for false alarms related to the multiple comparisons. In this way one can test whether the null hypothesis that the compared conditions are exchangeable in the whole chosen epoch and the spatial region can be rejected. This method was selected due to a low level of specificity of our predictions.

We performed planned comparisons, i.e., three comparisons within the given Context level: Unprimed vs. Subset, Unprimed vs. Fullset and Fullset vs. Subset, both for the Table and Game contexts. In addition, we compared the two contexts for the same Set level (Game-Unprimed vs. Table-Unprimed, Game-Fullset vs. Table-Fullset, Game-Subset vs. Table-Subset). Given that differential effects were expected depending on the response pattern, the analysis was planned for each observed group of responders, assuming a representative sample size. Foreshadowing the results, we performed a separate analysis for the two main groups of responders (logicians and weak pragmatists).

2.7. The Analysis of Appropriateness Judgments

The accuracy rates (**Table 2**) in the unambiguous conditions Game-Subset, Table-Subset and Table-Unprimed were at ceiling level; however, the Game-Unprimed condition created difficulties for some of the subjects. In this condition the expected response is "inappropriate," since the sentence truth-value is unknown. However, three subjects consistently (for more than 95% of trials) judged the target utterances as appropriate in this condition. Two of these subjects explained afterwards that they responded in this way since they did not want to "tell Lena" (the speaker) that she is wrong in a situation when they themselves could not know that, whereas the third subject admitted that they knew what our intended correct response was but they decided to respond differently. These three subjects were excluded from further analysis.

For the remaining subjects, the analysis of appropriateness judgments in the critical conditions revealed that the majority of subjects (33 out of 44; 75%) accepted the target sentences as appropriate in both the Table-Fullset as well as the Game-Fullset condition, which indicates the logical interpretation of *some*. These participants (logicians) were relatively consistent in their responses: between 80 and 100% of trials judged as appropriate in condition Table-Fullset, and between 92 and 100% in condition Game-Fullset. Only 11 out of 44 subjects (27.3%) rejected the target utterances as inappropriate in condition Table-Fullset, with consistency varying between-subject from 62 to 100%. Out of these 11 subjects, 10 consistently accepted the target utterances as appropriate in condition Game-Fullset (weak pragmatists) and one consistently rejected such trials as inappropriate (strong pragmatist).

We compared accuracy across conditions for logicians and weak pragmatists separately: In the Table-Fullset condition accuracy was defined according to Group: the response "appropriate" was defined as correct for logicians and the response inappropriate for weak pragmatists. For weak pragmatists, accuracy differed across conditions [$\chi^2(5, N = 10) = 19.98, p < 0.001, W = 0.399$], but there was no significant effect for logicians. Based on pairwise comparisons of the Set conditions within each context (Wilcoxon signed ranks), pragmatists' mean accuracy was lower in condition Table-Fullset compared to both Table-Subset ($z = -2.371$, $p = 0.018$, $r = -0.433$) and Table-Unprimed ($z = -2.524$, $p = 0.012$, $r = -0.461$). Also the mean accuracy in the Game-Fullset condition was lower relative to Game-Subset ($z = -2.232, p = 0.026, r = -0.407$).

Since all the subjects were screened with respect to a number of cognitive and personality traits (**Table 3**), we tested whether any of these values would be predictors of the response patterns. Independent *t*-tests showed no significant differences between weak pragmatists and logicians with respect to their working memory, AQ, reasoning capabilities or age ($p > 0.05$ for each test).

2.8. Reaction Times Analysis

The full-factorial ANOVA of the RTs (**Table 4** and **Figure 3**) showed a significant main effect of Set [$F(2, 82) = 15.704, p < 0.001, \eta_p^2 = 0.277$]. The interaction Set × Context was significant [$F(1.383, 56.700) = 19.206, p < 0.001, \eta_p^2 = 0.319$], as well as the Set × Group interaction [$F(2, 82) = 9.337, p < 0.001, \eta_p^2 = 0.187$].

The significant Context × Set interaction was subsequently broken down both by Context and by Set. First, we analyzed the RTs for each Context separately. For the Game context, there was a main effect of Set [$F(1.356, 55.599) = 15.835, p < 0.001, \eta_p^2 = 0.279$]. Pairwise comparisons (*p*-values were Bonferroni corrected) showed that condition Game-Subset received significantly faster responses than conditions Game-Fullset [$F(1, 41) = 12.415, \eta_p^2 = 0.232, p = 0.003, \Delta_{Full,Sub} = 37.558$] or Game-Unprimed [$F(1, 41) = 19.856, \eta_p^2$=0.326, $p < 0.001$], $\Delta_{Un,Sub} = 96.658$. The contrast between conditions Game-Unprimed and Game-Fullset was also significant [$F(1, 41) = 11.071, \eta_p^2 = 0.213, p = 0.006, \Delta_{Un,Full} = 59.100$] with the former receiving slower responses.

For the Table context, there was a significant effect of Set [$F(1.668, 68.387) = 19.722, p < 0.001, \eta_p^2 = 0.325$] as well as a significant Set × Group interaction [$F(1.668, 68.387) = 16.799, p < 0.001, \eta_p^2 = 0.291$]. Pairwise comparisons (Bonferroni corrected) showed that the response times in the Table-Fullset

TABLE 2 | The mean percentage (with standard deviation) of trials per condition judged as "appropriate," after excluding occasional missed trials.

		Table			Game		
		Fullset	**Subset**	**Unprimed**	**Fullset**	**Subset**	**Unprimed**
Experiment 1							
Logicians	Approp.	98.41 (4.04)	99.02 (1.87)	0.76 (1.32)	99.09 (1.63)	99.01 (1.76)	2.57 (5.25)
(*N* = 33)	Accuracy	99.24 (1.32)	97.42 (5.25)				
Weak pragmatists	Approp.	11.65 (11.45)	98.5 (1.75)	2.00 (3.87)	97.49 (2.63)	99.5 (1.05)	4.50 (12.52)
(*N* = 10)	Accuracy	88.35 (11.45)	98.00 (3.87)	95.50 (12.52)			
Strong pragmatist	Approp.	0	100	0	0	100	0
(*N* = 1)	Accuracy	100		100	100		100
Other	Approp.	50.83 (46.52)	99.17 (1.44)	3.33 (3.82)	97.50 (0.0)	100 (0.0)	95.79 (2.85)
(*N* = 3)	Accuracy	96.66 (3.33)	4.21 (2.85)				
Experiment 2							
Weak pragmatists	Approp.	05.33 (5,89)	98.98 (1.88)	1.00 (1.27)	97.62 (2.81)	99.00 (1.58)	1.14 (1.27)
(*N* = 15)	Accuracy	94.67 (5,89)	99.00 (1.27)	98.86 (1.27)			
Logicians	Approp.	97.19 (4.52)	99.38 (1.77)	0.94 (1.29)	99.06 (1.86)	99.38 (1.77)	2.19 (3.39)
N = (8)	Accuracy	99.06 (1.29)	97.81 (3.39)				
Strong pragmatists	Approp.	5.83 (5.20)	98.29 (2.96)	0	74,17 (31.35)	98.33 (1.44)	0.83 (1.44)
(*N* = 3)	Accuracy	94.17 (5.20)		100	25.83 (31.85)		99.17 (1.44)
Other	Approp.	100	100	5.00	94.87	97.50	95.00
(*N* = 1)	Accuracy	95.00	5.00				

In addition, the mean accuracy recalculated for cases where "inappropriate" is the accurate response: For conditions Table-Unprimed and Game-Unprimed, "inappropriate" responses are considered accurate; for logicians, "appropriate" is taken as accurate in condition Table-Fullset, for weak pragmatists, "inappropriate" is taken as accurate in this condition; for strong pragmatists, "inappropriate" is accurate in both the Table-Fullset and Game-Fullset conditions. In Experiment 2, strong pragmatists include 2 consistent subjects and one who was undecided between the weak and strong pragmatic interpretation. The subjects labeled as "other" responded "inappropriate" in condition Game-Unprimed, which is considered inaccurate and reflects a different and non-intended understanding of the task. These subjects showed mixed response patterns in condition Table-Fullset: In Experiment 1, there was one weak pragmatist in this group, one logician and one undecided subject; In Experiment 2, this group includes one logician.

TABLE 3 | Age and gender distribution, as well as the mean values with standard deviations [*M*(*SD*)] for all cognitive tests, in percentages, per group and per experiment.

Group	***N***	**Gender men/women**	**Age**	**Digit Span**	**Reading Span**	**Matrix Reasoning**	**AQ total**
Experiment 1							
Logicians	33	10/23	24.61 (4.97)	10.42 (2.21)	67.61 (9.88)	11.36 (1.99)	14.82 (4.68)
Weak pragmatists	10	3/7	26.50 (5.29)	10.00 (2.79)	69.20 (9.72)	11.50 (1.90)	11.80 (4.32)
Strong pragmatist	1	0/1	28	10	67	14	22
Other	3	2/1	27.00 (6.25)	13.00 (6.08)	69.33 (13.05)	13.00 (1.00)	14.33 (1.53)
Experiment 2							
Logicians	8	7/1	22.75 (2.25)	10.00 (3.02)	65.5 (10.41)	n/a	16.63 (5.37)
Weak pragmatists	15	5/10	23.73 (4.13)	10.07 (2.25)	69.13 (10.10)	n/a	14.20 (4.90)
Strong pragmatist	3	1/2	23.67 (2.08)	10.00 (1.00)	66.33 (4.04)	n/a	13.33 (6.11)
Other	1	0/1	23	6	60	n/a	16

condition were significantly longer relative to condition Table-Subset [$F(1, 41) = 41.834$, $\eta_p^2 = 0.505$, $\Delta_{Full,Sub} = 68.545$, $p < 0.001$] and relative to condition Table-Unprimed [$F(1, 41) = 12.268$, $\eta_p^2 = 0.230$, $\Delta_{Full,Un} = 46.616$, $p = 0.003$]. The analysis of repeated contrasts showed that the longer response times in the Table-Fullset condition were driven by the pragmatic responses: the difference between the two groups was larger in the Table-Fullset condition than in the Table-Subset condition [$F(1, 41) = 25.266$, $p < 0.001$, $\eta_p^2 = 0.381$, Table-Fullset $\Delta_{WPrag,Log} = 107.524$ and Table-Subset $\Delta_{WPrag,Log} = 0.984$], but the difference between the two groups was similar in conditions Table-Subset and Table-Unprimed ($p > 0.1$).

Second, we tested the effect of Context directly for each Set condition. This step was necessary in order to test some of the missing comparisons. No effects were observed for the Subset condition. For the Unprimed conditions there was a significant effect of Context [$F(1, 41) = 12.253$, $p = 0.001$, $\eta_p^2 = 0.230$], here the mean response times for the Game context were longer than for the Table context ($\Delta_{Game,Table} = 73.879$ ms). For the Fullset condition, the effect of Context was significant [$F(1, 41)$

TABLE 4 | The mean response time (with SD) in ms after excluding missed, incorrect responses and outliers (condition mean ±2 SD).

	Table			Game		
Group	**Fullset**	**subset**	**Unprimed**	**Fullset**	**subset**	**Unprimed**
Experiment 1						
Log ($N = 33$)	584.03 (99.70)	568.75 (102.32)	595.72 (119.95)	573.83 (109.93)	577.85 (109.36)	650.80 (153.36)
W Prag ($N = 10$)	691.55 (157.31)	569.74 (105.06)	586.63 (62.83)	638.07 (143.61)	558.94 (80.11)	679.29 (242.27)
Experiment 2						
W Prag ($N = 15$)	693.68 (218.35)	599.37 (134.74)	614.18 (165.44)	658.39 (208.00)	574.49 (120.56)	648.35 (186.42)

In condition Table-Fullset, for weak pragmatists a correct response is the pragmatic response. Log = logicians, W Prag = weak pragmatists.

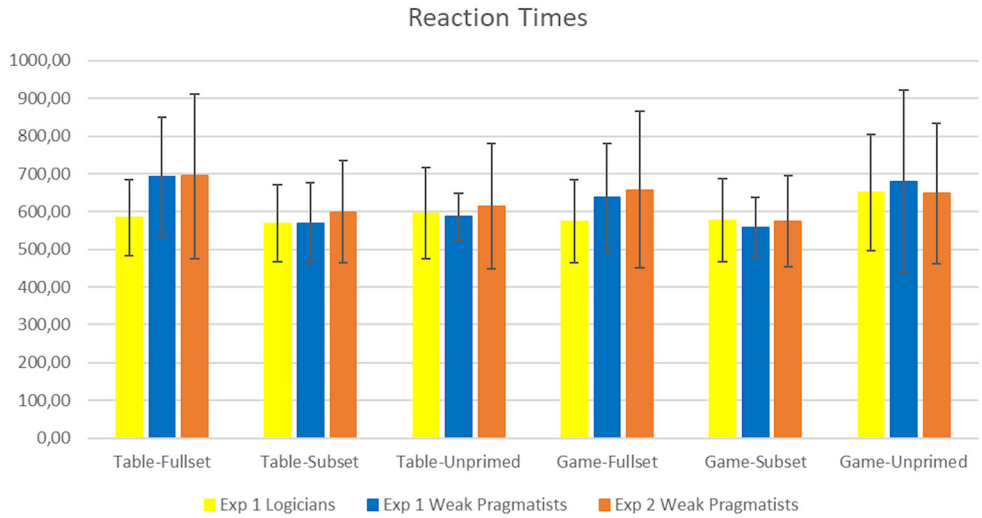

FIGURE 3 | The mean response times (in ms) with standard deviations (±1 SD), for each group and condition.

$= 8.00$, $p = 0.007$, $\eta_p^2 = 0.163$], namely, the Table context received on average longer response times than the Game context ($\Delta_{Table,Game} = 31.837$ ms). There was also an effect of Group [$F(1, 41) = 4.492$, $p = 0.04$, $\eta_p^2 = 0.099$], i.e., the mean response times of weak pragmatists were longer than those of logicians ($\Delta_{WPrag,Log} = 85.884$ ms).

2.9. EEG Results

The statistical analysis of the EEG data was performed separately for logicians (N = 33) and weak pragmatists (from the group of 10 weak pragmatists one had to be excluded from the ERP analysis due to broken electrode channels, so the number of subjects used was 9). The strong pragmatist was not included.

First, the three levels of the Set factor were compared for each Context level. For both contexts, a large negativity effect was observed for the Unprimed condition relative to the Subset and Fullset conditions. This effect started approximately 200 ms post-onset, lasted till the end of the epoch and had a global distribution ($p < 0.0001$ for each of the comparisons; see **Table 5** and **Figures 4**, **5**).

The comparison of the Table-Fullset and Table-Subset conditions did not show any significant effect for logicians as a separate group. However, for the weak pragmatists, this comparison lead to a marginally significant late negativity effect for the Table-Fullest condition, which based on the inspection of grand averages started around 600 ms post-stimulus onset and extended until the end of the epoch on the posterior sites: the effect's robustness is evaluated based on two adjacent clusters of 668–756 ms (marginally significant: $p = 0.068$) and 784–944 ms ($p = 0.040$).

For the Game context, differential effects are observed between the two groups. Logicians showed a negativity effect for the Game-Subset relative to Game-Fullset condition around the N400 time-window (344–662, $p = 0.0038$). By contrast, for weak pragmatists the Game-Subset vs. Game-Fullset contrast produced a positive cluster distributed over posterior channels (650–800, $p = 0.025$). The observed cluster has a positive polarity; yet, the cluster-based permutation is symmetric, so the polarity of the cluster depends on the direction of the comparison, but the polarity of the effect is a matter of interpretation. Based on the inspection of the temporal and topographical distribution of this effect, we interpret it rather as a late negativity for the Game-Fullset relative to Game-Subset condition: This effect appeared sustained until the end of the epoch and was most pronounced on the posterior sites.

TABLE 5 | The results of the cluster-based statistics for all comparisons: the temporal extension of the respective observed cluster (in ms) as well as its p-value evaluated by the permutation test.

Comparison	Cluster polarity	Experiment 1		Experiment 2
		logicians ($N = 33$)	weak pragmatists ($N = 9$)	*weak pragmatists* ($N = 15$)
Table unprimed vs. Subset	Negative	214–1,000, $p < 0.0001$	224–100, $p = 0.0032$	186–1,000, $p < 0.0001$
Table unprimed vs. Fullset	Negative	150–1,000, $p < 0.0001$	186–100, $p < 0.0001$	186–1,000, $p < 0.0001$
Table fullset vs. Subset	Negative	–	[668–756, $p = 0.068$]; 784–944, $p = 0.04$	–
Game unprimed vs. Subset	Negative	266–1000, $p < 0.0001$	250–1,000, $p < 0.0001$	222–1,000, $p < 0.0001$
Game unprimed vs. Fullset	Negative	204–1,000, $p < 0.0001$	244–636, $p = 0.0046$	222–946, $p < 0.0002$
Game subset vs. Fullset	Negative	344–662, $p = 0.0038$	–	–
	positive	–	650–800, $p = 0.025$	[596–736, $p = 0.069$]
Subset game vs. Table	Negative	372–582, $p = 0.0095$	374–540, $p = 0.0135$	–
Fullset game vs. Table	Negative	–	[270–368, $p = 0.08$]; 858–100, $p = 0.0559$	612–804, $p = 0.0074$
Unprimed game vs. Table	Neg/pos	–	–	–

Given that the weak pragmatists groups are relatively small, also marginally significant clusters are reported (0.5< α <0.1) in []. No extra rows are added in those cases where no significant positive clusters were found.

Comparing the ERPs across the contexts Game vs. Table showed that the ERPs in the Game-Subset condition were more negative than in the Table-Subset condition, both for logicians (372–582, $p = 0.0095$) and for weak pragmatists (374–540, $p = 0.0135$). No significant effects were found for the comparison between conditions Table-Unprimed and Game-Unprimed. For the comparison between Game-Fullset and Table-Fullset only a marginally significant late negativity effect is observed for weak pragmatists after 800 ms post-onset (858–100 ms, $p = 0.056$) as well as a non-significant trend of an effect around the N400 time window (270–368 ms, $p = 0.08$).

3. EXPERIMENT 2

Given that the number of weak pragmatist in the first experiment was rather low (only 9 usable subjects), some of the observed effects could be underpowered. Therefore, we decided to run a second experiment with the aim of replicating these effects. Given that the number of logicians in the first experiment was sufficient, we decided to focus on the pragmatists only. However, we did not want to give our subjects any explicit instruction with respect to how they should interpret the sentences, as this would have significantly changed the character of the experiment. Thus, we wanted to stick to the same procedure that allowed for spontaneous responses and interpretations. Furthermore, there is no clear predictor (e.g., personality test) that would allow us to determine who will turn out to respond pragmatically in such a task. Thus, we decided to use a similar task in a form of a short questionnaire to pre-screen potential subjects with respect to their tendency to respond pragmatically or logically.

3.1. Pre-screening

An online questionnaire was sent out to all the participants from the lab pool who did not participate in Experiment 1 or any similar experiment. They were told that we would like them to try out a sample of tasks used in the lab experiments and that everyone will be offered a testing date, independently of their responses. The questionnaire included 16 example questions similar to those from the experiment: The same type of visual scenarios was used, but for technical reasons the sentences were presented visually under the scenarios. The examples were mainly of the filler type, but one question corresponding to the Table-Fullset condition was used. Out of 108 participants who filled out the questionnaire, 50 responded pragmatically to this critical question, which is close to the usual pragmatic vs. logical response ratio in similar experiments on scalar implicatures (e.g., Spychalska et al., 2016), and were invited to the lab for the experiment. The remaining participants were invited for other, not related to pragmatics, experiments in our lab.

3.2. Participants

We tested 28 participants who were pre-screened as pragmatists[7]. One participant had to be excluded due to technical problems resulting in the experiment not being completed.

3.3. Behavioral Results

Out of 27 subjects who finished the experiment, one gave incorrect responses in the Game-Unprimed condition. Unexpectedly, as many as 8 subjects responded logically in the Table-Fullset condition in spite of giving a pragmatic answer in the questionnaire. Out of the remaining 18 subjects who responded pragmatically in the Table-Fullset condition, 2 were consistent strong pragmatists (responded pragmatically in the Game-Fullset condition) and one had a 40/60% ratio of weak vs. strong pragmatic responses. Thus, only 15 weak pragmatists could be included in the statistical analysis, whereas the other subjects were left out (see **Table 2**).

[7] Although every participant who pre-screened as pragmatist was offered a date, not all wanted to come for the EEG experiment; in addition, some of the scheduled dates had to be canceled due to the covid-19 pandemic outbreak.

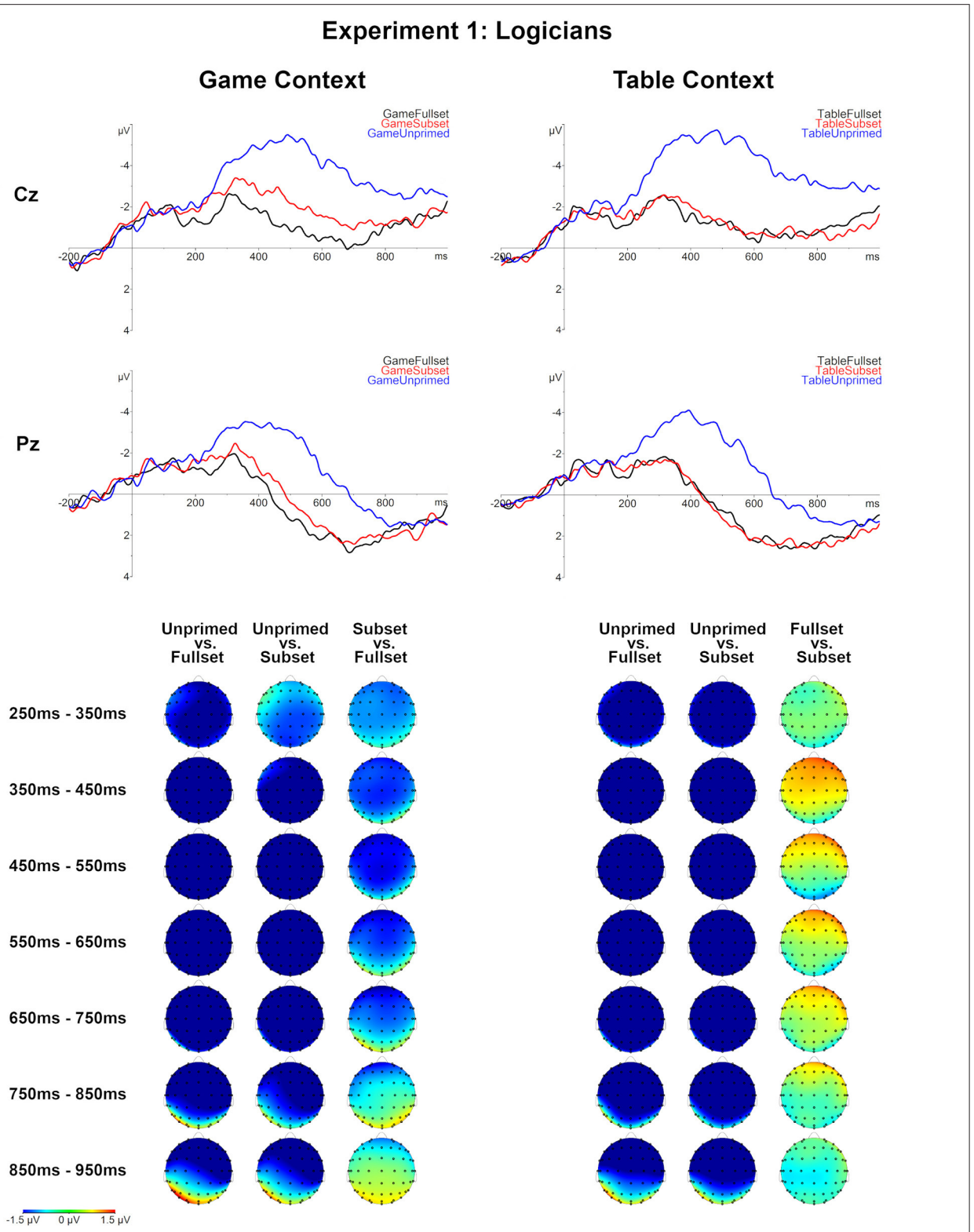

FIGURE 4 | Grand averages for all conditions for logicians ($N = 33$), Experiment 1. Topographical maps of the differences between the compared Set levels separately for the Table and Game context in consecutive 100 ms time-windows, between 250 and 950 ms post-onset.

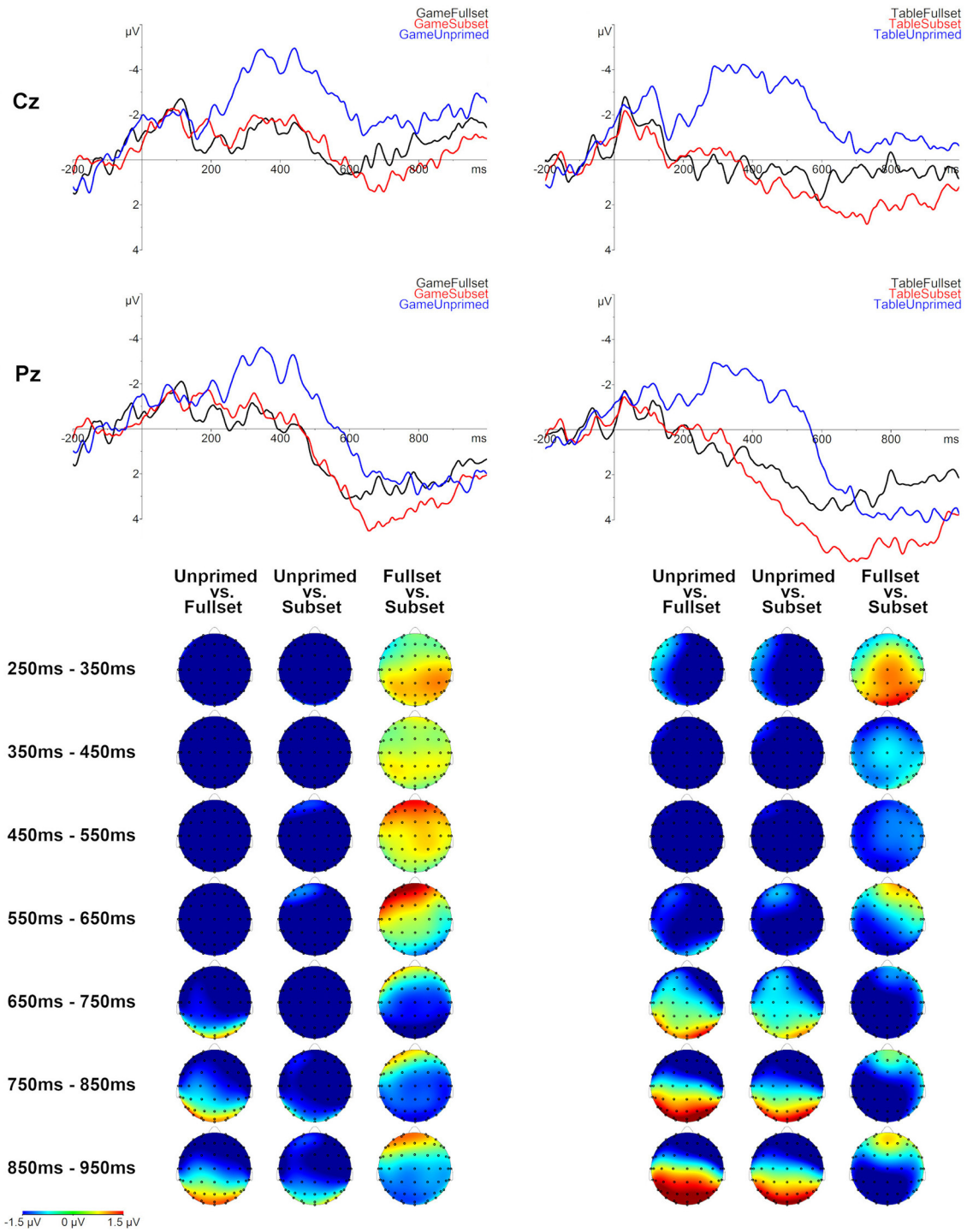

FIGURE 5 | Grand averages for all conditions for weak pragmatists ($N = 9$), Experiment 1. Topographical maps of the differences between the compared Set levels separately for the Table and Game context in consecutive 100 ms time-windows, between 250 and 950 ms post-onset.

Friedman test showed a marginally significant effect of condition for weak pragmatists' accuracy (in condition Table-Fullset the pragmatic response was defined as correct): $\chi^2(5, N = 15) = 11.003$, $p = 0.051$, $W = 0.147$.

3.4. Reaction Times

The analysis of response times (**Table 4** and **Figure 3**) for the 15 weak pragmatists showed a main effect of Set [$F(2, 28) = 8.758$, $p = 0.001$, $\eta_p^2 = 0.385$]. Context × Set interaction was only marginally significant [$F(1.42, 19.91) = 3.74$, $p = 0.055$, $\eta_p^2 = 0.211$, based on the Greenhouse Geisser correction]. Pairwise comparisons of the Set levels showed that the response times in the Fullset condition were significantly slower than in the Subset condition [$F(1, 41) = 11.725$, Bonferroni corrected $p = 0.012$, part. $\eta_p^2 = 0.456$, $\Delta_{Full,Sub} = 89.103$ ms]. The other contrasts were not significant.

3.5. EEG Results

The ERPs were analyzed for 15 weak pragmatists. Similar as in Experiment 1, the Unprimed conditions showed a larger negativity effect relative to Subset and Fullset conditions for both context types (**Table 5** and **Figure 6**). The negativity started around 200 ms post-onset and lasted almost till the end of the epoch. Unlike in Experiment 1, the comparison between Table-Fullset and Table-Subset did not show any significant effect. However, similar to the Experiment 1, the comparison between Game-Fullset and Game-Subset showed a late posterior negativity effect, which was marginally significant (the corresponding cluster latency and significance: 596–736 ms, $p = 0.069$). In addition, the comparison between Game-Fullset and Table-Fullest also showed a posterior negativity effect (612–804 ms, $p = 0.0074$).

4. DISCUSSION

In this paper we present the first ERP experiments testing the online processing of sentences with the weak scalar *some* in contexts where the speaker competence assumption does not hold. On the behavioral level we tested how the epistemic access to the quantified domain modulates the interpretation of the quantifier. By measuring the response times and ERPs we further tested how the online predictive processes are modulated by the quantifier interpretation on the one hand and the epistemic access to the domain on the other hand.

4.1. Appropriateness Judgments: Interpreting Truth and Appropriateness

Perhaps the most striking result at the behavioral level is a low ratio of pragmatic responses in the current experiments. In Experiment 1, only about 1/4 of the participants responded pragmatically, i.e., evaluated the target utterances with *some* as not appropriate in the Table-Fullset condition. Notably, practically all of these pragmatic responders accepted *some* as appropriate in the Game-Fullset condition, which means that they derived only the primary implicature. The strong pragmatic reading was adopted by only one participant in Experiment 1, and three more in Experiment 2, who judged the target *some*-utterances as inappropriate not just in the Table-Fullset but also in the Game-Fullset condition. Although the strong pragmatic interpretation is inconsistent with the neogricean view but predicted by the grammatical theory, it is not clear whether one can take such isolated responses as conclusive evidence to decide between the two approaches. Idiosyncratic response patterns may be observed due to various factors, for instance, as a strategic choice of an individual.

For Experiment 2, we only selected those participants who, based on a short questionnaire, were likely to display a pragmatic response pattern. In spite of that, almost one third of the tested individuals still responded logically during the actual EEG experiment. This result indicates that some features of the design made our participants more prone toward the logical interpretation, perhaps suppressing the implicature that they initially considered.

The low proportion of pragmatic relative to logical responses stays in contrast to the usual outcome of studies on scalar implicatures. For instance, Spychalska et al. (2016) reported an almost 50/50 divide between *pragmatists* and *logicians*. The reasons why so few subjects derived scalar implicatures in the current study may be manifold, including the presence of the partial access scenarios in the design, the proportion of other quantifiers used in the experiment, the modality of the stimuli presentation, as well as the nature of the task itself. Based on the oral feedback given by many of our participants, they found the question of whether the sentence is appropriate as more, rather than less, tolerant than the question about the sentence's truth-value. Some of the logical responders spontaneously and explicitly said that, strictly speaking, *Some of the cards on the table contain Xs* is not true when there are Xs on all of the cards; yet, in spite of not being true this sentence was still, according to them, "kind of appropriate." From a philosophical perspective such an interpretation is puzzling, since pragmatically infelicitous statements are on theoretical grounds considered semantically true but somewhat "not good enough." However, pre-theoretical intuitions of ordinary language users regarding truth may be dramatically different from the notions used by philosophers or semanticists: It seems that according to this alternative view, a sentence might not be considered true, but still "good enough." In other words, the fact that a sentence was "false" seemed, for some subjects, not sufficient to judge it as necessarily inappropriate. At the same time, sentences that were known to be *semantically false* were almost unanimously judged as inappropriate, except for the occasional misinterpretation cases, that we discuss below.

Epistemic uncertainty seemed to also play a significant role for the response patterns. Some subjects struggled with the intended interpretation of the task, namely, they accepted the target utterances with *some* as appropriate in the Game-Unprimed condition. After the experiment they explained that they were not willing to judge a speaker's statement as not appropriate if they themselves could not know the truth-value. Such an interpretation of the Game-Unprimed condition was surprising since all of our subjects seemed to have understood the task when given the instructions and they responded correctly during the exercise session. They were explicitly told that an utterance can only be regarded as appropriate if the *speaker knows* that it is

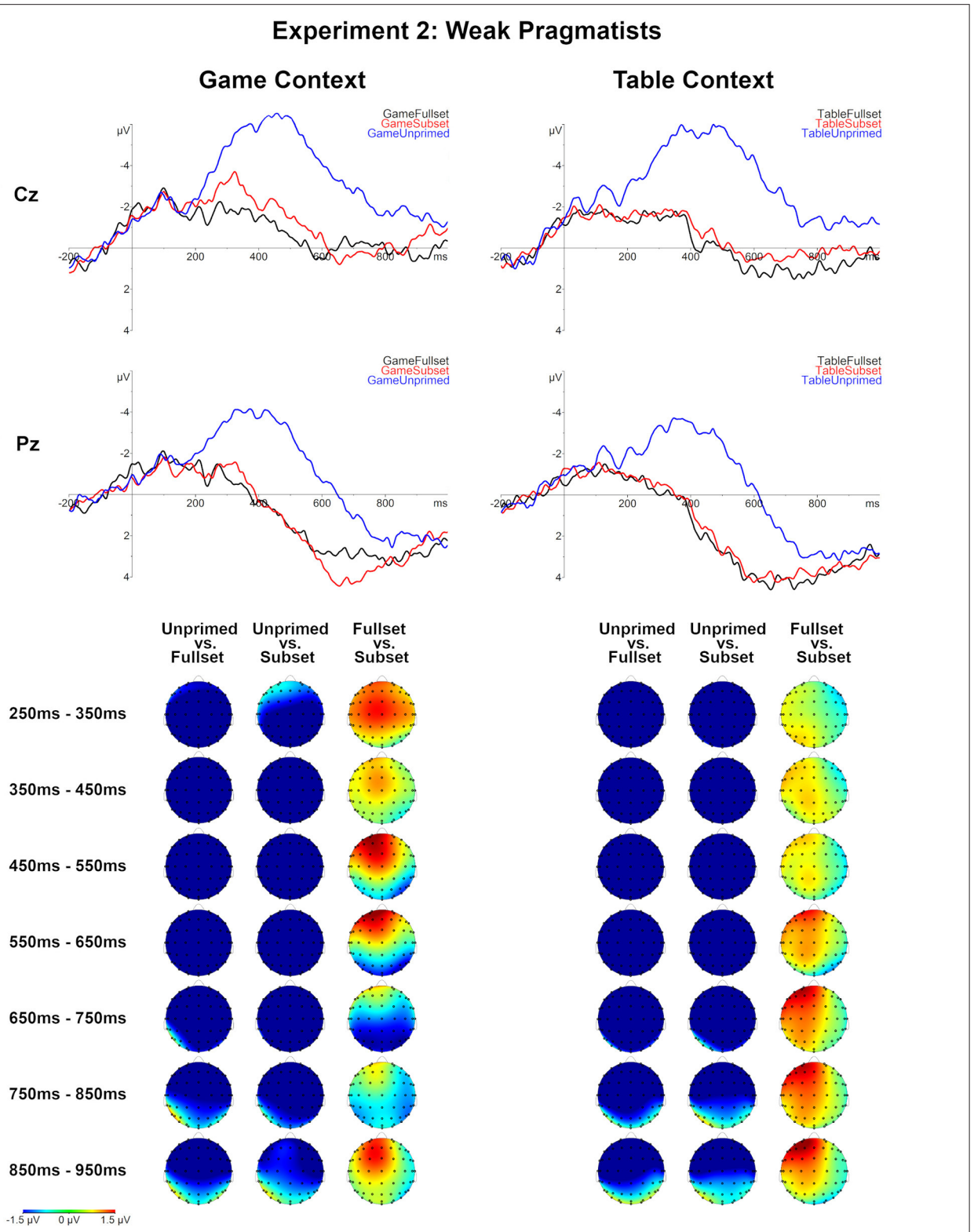

FIGURE 6 | Grand averages for all conditions for weak pragmatists ($N = 15$), Experiment 2. Topographical maps of the differences between the compared Set levels separately for the Table and Game context in consecutive 100 ms time-windows, between 250 and 950 ms post-onset.

true. However, this instruction did not seem to be sufficiently internalized. For some of the participants it was still not acceptable to "tell Lena" that she is wrong if they could not know it themselves. Other participants occasionally expressed doubt whether Lena had some hidden knowledge about the closed cards, for instance, she might have cheated. The reasons for such interpretations could be manifold, e.g., the former could be due to politeness or face-saving strategies (Brown and Levinson, 1987), the latter could be a case of epistemic vigilance as discussed by Sperber et al. (2010). In general, it appears that the inability to know the sentence's truth-value made some subjects more careful in rejecting the statement as inappropriate. This confusion in the task interpretation also suggests that a setting, where the epistemic access to the scenario is manipulated, requires more complex reasoning than a simple sentence-picture verification paradigm does.

The fact that our participants lacked any privileged access to the scenario might also explain the discrepancy between our results and those by Dieuleveut et al. (2019), who report a significant rate of strong pragmatic readings. Unlike in our experiment, their participants always had access to the whole information and, thus, they always knew whether all cards are hearts in the real world. This factor might have endorsed the egocentric perspective in interpreting pragmatically ambiguous statements, resulting in a higher rate of strong pragmatic readings. Moreover, their trials were constructed in such a way that the ignorant speaker was always contrasted with the knowledgeable one. This contrast might have also biased the subjects toward more accurate statements over pragmatically ambiguous ones ("Mary can't say that but Peter can"). Another important difference between their design and ours is the type of question used: *Can she say that?* vs. *Is it appropriate?* However, the instructions provided by Dieuleveut et al. (2019) and especially the feedback given to the exercises is very similar to ours, i.e., in their study it is explained that Mary (or John) can say a statement if they are sure that it is true, whereas the reasons for a negative response could be either that a statement is known to be false or that the speaker does not have enough information to know whether it is true or false. We introduced the same limitation to the interpretation of "appropriate" by instructing the subjects that statements were appropriate if the speaker knew that they were true, and inappropriate if they were visibly false or if the speaker lacked sufficient evidence. Thus, in principle, the indented interpretation of the question was very similar across the two studies, although one cannot exclude that a different formulation had an effect on the divergent response patterns.

4.2. Reaction Times Reveal a Processing Cost of the Primary Implicature as Well as a Cost of Processing Partial Access Scenarios

Multiple prior studies have demonstrated that SIs are computed at a significant cost, resulting in longer response times for the pragmatic relative to logical interpretation of *some*. In line with these results, in Experiment 1 we also observed that the weak pragmatic interpretation led to a significant delay in responses for the Table-Fullset condition. Given that the pragmatic responses in the Table-Fullset condition were also slower than the "inappropriate" judgments in the Table-Unprimed condition, this effect cannot be simply attributed to the type of judgment ("inappropriate") and may indicate a processing cost in deriving the primary implicature.

In addition, in Experiment 1, both the Game-Fullset and Game-Unprimed conditions showed slower responses relative to the Game-Subset condition. In the Game-Unprimed condition these slower responses could, in principle, be explained by assuming that the "inappropriate" judgments are more effortful than the "appropriate" ones. However, the response times in the Game-Unprimed condition proved not only longer relative to the other Game conditions but also relative to the Table-Unprimed condition, where the same type of behavioral judgments ("inappropriate") were given. As in the Table-Unprimed condition the target utterances were visibly false, whereas in the Game-Unprimed condition their truth-value was unknown, the significantly slower responses in the latter relative to the former condition seem to be driven by the increased costs related to the processing of the unknown semantic status of an utterance. By contrast, in the Game-Fullset condition both logicians and weak pragmatists gave "appropriate" judgments. Thus, the longer response times in this case relative to the Game-Subset condition can be explained in terms of some level of epistemic uncertainty related to the partial access scenario.

These response time results partially replicate in Experiment 2. In this case, only weak pragmatists are included in the analysis and we observe slower responses in the Fullset relative to Subset condition in both contexts. This supports the prior finding that the Fullset condition, in which the implicature has to be considered (and is derived in the full access but inhibited in the partial access context), leads to a delay in sentence evaluation.

4.3. ERP Results: Priming-Related Effect

The N400 effect observed for the Unprimed conditions relative to the Subset and Fullset conditions, for both context types, is consistent with the literature that predicts the modulation of the N400 by priming. In our study, the critical words in the Unprimed conditions were not presented in the respective scenarios, unlike in the Subset or Fullset conditions. As primed words tend to trigger smaller N400 ERPs relative to unprimed words, it is unsurprising that the Unprimed conditions showed an N400 effect relative to the other conditions. What is less expected is the lack of any ERP effect between the Game- and Table-Unprimed conditions, especially in the light of longer response time found for the former relative to the latter. One possibility is that the effect related to epistemic reasoning does not show in the early processing of naturally spoken sentences, and occurs at a later stage, when the behavioral response is given.

4.4. ERP Results: Shallow Processing of the Primary Implicature

The comparison between Table-Fullset and Table-Subset conditions allowed us to test the implicature processing in the full access context. Let us recall that in the study by Spychalska

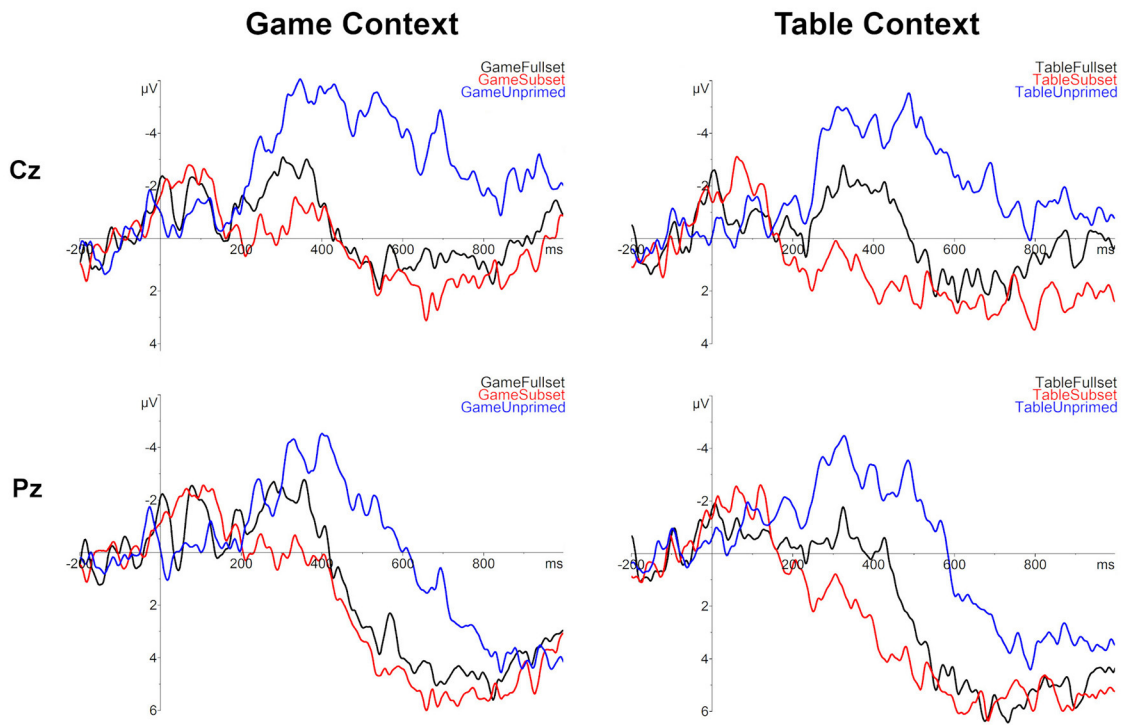

FIGURE 7 | The comparison of subject averages in all conditions for the strong pragmatists, including the subject who displayed a mixed weak/strong response pragmatic response pattern ($N = 4$).

et al. (2016), in a similar comparison, i.e., between *Some-infelicitous* and *Some-True* conditions, pragmatists showed a biphasic N400/P600 effect, whereas no effect was observed for logicians. In the current study, similar to the prior one, no effect was observed for logicians. Although, in Experiment 1, weak pragmatists showed a late posterior negativity effect for the Table-Fullset condition, starting after 600 ms post-stimulus onset, this effect did not replicate in Experiment 2. This result indicates that implicature was processed shallowly and with delay. The signature of the cost related to deriving the primary implicature could only be observed in longer response times in the Table-Fullset condition recorded for weak pragmatists.

4.5. ERP Results: The Cost of the Epistemic Step and the Role of Alternative Contexts

To test the effect of epistemic access for the processing of pragmatically ambiguous sentences with *some*, we compared Game-Fullset with Game-Subset conditions, as well as each of the Set conditions across the context type (Game vs. Table). We expected that the observed effects should be modulated by the quantifier reading, i.e., by the appropriateness judgments in the Table-Fullset and Game-Fullset conditions. For instance, in the case of strong pragmatic interpretation, an N400 effect was expected for the Game-Fullset relative to Game-Subset context. Unfortunately, as only three participants consistently adopted this reading (one in Experiment 1 and two in Experiment 2) and one subject in Experiment 2 partially adopted this reading, no statistical evaluation of this hypothesis was possible[8].

For weak pragmatists, the comparison between Game-Fullset and Game-Subset conditions showed a late, sustained negativity effect. This effect was significant in Experiment 1 and marginally significant in Experiment 2. The Game-Fullset condition also showed a late negativity effect relative to the Table-Fullset condition, which was marginally significant in Experiment 1 and significant in Experiment 2. It is important to note that there were no differences between the two experiments other than the method of subjects' selection: For the second experiment, the participants were pre-selected based on a short questionnaire that included one task corresponding to the Table-Fullset condition. This procedure was intended to reduce the number of unnecessary recordings (i.e., disregard those subjects who were likely to display a logical response pattern). Thus, the participants tested in Experiment 2 had seen a few questions similar to those used in the EEG experiment already before coming to the lab. Although it cannot be ruled out, it is unlikely that this pre-exposure to the experimental task would have had a

[8]Note that based on the visual inspection, the ERPs observed for these participants are consistent with this hypothesis (see **Figure 7**).

significant effect on the observed ERP patterns. However, due to the interruption of the testing caused by the pandemic outbreak we were able to record only 15 consistent weak pragmatists, which is also suboptimal from the point of view of statistical power. Given that the reported effects showed a consistent trend across experiments both with respect to the topography and latency, they are likely to reflect real effects and the marginally significant *p*-values are due to small sample sizes.

Late sustained negativities of posterior distribution were observed earlier in various linguistic contexts. For instance, Politzer-Ahles et al. (2012) observed an effect of this sort in response to pragmatically infelicitous quantifiers. In this study, sentences such as *In this picture, some of the girls are sitting on the blankets sunbathing* were read as descriptions of pictures that matched (with a subset of the girls on the blankets) or mismatched the interpretation with the scalar implicature (with all of the girls on the blankets). The ERPs measured at the onset of sentence-initial quantifiers showed a posterior negativity effect after 500 ms post-stimulus onset for the mismatching relative to the matching scenario. This effect was interpreted as a signature of "effortful pragmatic reanalysis," which the authors further suggested specifically involved "inhibiting the pragmatic reading of *some* of and retrieving the semantic reading." The effect was observed for the whole tested population without differentiating between pragmatic and logical responders. Acceptability judgments that would allow to distinguish between the pragmatic and logical interpretation followed only 6 trials per condition. Given that most of the participants were considered to be inconsistent responders, no between-group analysis was conducted.

Late posterior negativity (LPN) effects have also been observed in memory research: A large number of ERP studies on recognition memory observed that responses for old relative to new, correctly classified items, tend to trigger a sustained negativity effect over posterior sites beginning at approximately 600 or 800 ms post-onset and lasting till the end of the 1,000 ms or even 2,000 ms long epoch (for a review see Mecklinger et al., 2016). However, due to heterogeneity of experimental manipulations giving rise to this effect, it has been suggested that the LPN may not reflect a single process and is unlikely to constitute a direct correlate of episodic recollection, especially that this effect was shown to be sensitive not just to episodic but also to semantic memory tasks, that require more specific reconstructive processes (cf. Johansson and Mecklinger, 2003; Herron, 2007). In attempts to describe the functional sensitivity of LPN, many authors have linked it, inter alia, to late/extended retrieval processes (Dzulkifli and Wilding, 2005; Bergström et al., 2013), post-retrieval maintenance/evaluation of contextual information (Johnson et al., 2008), contextual familiarity (Addante et al., 2012), context monitoring and contextual retrieval of task relevant attributes (Goffaux et al., 2008), or enhanced need to monitor response conflict between suppression and automatic retrieval (Hu et al., 2015). As a common denominator of these proposal, the LPN seems to be related to late processes that have to do with (re)evaluation/monitoring of the context. Assuming a functional similarity between the late posterior negativity observed in the current experiments for weak pragmatists and the memory-related LPN, one could argue that in our studies this effect arises as a result of extended and possibly inference-driven context monitoring processes that may be related to reevaluation of the scenario, in particular, reconsidering the speaker's epistemic access. As weak pragmatists are sensitive to the implicature in the full access context, one can hypothesize that for this group the partial access context engages processes related to the evaluation of speaker competence assumption, and eventually to implicature inhibition.

By contrast, for logicians a differential pattern of effects is observed, namely, a robust N400-like negativity for the Game-Subset relative to Game-Fullset condition, as well as relative to the Table-Subset condition. At first, this result may appear puzzling but it can be explained by taking into account global pragmatic effects arising from the competition between two alternative contexts contrasted in the experiment. Based on the monotonicity properties of *some*, in the Subset condition, the Game restriction is somewhat less informative than the Table restriction, and in addition the Game restriction is more informative when used in the Fullset rather than in the Subset condition. Accordingly, this informativity relation was also expected to modulate predictive processes during sentence comprehension, leading to larger N400 ERPs for the less informative utterances. The effects observed for logicians are precisely in the line with this prediction. Although for weak pragmatists the N400-like effect for the Game-Subset vs. Table-Subset condition was also observed in Experiment 1, Experiment 2 did not show a similar, even marginally significant, effect for this group. Thus, the N400-like negativity in the Game-Subset condition appears to be primarily modulated under the logical interpretation. Such an outcome is expected if one observes that, under the logical interpretation, *some* may be considered the most "optimal" quantifier (out of all contextually provided in the experiment) to use in cases where insufficient information is provided about the domain, which is exactly the Game-Fullset context. Thus, in spite of equivalent appropriateness judgements in all of the Subset and Fullset conditions, under the logical interpretation, *some cards in the game* may be primarily used as means of expressing uncertainty about the whole domain, and consequently perceived as less optimal in the Subset condition, where *some cards on the table* would be more informative. Most importantly, logicians are defined as those participants who, based on their appropriateness judgments, appear not to have derived the primary implicature. This does not mean that they are insensitive to pragmatic mechanisms as such. In this case, we observe that the global pragmatic effects, which are based on contextually provided alternatives, played a primary role in modulating their processing pattern.

4.6. Response Times and ERPs Reveal Processes at Different Stages

It is noteworthy that some of the contrasts that showed significant reaction time differences did not yield significant ERP effects. For instance, the weak pragmatic response was associated with longer

response times in the Table-Fullset condition relative to Table-Subset, the late negativity ERP effect in this comparison was only observed in Experiment 1 and did not replicate in Experiment 2. This result suggests that the primary implicature processing did not occur incrementally but rather with delay, and only incurred cost at the stage of the behavioral judgment. In addition, the response times in condition Game-Unprimed turned out significantly longer than those in condition Table-Unprimed, no significant ERP effect is observed in this comparison, at least not in the analyzed epoch, i.e., up to the 1,000 ms post-stimulus onset. Given the natural and relatively fast pace of presenting the auditory stimuli, some of the processes related to epistemic reasoning might have been impossible to detect in the early time-window (up to 1,000 ms), when the ERPs are measured. These processes, however, still left a mark in the responses times, as the responses were given at a later stage.

4.7. Differences Between the Current and the Prior ERP Study on Scalar Implicatures

The results observed in the current study differ to a large extent from those observed in Spychalska et al. (2016), including a different distribution of behavioral responses (a lower proportion of pragmatists in the current study) and a different pattern of ERP results in the comparable conditions. The most striking result is the lack of any robust effect for weak pragmatists in the Table-Fullset vs. Table-Subset comparison. Let us recall that in a similar comparison in Spychalska et al. (2016) pragmatists had a combination of the N400 and the P600 effect. These differences in results may be linked to some important aspects of both designs, including the distribution of filler trials, the modality of the stimuli presentation (visual vs. auditory sentence presentation), the type of task and, finally, the very presence of the partial access scenarios in the current experimental setting.

First, although the current study used filler trials of a similar sort as the prior one, namely, trials with *all* and *no* as well as with numerals, the probability of encountering such items was lower compared to the prior study. In Spychalska et al. (2016), there was an equal number of trials with *all* and *some*, each of which constituted 40% of all trials, whereas the remaining 20% were fillers with such quantifiers as *no*, *most*, and with bare numerals. In the current study, due to the fact that we needed as many as six different conditions with *some*, as well as additional fillers with *some*, trials with *some* altogether constituted approximately 66% of all trials, whereas those with *all* only about 10% (10% of all trials were those with *no* and 14% were other fillers). Prior studies have shown that the type and proportion of filler items may have a significant effect on the time-course of scalar implicature processing: For instance, Dieussaert et al. (2011) showed that participants tended to be less consistent in the chosen logical or pragmatic interpretation if the filler ratio was higher. Degen and Tanenhaus (2015) showed that implicatures were processed more costly if other scalar terms such as number words were available in the context. Finally, Augurzky et al. (2019) observed that the contrast between *all* vs. *some* may prime the scalar implicature, more specifically, if such a contrast was not present in the context, the implicature was processed more shallowly. This last result appears to be particularly relevant for our study, namely, the lower proportion of trials with *all* in the current design could have contributed to some extent to the observed lower proportion of pragmatic responders as well as to the more shallow processing of the implicature. Still, the contrast with *all* was not absent in the design and the difference was only in the probability of such items. To evaluate the role of this factor we can refer to the study by Hunt et al. (2013), which used a very similar design to Spychalska et al. (2016), but a lower proportion of *all* vs. *some* trials: In this study there were 171 target trials with *some* (divided into three conditions: *true/false/infelicitous*) and additional 171 filler trials that were distributed between *all*, *no* and *some*. Although the precise proportion is not reported, it is clear that the probability of trials with *all* was much lower than those with *some*[9]. In spite of having a lower ratio of *all* vs. *some* items, Hunt et al. (2013) observed a very similar pattern of results as Spychalska et al. (2016), namely, a biphasic N400/P600 effect for the *infelicitous* relative to the *true* condition, that was only apparent for the pragmatic responders. Thus, although a diminution of the expected N400/P600 effect relative to the probability of *all* vs. *some* items seems plausible, given that the *all* items were still contextually active in the current experiment, the filler distribution is unlikely to explain the complete lack of the expected N400/P600 effect.

Second, as discussed earlier, the different form of the judgment task, namely, appropriateness rather than truth-value judgements, appears to have affected the interpretation of the utterances resulting in a lower proportion of pragmatic responders. This is evident from the oral feedback provided by the participants as well as from some of the misinterpretation cases of the partial access conditions. It is, however, unlikely that the effect of the question type reached beyond the distribution of the behavioral responses and also modulated the time-course of the implicature processing. Although the P600 effect is considered task-dependent, the N400 effect is generally taken to occur independently of the task. Thus, the different type of task should not prohibit the N400 from occurring in response to the condition inconsistent with the implicature.

The third important aspect is the modality in which the linguistic stimuli were presented, which also determined the time-course of the stimuli presentation. It is unlikely that the auditory vs. visual presentation of the sentence made a significant difference to the general pattern of the effects, since language-related ERP components, such as the N400 and the P600 are generally modality-independent (Osterhout and Holcomb, 1993; Kutas and Federmeier, 2011). Auditorily presented sentences tend to trigger P600 effects in response to syntactic violations similar to visually presented ones (Osterhout and Holcomb, 1993). N400 effects were also observed both for visual and auditory words, although some cross-modality differences in the time-course and topography of these effects have been observed: Auditory N400s tend to begin earlier, last longer, and have a slightly more frontal and less right-hemisphere biased topography (Kutas and Van Petten, 1994; Kutas and

[9] Assuming the likely even distribution of *all*/*some*/*no* across the filler trials, the proportion of *all* would be ca. 16% and the proportion of *some* approximately 66%.

Federmeier, 2011). Different patterns of priming-driven N400 effects between auditory and visual words have been reported by Holcomb et al. (1992) and Holcomb and Anderson (1993), who argued for a modality-specific modulation of the semantic processing system. Yet, these modality specific differences seem to concern only minor spatio-temporal variation in the component's characteristics, rather than the occurrence of the effects under similar experimental manipulations. Nevertheless, the auditory modality had also consequences for the time-course and pace of presenting the linguistic stimulus. In the study by Spychalska et al. (2016), the quantifying phrase was presented before the pictures, which means that during the inspection of the scenario the subjects could have already computed the implicature and shape their expectations regarding the sentence-final noun. From the quantifier onset until the critical word onset there were in total 5,000 ms. In the current study, the sentence was heard during the inspection of the scenario: The average sentence duration was approximately $2,502ms$ (2,702 ms was the average length of the audio file, which includes the approximate 200 ms of the pre-stimulus silence). From the onset of the sentence until the onset of critical word there were on average about 1,939 ms (based on the 2,139 average onset of the critical noun minus the 200 ms silence onset), whereas from the onset of the scenario until the onset of the critical noun there were on average 3,939 ms. If the implicature calculation is indeed an effortful process, then it is possible that in the current study there was not enough time for the implicature to be computed and, thus, it could not modulate the predictions for sentence-final words. To discuss the role of this factor, we can again refer to the study by Hunt et al. (2013) for a comparison. In this experiment, the scenario was presented first and consisted of two screens: the first screen with a set of items of various sorts, e.g., 4 steaks, 4 apples and 4 brownies, and the second screen, where some items from one category (e.g., steaks) and all items from another category (e.g., apples) were cut. Sentences of the form *The student has cut some of the apples in this story* were presented word-by-word after the scenario, using rapid serial visual presentation with 300 ms for each word and 200 ms between words. Thus, from the onset of the quantifier until the onset of the critical noun there was only a time span of about 1,500 ms. This would suggest that the N400 effect in response to the scenario-based implicature mismatch may also be observed for stimuli presented in a natural pace. Yet, one must also take into account that in Hunt et al. (2013) the scenario preceding the sentence presentation was shown for a total of 13,000 ms (the duration of the two screens was 7,500 and 5,500 ms respectively), i.e., much longer than in our current experiment. Given that in Hunt et al. (2013) the majority of trials used the quantifier *some*, and the set of all other quantifiers used was limited to *some*/*all*/*no*, one cannot exclude that some strategic anticipation of the nouns' match/mismatch in relation to each of the potential quantifiers happened already during the scenario inspection to facilitate the processing of the upcoming sentence.

The last factor to be considered is the presence of the partial access scenarios. This aspect of the design distinguishes the current study both from the study by Spychalska et al. (2016) and the one by Hunt et al. (2013). Thus, it is a likely candidate to explain the discrepancy between the current results and the prior ones. The competition between the two alternative contexts and the presence of the closed cards in the scenarios possibly triggered processes related to the reevaluation of the speaker epistemic access. This epistemic component led to non-incremental implicature processing: Primary implicatures were derived with delay and post-propositionally, which explains the absence of any clear ERP effect for weak pragmatists in the critical Table-Fullset vs. Table-Subset comparison. The LPN effect observed for this group in the partial access context (Game-Fullset vs. Subset) may indicate engagement of processes related to the increased context evaluation/monitoring. This effect may be explained by the hypothesis that weak pragmatists perform epistemic reasoning related to the evaluation of the competence assumption, which leads then to implicature inhibition in the partial access context. In the studies by Hunt et al. (2013) or Spychalska et al. (2016), where only full access scenarios were presented, such epistemic processes were not contextually induced. By contrast, the processing patterns of logicians appear to be modulated rather by global pragmatic mechanisms related to informativeness of each of the alternative quantifying expressions (*some cards on the table* vs. *some cards in the game*) as applied to the Fullset or Subset condition. Since, in this particular setting, *some cards in the game* is semantically weaker than *some cards on the table*, the use of the former expression may be considered less informative in the Subset condition. This effect is further strengthened if the logical reading is adopted, since in this case, *some* can be taken as the most optimal means of expressing uncertainty about the whole domain, resulting in the *some cards in the game* quantifying phrase being the most optimal one in the Fullset scenario.

4.8. Conclusion

Prior studies on the role of speaker competence assumptions in deriving scalar implicatures have been rather scarce and up to date there have been no ERP studies investigating the real-time processing of scalar implicatures in partial access contexts, i.e., contexts where the speaker's competence cannot be assumed. In this paper we present both behavioral and ERP data to fill this gap.

First, we observe a very low percentage of pragmatic responses in the full access contexts, where the speaker competence assumption holds. Thus, primary scalar implicatures were derived less frequently than in other experiments reported in the literature. Moreover, almost all those subjects who did derive the primary implicature did not derive the secondary implicature in the partial access context: Only three subjects in total applied the strong pragmatic interpretation (four if we include the one additional subject who was switching between the weak and strong interpretation). This result is striking as it indicates, on the one hand, that the strong pragmatic interpretation appears at best as an isolated response pattern, and on the other hand, that the presence of partial access contexts in the design suppresses the pragmatic interpretation

as such. This effect may be also due to the type of judgment task used in our study: Appropriateness judgments may have been interpreted less rigidly than truth-value judgments which led some subjects to treat sentences as appropriate in spite of the fact that they would not have evaluated them as true in the strict sense. It is also interesting that partial access contexts showed problematic also in other cases: In the Game-Unprimed condition, where the sentence semantic status is unknown, less accurate judgments were observed leading to a number of systematic misinterpretation cases. These results also show that speaker epistemic status manipulations are experimentally problematic, possibly since they require from the participants to perform belief reasoning and to represent the presumed belief state of the virtual agent. This may lead to a number of errors both in the sense of varied interpretations of what the indented (in the experiment) agent's belief state is as well as errors of misidentification of one's own belief with that of the agent.

As the most important result we showed that partial access contexts involve a processing cost, which left mark both in the accuracy, response times and elicited ERPs. The condition Game-Unprimed, which involves the highest level of epistemic uncertainty level, was associated not only with the highest level of response errors but also with longer response times both relative to Game-Subset and Game-Fullest conditions, as well as relative to the Table-Unprimed condition. This result was comparable for both logicians and weak pragmatists. Although no ERP effect was observed when comparing the Table- and Game-Unprimed conditions, the difference in accuracy and response times is interpreted as evidence of increased cognitive demands related to epistemic uncertainty in the partial access context. Furthermore, our results suggest that, in the current experimental setting, deriving implicatures was cognitively costly: Weak pragmatist responded slower in the Table-Fullset condition relative to other conditions but showed no robust ERP signature of implicature processing. This indicates that deriving the primary implicature occurred not incrementally and late. Longer response times are also observed for weak pragmatists in the Game-Fullset relative to the Game-Subset condition. In addition, the Game-Fullset condition shows a late posterior negativity effect relative to the Game-Subset and Table-Fullset conditions, which we interpret as a signature of epistemic context-reevaluation that led to implicature inhibition. Due to the small number of weak pragmatists in both experiments, some of these effects are marginally significant and, hence, should be treated with caution. However, as we can see a consistent trend between the two experiments, the observed patterns are likely to reveal real effects. We argue that the observed processing patterns are inherently related to the contrast between the partial and full access contexts present in the experiment. In the case of weak pragmatists, who are sensitive to the implicature at the quantifier level, this contrast leads to non-incremental implicature processing and epistemic context-reevaluation. For logicians, who do not derive the primary implicature, the processing patterns are primarily modulated by the informativity relation between the two domain restrictions.

To sum up, our experiment shows that if the general context raises the question of whether or not the speaker has sufficient information to make the statement, the implicatures are processed as postpropositional inferences rather than as automatic and incremental. Thus, the contrast between full and partial access contexts seems to enforce on the listeners, at least on those who choose the pragmatic interpretation, taking the "epistemic step": reconsidering whether or not the speaker is epistemically competent. Although, our conclusions are not claimed to provide any definite answer in the debate, our results appear more in line with the traditional, pragmatic account rather than with the grammatical view on scalar implicatures.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

MS designed and programmed the experiment, supervised the stimuli preparation and the data collection, pre-processed and analyzed the data, and wrote the manuscript. LR contributed to the design, prepared the visual and auditory stimuli, collected the data, and contributed to the manuscript preparation. PS provided the infrastructure for the auditory stimuli recording, contributed to the design, and results interpretation. MW provided the infrastructure and funding for the EEG recording, contributed to the design, and results interpretation. All authors contributed to the article and approved the submitted version.

FUNDING

The research for this paper has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) with the following grants (indicated by project numbers): 366639214 (PI: MS), 254822891 (PI: PS), 367110651 (PI: MW), within the Priority Program XPrag.de (SPP1727); as well as 274877981, LING 2.1 (PI: MW; LR) within the RTG Situated Cognition.

ACKNOWLEDGMENTS

We acknowledged support by the Open Access Publication Funds of the Ruhr-Universität Bochum. We are grateful for the support during stimulus recording from the Institute of Linguistics at the University of Cologne. We would also like

to thank Lena Straßburger for assistance in the recordings and Janina Kalbertodt for assistance and advise during the recording and preprocessing of the acoustic stimuli. We are also grateful to the reviewers for their valuable feedback on the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.679491/full#supplementary-material

REFERENCES

Addante, R. J., Ranganath, C., and Yonelinas, A. P. (2012). Examining ERP correlates of recognition memory: evidence of accurate source recognition without recollection. *Neuroimage* 62, 439–450. doi: 10.1016/j.neuroimage.2012.04.031

Augurzky, P., Franke, M., and Ulrich, R. (2019). Gricean expectations in online sentence comprehension: an ERP study on the processing of scalar inferences. *Cogn. Sci.* 43:e12776. doi: 10.1111/cogs.12776

Bergen, L., and Grodner, D. (2012). Speaker knowledge influences the comprehension of pragmatic inferences. *J. Exp. Psychol. Learn. Mem. Cogn.* 38, 1450–1460. doi: 10.1037/a0027850

Bergen, L., Levy, R., and Goodman, N. (2016). Pragmatic reasoning through semantic inference. *Semant. Pragmat.* 9. doi: 10.3765/sp.9.20

Bergström, Z. M., Henson, R. N., Taylor, J. R., and Simons, J. S. (2013). Multimodal imaging reveals the spatiotemporal dynamics of recollection. *Neuroimage* 68, 141–153. doi: 10.1016/j.neuroimage.2012.11.030

Bornkessel-Schlesewsky, I., and Schlesewsky, M. (2008). An alternative perspective on "semantic P600" effects in language comprehension. *Brain Res. Rev.* 59, 55–73. doi: 10.1016/j.brainresrev.2008.05.003

Bott, L., and Noveck, I. A. (2004). Some utterances are underinformative: the onset and time course of scalar implicatures. *J. Mem. Lang.* 51, 437–457. doi: 10.1016/j.jml.2004.05.006

Breheny, R., Ferguson, H. J., and Katsos, N. (2013). Taking the epistemic step: toward a model of on-line access to conversational implicatures. *Cognition* 126, 423–440. doi: 10.1016/j.cognition.2012.11.012

Brouwer, H., Fitz, H., and Hoeks, J. (2012). Getting real about semantic illusions: rethinking the functional role of the p600 in language comprehension. *Brain Res.* 29, 127–143. doi: 10.1016/j.brainres.2012.01.055

Brown, P., and Levinson, S. C. (1987). *Politeness: Some Universals in Language Usage. Studies in interactional sociolinguistics*. Cambridge: Cambridge University Press.

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

Chevallier, C., Bonnefond, M., Van der Henst, J.-B., and Noveck, I. A. (2010). Using ERP's to capture inferential processes guided by prosodic cues. *Ital. J. Linguist.* 22, 125–152.

Chierchia, G. (2004). "Scalar implicatures, polarity phenomena, and the syntax/pragmatics interface," in *Structures and beyond,* ed A. Belletti, Vol. 3, (Oxford: Oxford University Press), 39–103.

Chierchia, G., Fox, D., and Spector, B. (2011). "Scalar implicature as a grammatical phenomenon," in *Semantics: An International Handbook of Natural Language Meaning,* Vol. 3, eds P. Portner, C. Maienborn, and K. von Heusinger (Berlin; Boston, MA: Mouton de Gruyter), 2297–2331.

Degen, J., and Tanenhaus, M. K. (2014). Processing scalar implicature: a constraint-based approach. *Cogn. Sci.* 39, 667–710. doi: 10.1111/cogs.12171

Degen, J., and Tanenhaus, M. K. (2015). Availability of alternatives and the processing of scalar implicatures: a visual world eye-tracking study. *Cogn. Sci.* 40, 172–201. doi: 10.1111/cogs.12227

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

Dieuleveut, A., Chemla, E., and Spector, B. (2019). Distinctions between primary and secondary scalar implicatures. *J. Mem. Lang.* 106, 150–171. doi: 10.1016/j.jml.2019.02.008

Dieussaert, K., Verkerk, S., Gillard, E., and Schaeken, W. (2011). Some effort for some: further evidence that scalar implicatures are effortful. *Q. J. Exp. Psychol.* 64, 2352–2367. doi: 10.1080/17470218.2011.588799

Dzulkifli, M. A., and Wilding, E. L. (2005). Electrophysiological indices of strategic episodic retrieval processing. *Neuropsychologia* 43, 1152–1162. doi: 10.1016/j.neuropsychologia.2004.11.019

Federmeier, K., Wlotko, E., De Ochoa-Dewald, E., and Kutas, M. (2007). Multiple effects of sentential constraint on word processing. *Brain Res.* 18, 75–84. doi: 10.1016/j.brainres.2006.06.101

Fox, D. (2014). Cancelling the maxim of quantity: another challenge for a Gricean theory of scalar implicatures. *Semant. Pragmat.* 7, 5–1. doi: 10.3765/sp.7.5

Gazdar, G. (1979). *Pragmatics: Implicature, presupposition, and logical form*. New York, NY; Academic Press.

Geurts, B. (2010). *Quantity Implicatures*. Cambridge; New York, NY: Cambridge University Press.

Goffaux, P., Phillips, N., Sinai, M., and Pushkar, D. (2008). Neurophysiological measures of task-set switching: effects of working memory and aging. *J. Gerontol. Ser. B Psychol. Sci. Soc. Sci.* 63, P57–P66. doi: 10.1093/geronb/63.2.P57

Goodman, N. D., and Stuhlmüller, A. (2013). Knowledge and implicature: modeling language understanding as social cognition. *Top. Cogn. Sci.* 5, 173–184. doi: 10.1111/tops.12007

Grodner, D. J., Klein, N. M., Carbary, K. M., and Tanenhaus, M. K. (2010). "Some," and possibly all, scalar inferences are not delayed: evidence for immediate pragmatic enrichment. *Cognition* 116, 42–55. doi: 10.1016/j.cognition.2010.03.014

Hagoort, P., Brown, C. M., and Groothusen, J. (1993). The syntactic positive shift (SPS) as an ERP measure of syntactic processing. *Lang. Cogn. Processes* 8, 439–483. doi: 10.1080/01690969308407585

Hagoort, P., Hald, L. A., Bastiaansen, M. C. 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

Hanna, J. E., Tanenhaus, M. K., and Trueswell, J. C. (2003). The effects of common ground and perspective on domains of referential interpretation. *J. Mem. Lang.* 49, 43–61. doi: 10.1016/S0749-596X(03)00022-6

Hartshorne, J. K., Snedeker, J., Azar, S. Y.-M. L., and Kim, A. E. (2015). The neural computation of scalar implicature. *Lang. Cogn. Neurosci.* 30, 620–634. doi: 10.1080/23273798.2014.981195

Herron, J. E. (2007). Decomposition of the ERP late posterior negativity: effects of retrieval and response fluency. *Psychophysiology* 44, 233–244. doi: 10.1111/j.1469-8986.2006.00489.x

Hoeks, J. C., Stowe, L. A., and Doedens, G. (2004). Seeing words in context: the interaction of lexical and sentence level information during reading. *Cogn. Brain Res.* 19, 59–73. doi: 10.1016/j.cogbrainres.2003.10.022

Holcomb, P. J., and Anderson, J. E. (1993). Cross-modal semantic priming: a time-course analysis using event-related brain potentials. *Lang. Cogn. Processes* 8, 379–411. doi: 10.1080/01690969308407583

Holcomb, P. J., Coffey, S. A., and Neville, H. J. (1992). Visual and auditory sentence processing: a developmental analysis using event-related brain potentials. *Deve. Neuropsychol.* 8, 203–241. doi: 10.1080/87565649209540525

Horn, L. R. (1972). *On the Semantics of Logical Operators in English*. Ph.D. thesis, UCLA.

Horn, L. R. (1989). *A Natural History of Negation*. Chicago: University of Chicago Press.

Hu, X., Bergström, Z. M., Bodenhausen, G. V., and Rosenfeld, J. P. (2015). Suppressing unwanted autobiographical memories reduces their automatic influences: evidence from electrophysiology and an implicit autobiographical memory test. *Psychol. Sci.* 26, 1098–1106. doi: 10.1177/0956797615575734

Huang, Y. T., and Snedeker, J. (2009). Online interpretation of scalar quantifiers: insight into the semantic-pragmatics interface. *Cogn. Psychol.* 58, 376–415. doi: 10.1016/j.cogpsych.2008.09.001

Hunt, L., Politzer-Ahles, S., Gibson, L., Minai, U., and Fiorentino, R. (2013). Pragmatic inferences modulate N400 during sentence comprehension:

evidence from picture-sentence verification. *Neurosci. Lett.* 534, 246–251. doi: 10.1016/j.neulet.2012.11.044

Johansson, M., and Mecklinger, A. (2003). The late posterior negativity in ERP studies of episodic memory: action monitoring and retrieval of attribute conjunctions. *Biol. Psychol.* 64, 91–117. doi: 10.1016/S0301-0511(03)00104-2

Johnson, J. D., Minton, B. R., and Rugg, M. D. (2008). Content dependence of the electrophysiological correlates of recollection. *Neuroimage* 39, 406–416. doi: 10.1016/j.neuroimage.2007.08.050

Keysar, B., Barr, D. J., Balin, J. A., and Brauner, J. S. (2000). Taking perspective in conversation: the role of mutual knowledge in comprehension. *Psychol. Sci.* 11, 32–38. doi: 10.1111/1467-9280.00211

Keysar, B., Lin, S., and Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition* 89, 25–41. doi: 10.1016/S0010-0277(03)00064-7

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

Kuperberg, G. R., and Jaeger, T. F. (2015). What do we mean by prediction in language comprehension? *Lang Cogn. Neurosci.* 31, 32–59. doi: 10.1080/23273798.2015.1102299

Kuperberg, G. R., Sitnikova, T., Caplan, D., and Holcomb, P. J. (2003). Electrophisiological distinction in processing conceptual relationship within simple sentences. *Cogn. Brain Res.h* 17, 117–129. doi: 10.1016/S0926-6410(03)00086-7

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 (ERP). *Ann. 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 Van Petten, C. (1994). "Psycholinguistics electrified: event-related potential investigations," in *Handbook of Psycholinguistics*, ed M. A. Gernsbacher (San Diego, CA: Academic Press), 83–143.

Lau, E. F., Holcomb, P. J., and Kuperberg, G. R. (2013). Dissociating N400 effects of prediction from association in single word contexts. *J. Cogn. Neurosci.* 25, 484–502. doi: 10.1162/jocn_a_00328

Maris, E., and Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *J. of Neurosci. Methods* 164, 177–190. doi: 10.1016/j.jneumeth.2007.03.024

Mecklinger, A., Rosburg, T., and Johansson, M. (2016). Reconstructing the past: the late posterior negativity (LPN) in episodic memory studies. *Neurosci. Biobehav. Rev.* 68, 621–638. doi: 10.1016/j.neubiorev.2016.06.024

Nadig, A. S., and Sedivy, J. C. (2002). Evidence of perspective-taking constraints in children's on-line reference resolution. *Psychol. Sci.* 13, 329–336. doi: 10.1111/j.0956-7976.2002.00460.x

Nieuwland, M. S. (2016). Quantification, prediction and the online impact of sentence truth-value: evidence from event-related potentials. *J. Exp. Psychol. Learn. Mem. Cogn.* 42, 316–334. doi: 10.1037/xlm0000173

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

Nieuwland, M. S., and Kuperberg, G. R. (2008). When the truth isn't too hard to handle: an event-related potential study on the pragmatics of negation. *Psychol. Sci.* 19, 1213–1218. doi: 10.1111/j.1467-9280.2008.02226.x

Nieuwland, M. S., and Martin, A. E. (2012). If the real world were irrelevant, so to speak: the role of propositional truth-value in counterfactual sentence comprehension. *Cognition* 122, 102–109. doi: 10.1016/j.cognition.2011.09.001

Nieuwland, M. S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., et al. (2018). Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *eLife* 7:e33468. doi: 10.7554/eLife.33468

Noveck, I. A., and Posada, A. (2003). Characterizing the time course of an implicature: an evoked potentials study. *Brain Lang.* 85, 203–210. doi: 10.1016/S0093-934X(03)00053-1

Noveck, I. A., and Reboul, A. (2008). Experimental pragmatics: a gricean turn in the study of language. *Trends Cogn. Sci.* 12, 425–341. doi: 10.1016/j.tics.2008.07.009

Oostenveld, R., Fries, P., Maris, E., and Schoffelen, J. (2011). FieldTrip: open source software for advanced analysis of MEG, EEG and invasive electrophysiological data. *Comput. Intell. Neurosci.* 2011:156869. doi: 10.1155/2011/156869

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 Holcomb, P. J. (1993). Event-related potentials and syntactic anomaly: evidence of anomaly detection during the perception of continuous speech. *Lang. Cogn. Processes* 8, 413–437. doi: 10.1080/01690969308407584

Osterhout, L., Holcomb, P. J., and Swinney, D. A. (1994). Brain potentials elicited by garden-path sentences: evidence of the application of verb information during parsing. *J. Exp. Psychol. Learn. Mem. Cogn.* 20, 786–803. doi: 10.1037/0278-7393.20.4.786

Politzer-Ahles, S., and Fiorentino, R. (2013). The realization of scalar inferences: context sensitivity without processing cost. *PLoS ONE* 8:e63943. doi: 10.1371/journal.pone.0063943

Politzer-Ahles, S., Fiorentino, R., Jiang, X., and Zhou, X. (2012). Distinct neural correlates for pragmatic and semantic meaning processing: an event-related potential investigation of scalar implicature processing using picture-sentence verification. *Brain Res.* 1490, 134–152. doi: 10.1016/j.brainres.2012.10.042

Rabovsky, M., Hansen, S. S., and McClelland, J. L. (2018). Modelling the N400 brain potential as change in a probabilistic representation of meaning. *Nat. Hum. Behav.* 2, 693–705. doi: 10.1038/s41562-018-0406-4

Regel, S., Gunter, T., and Friederici, A. (2011). Isn't it ironic? an electrophysiological exploration of figurative language processing. *J. Cogn. Neurosci.* 23, 277–293. doi: 10.1162/jocn.2010.21411

Sauerland, U. (2004). Scalar implicatures in complex sentences. *Linguist. Philosophy* 27, 367–391. doi: 10.1023/B:LING.0000023378.71748.db

Schulz, K., and Van Rooij, R. (2006). Pragmatic meaning and non-monotonic reasoning: the case of exhaustive interpretation. *Linguist. Philosophy* 29, 205–250. doi: 10.1007/s10988-005-3760-4

Schumacher, P. B. (2013). When combinatorial processing results in reconceptualization: towards a new approach of compositionality. *Front. Psychol.* 4:677. doi: 10.3389/fpsyg.2013.00677

Schumacher, P. B. (2014). Content and context in incremental processing: "the Ham Sandwich" revisited. *Philos. Stud.* 168, 151–165. doi: 10.1007/s11098-013-0179-6

Sedivy, J. C. (2003). Pragmatic versus form-based accounts of referential contrast: evidence for effects of informativity expectations. *J. Psycholinguist. Res.* 32, 3–23. doi: 10.1023/A:1021928914454

Sperber, D., Clment, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., et al. (2010). Epistemic vigilance. *Mind Lang.* 25, 359–393. doi: 10.1111/j.1468-0017.2010.01394.x

Spychalska, M., Kontinen, J., and Werning, M. (2016). Investigating scalar implicatures in a truth-value judgment task: evidence from event-related brain potentials. *Lang. Cogn. Neurosci.* 31, 817–840. doi: 10.1080/23273798.2016.1161806

Swaab, T. Y., Ledoux, K., Camblin, C. C., and Boudewyn, M. A. (2012). "Language-related ERP components," in *The Oxford Handbook of Event-Related Potential Components*, eds S. J. Luck and E. S. Kappenman (Oxford University Press), 397–440.

Van den Noort, M., Bosch, P., Haverkort, M., and Hugdahl, K. (2008). A standard computerized version of the reading span test in different languages. *Eur. J. Psychol. Assess.* 24, 35–42. doi: 10.1027/1015-5759.24.1.35

Van Herten, M., Kolk, H., and Chwilla, D. (2005). An ERP study of P600 effect elicited by semantic anomalities. *Cogn. Brain Res.* 22, 241–255. doi: 10.1016/j.cogbrainres.2004.09.002

Conflict of Interest: 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 © 2021 Spychalska, Reimer, Schumacher and Werning. 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) and the copyright owner(s) 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 August 2021
doi: 10.3389/fpsyg.2021.629177

Processing Non-at-Issue Meanings of Conditional Connectives: The *wenn/falls* Contrast in German

*Mingya Liu**

Department of English and American Studies, Humboldt University of Berlin, Berlin, Germany

Logical connectives in natural language pose challenges to truth-conditional semantics due to pragmatics and gradience in their meaning. This paper reports on a case study of the conditional connectives (CCs) *wenn/falls* 'if/when, if/in case' in German. Using distributional evidence, I argue that *wenn* and *falls* differ in lexical pragmatics: They express different degrees of speaker commitment (i.e., credence) toward the modified antecedent proposition at the non-at-issue dimension. This contrast can be modeled using the speaker commitment scale (Giannakidou and Mari, 2016), i.e., More committed <WENN p, FALLS p> Less committed. Four experiments are reported which tested the *wenn/falls* contrast, as well as the summary of an additional one from Liu (2019). Experiment 1 tested the naturalness of sentences containing the CCs (*wenn* or *falls*) and conditional antecedents with varying degrees of likelihood (very likely/likely/unlikely). The starting prediction was that *falls* might be degraded in combination with very likely and likely events in comparison to the other conditions, which was not borne out. Experiment 2 used the forced lexical choice paradigm, testing the choice between *wenn* and *falls* in the doxastic agent's conditional thought, depending on their belief or disbelief in the antecedent. The finding was that subjects chose *falls* significantly more often than *wenn* in the disbelief-context, and vice versa in the belief-context. Experiment 3 tested the naturalness of sentences with CCs and an additional relative clause conveying the speaker's belief or disbelief in the antecedent. An interaction was found: While in the belief-context, *wenn* was rated more natural than *falls,* the reverse pattern was found in the disbelief-context. While the results are mixed, the combination of the findings in Experiment 2, Experiment 3 and that of Experiment 4a from Liu (2019) that *falls* led to lower speaker commitment ratings than *wenn,* provide evidence for the CC scale. Experiment 4b tested the interaction between two speaker commitment scales, namely, one of connectives (including *weil* 'because' and *wenn/falls*) and the other of adverbs (factive vs. non-factive, Liu, 2012). While factive and non-factive adverbs were rated equally natural for the factive causal connective, non-factive adverbs were preferred over factive ones by both CCs, with no difference between *wenn* and *falls*. This is discussed together with the result in Liu (2019), where the *wenn/falls* difference occurred in the absence of negative polarity items (NPIs), but disappeared in the presence of NPIs. This raises further questions on how different speaker commitment scales interact and why.

Keywords: conditional connectives, German, experiment, speaker commitment, non-at-issue meaning

OPEN ACCESS

Edited by:
Yury Y. Shtyrov,
Aarhus University, Denmark

Reviewed by:
Karolina Krzyżanowska,
University of Amsterdam, Netherlands
Mailin Antomo,
University of Göttingen, Germany
Laia Mayol,
Pompeu Fabra University, Spain
Jacqueline Visconti,
University of Genoa, Italy

***Correspondence:**
Mingya Liu
mingya.liu@hu-berlin.de

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

Received: *13 November 2020*
Accepted: *23 June 2021*
Published: *10 August 2021*

Citation:
Liu M (2021) Processing Non-at-Issue Meanings of Conditional Connectives: The wenn/falls Contrast in German. Front. Psychol. 12:629177. doi: 10.3389/fpsyg.2021.629177

INTRODUCTION

Attitudinal expressions conveying speaker's beliefs or preferences are pervasive in natural language and communication. However, the related expressions can pose challenges to formal theories of grammar due to pragmatics (e.g., multidimensionality, context-dependence, and subjectivity) and gradience. Their formal modeling presupposes an empirically adequate characterization, for which experimental methods are useful, and sometimes, indispensable. This paper reports on a case study of German conditional connectives (CCs), as those in (1)[1]. While conditionals are one of the most studied topics in cognitive science and linguistics, CCs have drawn attention to a much lesser extent than the other related lexical and grammatical devices. In the formal semantic literature, CCs as the English *if* are claimed to have no semantics in Kratzer's (1991) restrictor analysis of conditionals. The existing vast linguistic literature on the interpretation of conditionals (to just name a few, e.g., Iatridou, 1991; von Fintel, 1999, 2007, 2011; Arregui, 2005; Grosz, 2012; Elder and Jaszczolt, 2016) shows effects of various factors (tense, mood, and polarity items) on the interpretation of conditionals, as well as the effect of CCs (e.g., Dostie, 1987; Léard, 1987 on CCs in French, Montoliìo, 2000; Schwenter, 2001 on CCs in Spanish, Ippolito and Su, 2014 on the Mandarin counterfactual CC *yaobushi* 'if-not', Hoeksema, 2012 on *unless* and among many others, also Declerck and Reed, 2001 on a comprehensive analysis of conditionals in English and Breindl et al., 2014 on connectives in German).

(1) Examples of German CCs[2]:
wenn 'when, if', *falls* 'if, in case,' *insofern/sofern* 'provided that,' *vorausgesetzt dass/unter der Voraussetzung, dass* 'provided that,' *angenommen dass* 'assuming that,' *im Falle, dass/für den Fall, dass/gesetzt (den Fall), dass* 'in the event that,' *unter der Bedingung, dass* 'on condition that,' *unter der Annahme, dass* 'assuming that,' *ausgenommen, dass* 'except if'

As is known from the literature, conditionals are non-veridical (Giannakidou, 1998, 1999), that is, *if*-clauses do not entail the truth of the antecedent proposition. In addition, the literature also shows that the non-veridicality property of conditionals can be influenced by various factors. The first, and probably most studied, is tense and mood choice, which reflects subjective (non-veridical) judgments. Conditionals in languages with tense and mood morphology come in two sorts: indicative and subjunctive. While the former is non-veridical, the latter is antiveridical, i.e., it presupposes (or implicates) the falsity of the antecedent proposition. That is, in (2a) the speaker does not know if John gets a promotion or not, but in (2b) the speaker presupposes that John did not get a promotion[3].

(2) (a) *If John gets a promotion, he will do a big party.*
(b) *If John had got a promotion, he would have done a big party.*

CCs, just as tense or mood choice, can reflect the speaker's doxastic assumptions at semantic and pragmatic levels. In this paper, I will use distributional and experimental evidence to argue that apparently similar CCs differ in lexical pragmatics (see Visconti, 1996 on CCs in Italian; Liu, 2019; Liu and Wang, 2021 on CCs in Mandarin)[4]. More specifically, they can express different degrees of credence toward the modified proposition. The meaning difference between various CCs in this regard can be formally modeled using speaker commitment[5] scales (Giannakidou and Mari, 2016) and as non-at-issue meanings (Simons et al., 2010) or, more precisely, an implicature resulting from the lexical choice between similar CCs. The paper focuses on the German CCs *wenn* vs. *falls*. It is organized as follows: Section "Non-at-Issue Meanings of *wenn/falls* in German" presents the distributional properties of *wenn/falls,* and provides an analysis relating *falls* to a weakened speaker commitment in contrast to *wenn*. Section "Experiments" reports on four experiments testing the analysis. Section "General Discussion and Conclusion" discusses the results and concludes the paper.

NON-AT-ISSUE MEANINGS OF *wenn/falls* IN GERMAN

In German, *wenn* is a more frequent word than *falls*[6], but researchers do not have a consensus regarding the question whether *wenn* or *falls* is the prototypical CC. The handbook of Breindl et al. (2014) contains a comprehensive description of the German CCs in comparison to one another and also to other connectives. I will not go through the entire list, which also includes the discussion of *wenn/falls*-complement clauses, irrelevance conditionals (*selbst/auch wenn/*falls* 'even if') and except-conditionals (*außer wenn/falls* 'except if'). The authors

[1]These expressions have different morphosyntactic properties (see Iatridou, 1991; Haegeman, 2003; Bhatt and Pancheva, 2006, a.o.). I put them under the unifying category of CC used in a semantic sense, due to the focus of the paper on the semantics and pragmatics of CCs and conditionals.

[2]The examples used throughout the paper do not reflect the author's personal opinions. Instead, they are only used to discuss grammar and linguistics.

[3]It is worth noting that subjunctive conditionals do not always presuppose or implicate the falsity of the antecedent proposition, and thus, are not always counterfactual (see Anderson, 1951; Iatridou, 2000; Starr, 2019). On the other hand, Arregui and Biezma (2016) argue that the counterfactuality implicature cannot be canceled without good reason.

[4]The simple and complex CCs (used in purely descriptive terms) in natural language can differ in terms of among others, biconditionality, see Montoliìo (2000) on Spanish *a condición de* 'on condition that' and Liu and Barthel (2021) on German *nur wenn* 'only if'.

[5]Please note that the notion of commitment used in the paper is linked to the degree of the speaker's belief or credence in a proposition, as conveyed by the modifying expressions. This is different from the notion used in, for example, Krifka (2015) or Geurts (2019) for modeling human communication.

[6]The *Wortschatz* corpora (http://corpora.uni-leipzig.de, accessed October 27, 2020) of Leipzig University show that *wenn* has the frequency class 5, *falls* 10. The frequency class of a word FC(w) is calculated based on the occurrence frequency of the word "Fw" in comparison to the frequency of the most frequent word "Fmax," in this case, *der* 'the$_{MASC}$,' and is defined as FC(w) = [log[2](Fmax/Fw)]. The higher the frequency class, the rarer the word.

also discuss the availability of causal and concessive readings for *wenn* but not for *falls*, which I will not deal with in this paper as the semantic or pragmatic status of the causal inference in indicative conditionals is debatable (see Volodina, 2006, 2011; Krzyżanowska et al., 2017; Krzyżanowska, 2019; Skovgaard-Olsen et al., 2016), as well as that of the concessive reading. In a nutshell, syntactically speaking, *wenn* and *falls*, by and large, have similar distributions in terms of syntactic positions where they can occur, but there is a preference for *wenn* over *falls* in adverbial clauses in a sentence-final position (Breindl et al., 2014). Semantically, the most prominent difference between them lies in that *wenn* has both a conditional and a temporal reading[7] whereas *falls* only has a conditional reading, which makes the use of the latter more restricted. Furthermore, Volodina (2006) relates their meaning differences to factivity and specificity: the non-factive specific use of *wenn* gives rise to ambiguity between a temporal (similar to *sobald* 'as soon as') and a conditional reading (similar to *falls*), see (3a); a non-factive generic use is possible for *wenn* but *falls* only allows a specific use, see (3b).

(3) (a) *Wenn/Falls der Regen aufhört, gehen wir hin.*
when/if the rain stops, go we there
'If/When the rain stops, we will go there.'
[conditional: *wenn/falls*, temporal: *wenn/*falls*]

(b) *Wenn/Falls es wärmer wird, schmilzt das Eis.*
if it warmer gets melts the ice
'If it gets warmer, the ice melts/will melt.'
[specific: *wenn/falls*, generic: *wenn/*falls*]

In the following of this section, I will present additional distributional properties of *wenn* and *falls* in different kinds of conditionals (von Fintel, 2007, 2011) and provide an analysis capturing their contrast.

Distribution of *wenn* vs. *falls*

Both *wenn* and *falls* are fine in indicative conditionals (3) and biscuit conditionals (which assert the consequent proposition with no conditional dependence on the antecedent), see (4).

(4) *Wenn/Falls du Hunger hast, gibt es Kekse*
when/if you hunger have, gives it biscuits
auf dem Tisch.
on the table
'If you are hungry, there are biscuits on the table.'

But they differ, among others, in the following aspects. The first contrast (Contrast 1) is that *wenn* can, but *falls* cannot, be used in premise conditionals, such as in (5), which presupposes that someone other than the speaker, in this case A, believes the truth of the antecedent proposition (Iatridou, 1991). The speaker accommodates the presupposition by using *wenn*, for which *falls* is odd[8]. The same contrast holds for factive conditionals as in (6a), with the speaker or contextual presupposition that the antecedent is true, or (6b) from Breindl et al. (2014, p. 756). However, for the latter case, it seems more appropriate to translate the *wenn*-sentence using *since*; this point has been made in Volodina (2006, pp. 367, 368) who claims that a factive use of *wenn* does not allow a purely conditional reading or a temporal reading, but can receive a causal interpretation.

(5) A: *Kai ist krank.*
Kai is sick
'Kai is sick.'
B: *Wenn/?Falls Kai krank ist, muss das Seminar ausfallen.*
if Kai sick is, must the seminar fail
'If Kai is sick, the seminar must be canceled.'

(6) (a) *Wenn/?Falls ich es mir so recht überlege, könnte*
if I it me so right think, could
das stimmen.
it hold
'If I think about it seriously, this could be true.'

(b) (Max says to Moritz, who is present at the moment):
Wenn/?Falls du nun mal wieder hier bist, lass uns
if you now once again here are let us
doch Schach spielen!
DOCH chess play
'Since you are now here again, let's do play the chess!'

The second – controversial – contrast (Contrast 2) is that *falls* is degraded in counterfactual conditionals (indicated by subjunctive mood in German, henceforth "subj" in the examples) or less preferred than *wenn*, see (7). For example, according to the "grammis"[9], counterfactual use of *falls* is usually excluded, with some exceptions, as in their example (8) below. However, it is to note that there might be regional differences in this regard: Some native speakers I consulted with do not judge (7) with *falls* to be degraded.

(7) *Wenn/?Falls Kai krank gewesen wäre, hätte*
if Kai sick been be_subj has_subj
das Seminar ausfallen müssen.
the seminar fail must
'If Kai had been sick, the seminar would have been canceled.'

(8) *Solange die Boeing in Algier stand (. . .),*
as long as the Boeing in Algiers was (. . .),

[7]This does not mean that *wenn* is always ambiguous. For example, if we add a negative polarity item (NPI) such as *jemals* 'ever' into the antecedent, the resulting sentence (*wenn der Regen jemals aufhört* 'if the rain ever stops') only has a conditional reading. Furthermore, in this regard, the English *when* has also been argued to have both a temporal and a conditional reading, see Elder and Jaszczolt (2016) and their example (20) in it: *When you follow that through you've got the means to give rise to a change in the method.* Whether the ambiguity of the English *when* and the German *wenn* is comparable is an empirical question we will leave for future research.

[8]It is to note that the speaker can use *falls* in order to not accommodate the antecedent proposition, for example, to indicate their doubt on A's assertion. In comparison, (5) is supposed to show that in the case that the speaker does accommodate the antecedent proposition, they need to use *wenn* but not *falls*.

[9]https://grammis.ids-mannheim.de/systematische-grammatik/2101 (accessed on 28.10.2020).

wäre ein Kommandounternehmen denkbar
be_subj a commando conceivable
gewesen, falls die algerische Regierung die
been, if the Algerian government the
Zustimmung gegeben hätte. (Zeit, 21.6.1985, 2)
approval given has_subj
'As long as the Boeing was in Algiers (...), a commando would have been conceivable if the Algerian government had given its approval.'

The third, uncontroversial contrast (Contrast 3) is that *falls* is out in counterfactual optatives, see (9). Following Grosz (2012), I assume that counterfactual optatives have no descriptive but presuppositional and expressive content, as illustrated below.

(9) *Wenn/*Falls ich nur reich wäre!*
if I only rich be_subj
'If only I were rich!'
(a) Presupposition: the speaker is not rich.
(b) Expressive: the speaker desires that s/he be rich.

The fourth contrast (Contrast 4) is that unlike *wenn, falls* is degraded with the quantifying adverb *immer* 'always' (Zaefferer, 1991). For the minimal pair in (10a), Zaefferer (1991, p. 216) argues "that explicitly quantified c-constructures are plural forms, bare c-constructures with particles like *if* are transnumeral forms (unspecified with respect to number), and that bare c-constructures with *falls* are singular forms." Some speakers pointed out to me that the only, or the more prominent, reading of (10a) is temporal; this is in line with the claim made in Breindl et al. (2014, pp. 765, 766) that unlike *wenn, falls* cannot quantify over time points but is used only to hypothesize based on the truth or falsity of the antecedent proposition. However, as shown in (10b), the prominent reading of the sentence is clearly conditional, that is, the adverb quantifies over cases rather than times. In Section "Non-at-Issue Meanings of *wenn/falls* in German," I will argue instead that the contrast is due to the presupposition of *always,* which clashes with the meaning of *falls.*

(10) (a) *Immer wenn/*falls Steffi gewinnt, wird gefeiert.*
always if Steffi wins, will celebrated
'Always, if Steffi wins, there will be celebrations.'
(b) *Immer, wenn Sie dieses Symbol sehen,*
always if you this symbol see
können Sie den
can you the
Beitrag als PDF downloaden[10].
contribution as PDF download
'Whenever you see this symbol, you can download the contribution.'

Last but not least, while both *wenn/falls* license NPIs, such as *auch nur irgendein* 'even any' in (11), Liu (2012) claims that *falls* is degraded with factive evaluative adverbs, which show PPI (positive polarity item) behavior (Contrast 5), see her example in (12). The speakers I checked with have different intuitions about *wenn* in (12): It is fine for some, and for others, it is equally odd as *falls*.

(11) *Wenn/Falls Du auch nur irgendeinen*
if you also only any
Artzt kennst, schreibe mir bitte.
doctor know, write me please
'If you know any doctor, please write to me.'

(12) *Wenn/?Falls die Vorschule leider keinen*
if the preschool unfortunately no
Spielplatz hat, können die Kinder
playground has, can the kids
keinen Sport treiben.
no sports do
'If the preschool unfortunately has no playground, the kids cannot do sports.'

Below, I will provide an analysis to account for the *wenn/falls* contrast based on the above observations.

Analysis

In the analysis, I will use two theoretical components: One is the speaker commitment scales used in Giannakidou and Mari (2016, 2021). Following their works, I assume "non-veridical equilibrium" (implying that p and ¬p as equal possibilities) to be the default for epistemic possibility, questions, and conditionals. That is, the speaker does not convey any preference for p or ¬p. But the equilibrium of conditionals (as for questions) can be manipulated to produce bias (i.e., reduced or higher speaker commitment) through various lexical or grammatical devices (for German, see Reis and Wöllstein, 2010; Liu, 2019; Sode and Sugawara, 2019; Liu et al., 2021). In the following, I will provide several examples as triggers of speaker bias and then argue that the *wenn/falls* contrast can be captured along the lines. The other component is the notion of non-at-issue meanings (e.g., Simons et al., 2010; Tonhauser, 2012). I will argue using diagnostic tests from the theoretical literature that the speaker bias conveyed by *falls* in comparison to *wenn* is a non-at-issue meaning. While non-at-issue meanings can be semantic or conventional such as conventional presuppositions or conventional implicatures (Potts, 2005, a.o), I will show further that the non-at-issue of *falls* is of conversational nature as well as that the implicature is different from scalar implicatures.

Giannakidou (1998, 2014), in her (non)veridicality framework, has related attitudes (i.e., speaker's doxastic assumptions) to the notion of speaker commitment. In more recent works, Giannakidou and Mari (2016) argue that differences of attitudes can be modeled through speaker commitment scales (SCSs). For example, they apply the scale in (13) to capture the speaker's doxastic attitude toward the modified proposition. The necessity modal verb *must* conveys a stronger speaker commitment than the possibility modal adverb *possibly*, but a weakened speaker commitment in comparison to the unmodalized variant, which expresses full commitment.

[10] https://www.korrekturen.de/forum.pl/md/read/id/97832/sbj/komma-beiimmer-wenn/(accessed on 13.11.2020)

(13) More committed <unmodalized p, MUST p, POSSIBLY p> Less committed
(a) *It is raining.*
(b) *It must be raining.*
(c) *It is possibly raining.*

It has to be further explored whether the SCSs can encode not only doxastic attitudes but also deontic or bouletic ones, and whether the perspectival agent must be the speaker or can be a sentence subject or another discourse referent. Furthermore, Giannakidou and Mari (2016) remain non-committed as to the semantic or pragmatic nature of this meaning difference, namely, whether it is at-issue or non-at-issue, and in the latter case, whether it is a conventional implicature or a conversational implicature. For example, while the weaker speaker commitment meaning of the possibility modal verb seems to be its semantics, it is unclear how the weakened speaker commitment meaning of *must* relates to its necessity modal meaning. It seems that alternatives and their commitment strength of SCSs are determined by a variety of factors ranging from the at-issue as well as non-at-issue content. Thus, these scales might be different from Horn scales based on entailment relations. With these open questions kept in mind, I will show that SCSs are very useful for modeling grammar of speaker commitment in general, and provide experimental work testing these in Section "Experiments."

First, for example, Zimmermann (2004) argues that the German discourse particle *wohl* (roughly 'probably') expresses a higher speaker credence in the truth of the proposition than the possibility adverb *vielleicht* 'possibly.' Thus, we can fit these alternations into a SCS in (14).

(14) More committed <unmodalized p, WOHL p, VIELLEICHT p> Less committed
(a) *Hein ist auf dem See.*
'Hein is over the lake.'
(b) *Hein ist wohl auf dem See.*
'Hein is probably over the lake.'
(c) *Hein ist vielleicht auf dem See.*
'Hein is possibly over the lake.'

Second, Liu (2012) argues to distinguish between factive and non-factive evaluative adverbs in German, which show different distributions in entailment-canceling contexts (Simons et al., 2010). For example, *leider* and *unglücklicherweise* both mean roughly 'unfortunately,' but the latter can occur in questions, conditionals and modals, whereas the former is odd in these contexts. Liu (2012) thus labels *leider* as a factive adverb and *unglücklicherweise* a non-factive one. This idea can be equally translated into a SCS as in (15).

(15) More committed <LEIDER p, UNGLÜCKLICHERWEISE p> Less committed

Third, Liu (2019) shows experimental evidence that in German and English conditionals, NPIs (*jemals/überhaupt* 'ever/at all') led to lower ratings of speaker commitment to the antecedent proposition in comparison to sentences without NPIs. This finding can be put into the SCS in (16). Whether this scale holds for all NPIs or not is a question beyond the scope of this paper.

(16) More committed <~~NPI~~ p, NPI p> Less committed

Fourth, SCSs can also be used to model the difference between the clausal connectives. For example, in contrast to non-veridical CCs, causal connectives are veridical or factive operators, that is, they convey the speaker's full commitment to the truth of the antecedent (Giannakidou, 1998, et seq). This idea can also be put into a SCS, as shown in (17).

(17) More committed <BECAUSE p, IF p> Less committed

In the rest of this section, I will argue that SCSs are also useful for modeling the internal differences among the CCs.

That CCs can differ in degrees of speaker commitment is not new. For example, Visconti (1996, p. 555) claims that CCs can contribute secondary (in recent terms, 'non-at-issue') meanings concerning a 'propositional attitude' toward the modified propositions, such as the speaker's epistemic/doxastic/deontic/emotional evaluation toward the antecedent or the consequent. In Italian, Visconti claims that the CCs *nel caso che* 'in the case that,' *nell'eventualità che* 'in the eventuality that' and *casomai* 'if-ever' [made up of a simple CC *caso* 'in case, if' and a NPI *mai* 'ever'] differ in terms of the speaker's attitude toward the antecedent 'p' that is expressed at the level of conventional implicatures: While *nel caso che* is doxastically neutral, *nell'eventualità che* expresses a negative bias 'unlikely(p)' and *casomai* conveys an even stronger bias, namely, 'improbable(p).' Due to the different degrees of the bias, it is odd to use *nell'eventualità che* (or *casomai*) for modifying the antecedent that is simultaneously labeled as highly likely by the non-restrictive relative clause, whereas it is not a problem for *nel caso che*, as shown in her example (18) (Visconti, 1996, p. 559). The idea can be translated into a SCS in (19).

(18) *Nel caso che/?Nell'eventualità che Giampiero riesca ad*
in case that/?in the event that Giampiero manages to
affittare quella casa al mare
rent that house by the sea
– cosa che pare molto probabile – passeremo da lui una
– what that seem very likely– we will pass by him a
settimana in luglio.
week in July
'In the case/?In the eventuality that Giampiero manages to rent that house by the sea –
which he almost certainly will – we'll go and stay with him for a week in July.'

(19) More committed <NEL CASO CHE p, NELL'EVENTUALITÀ CHE p, CASOMAI p> Less committed

Visconti's proposal of treating the speaker assumption conveyed by these CCs as conventional implicature is a very insightful idea and is obviously useful for analyzing complex (i.e., multi-word) but compositional CCs (i.e., with transparent

semantics which can be derived compositionally with the subparts of the CC) such as those in (20). The adjectives provide information about the speaker's doxastic (20a) or bouletic assumptions (20b) about the antecedent proposition. These meanings are logically and compositionally independent of the conditional core in these sentences, and thus are indeed conventional implicatures in the sense of Potts (2005). On the other hand, conventional implicatures are neither cancellable nor reinforceable, compared to conversational implicatures, which are cancellable and reinforceable. This raises the empirical question whether all the CCs express speaker bias at the dimension of conventional implicatures, or whether they can encode weaker, i.e., non-conventional, meaning.

(20) (a) *In the impossible/possible/unlikely/likely event that. . .*
(b) *In the fortunate/unfortunate event that. . .*

Liu and Wang (2021) provide distributional and experimental evidence that the Mandarin Chinese CC *wanyi* (lit, 'one of ten thousand,' originally a numerical expression, used as a CC in modern Mandarin) conveys a weakened speaker commitment than *ruguo* 'if', as in (21)/(22). They treat this meaning difference at the dimension of non-at-issue-meanings (Simons et al., 2010).

(21) More committed <RUGUO p, WANYI p> Less committed

(22) *Ruguo/Wanyi wo xiawu you shijian, wo jiu he*
if I afternoon have time, I JIU with
pengyou qu he kafei.
friend go drink coffee
'If I have time in the afternoon, I will go have a coffee with my friends.'

The same scale can apply to the German CCs, such as in (23a): *wenn/falls* express speaker commitment of intermediate degree between *im wahrscheinlichen Fall, dass* 'in the probable event that' and *im unwahrscheinlichen Fall, dass* 'in the improbable event that.' Further, I argue that compared to *wenn, falls* expresses a weakened speaker commitment toward the antecedent proposition (*p*), see (23b). That is, *falls* indicates that the speaker takes *p* as not likely. This meaning acts at a separate layer of doxastic states, i.e., it does not target the question under discussion, and thus it is non-at-issue (Simons et al., 2010).

(23) (a) More committed < . . . *'im wahrscheinlichen Fall, dass', wenn/falls, 'im unwahrscheinlichen Fall, dass',* . . .> Less committed
(b) More committed <WENN p, FALLS p> Less committed

Following this, *falls* has an attitudinal meaning at a separate layer of doxastic states, i.e., λp.¬likely(*p,x*) with x as a free variable (for the attitudinal holder) whose value is to be determined by context (e.g., *x* is the speaker, or the sentential subject). This is what I call 'weak unlikelihood implicature' (WUI). A sentence such as (24) expresses an at-issue content as proposed by Kratzer (1986, 1991) and paraphrased in (24a), and additionally, a non-at-issue content as in (24b).

(24) *Falls es regnet, werden die Strassen nass.*
if it rains, become the streets wet
'If it rains, the streets will get wet.'
(a) At-issue content: the worlds (compatible with the speaker's knowledge) in which it rains are among the worlds in which the streets get wet.
(b) Non-at-issue content: ¬likely(rain,x)

I will first address the non-at-issue and the implicature part of the proposal and then explain why it should be "weak."

Tonhauser (2012) puts forward three criteria along which at-issue content differs from non-at-issue content: First, at-issue content can be directly assented or dissented with, but non-at-issue content cannot. Second, at-issue content addresses the question under discussion, but non-at-issue content does not. Third, at-issue content determines the relevant set of alternatives whereas non-at-issue content does not. I will apply one test Tonhauser proposes based on the first criterion in (25): As is shown, the conditional (i.e., at-issue) meaning in A's utterance can be assented or dissented with positive continuation (B1 and B2) but the speaker assumption about the antecedent proposition cannot (B3 and B4). This contrast speaks in favor of the non-at-issue status or pragmatic nature of the bias encoded in *falls*.

(25) Diagnostic: Assent/dissent with positive continuation
A: *Wenn/Falls es morgen regnet, kommt Paul*
if it tomorrow rains, comes Paul
mit dem Auto.
with the car
B1: *Ja, richtig. Bei Regnen kommt er mit dem Auto.*
yes, right. with rain comes he with the car
B2: *Nein, stimmt nicht. (Selbst) bei Regnen kommt*
no, holds not even with rain comes
er nicht mit dem Auto.
he not with the car
B3: *#Ja, richtig. Es regnet morgen wohl nicht.*
yes, right. it rains tomorrow probably not
B4: *#Nein, stimmt nicht. Es regnet morgen wohl.*
no, holds not. it rains tomorrow probably
['A: If it rains tomorrow, Paul will come by car. B1: Yes, right. If it rains, he will come by car. B2: No, not true. (Even) when it rains, he does not come by car. B3: #Yes, right. It probably won't rain tomorrow. B4: #No, not true. It probably will rain tomorrow.']

Furthermore, non-at-issue contents can project out of entailment canceling contexts (Simons et al., 2010; Liu, 2012). For example, in (26), the intuition is that the bias conveyed by *falls* survives embedding in the question operator, an entailment canceling context.

(26) *Kommst du noch mit, falls es regnet?*
come you still with if it rains
'Are you still coming, if it rains?'

In addition, consider the minimal pair in (27) from Liu (2019): The relative clause in the sentence indicates the speaker's commitment to the antecedent proposition, which does not go along with the CC *im unwahrscheinlichen Fall, dass* but

is ok with *falls*. This indicates that in the former case, the unlikelihood meaning component is semantic/conventional and thus uncancellable, whereas it is pragmatic/conversational and thus cancellable in the latter case.

(27) (a) *#Im unwahrscheinlichen Fall, dass es draußen regnet*,
in the improbable case, that it outside rains
was ich glaube, bleibt Susanne zu Hause.
which I believe stays Susanne at home
'In the improbable case that it is raining outside, which I believe, Susanne will stay at home.'

(b) *Falls es draußen regnet, was ich glaube, bleibt*
if it outside rains which I believe stays
Susanne zu Hause.
Susanne at home
'If it is raining outside, which I believe, Susanne will stay at home.'

These data taken together indicate that the weak bias created by *falls* is a secondary, i.e., non-at-issue, content that is lexically triggered but contextually cancellable, and more precisely, an implicature (Grice, 1975; Potts, 2005; see Zakkou, 2018 for discussion of the reliability of the cancellability test). Here, a natural question arises whether it is a scalar implicature. However, if we compare the CC scale in (23b) with a typical Horn scale such as <all, some>, there are at least the following three aspects where they differ: First, the scalar implicature, e.g., *not all students came* is computed based on the semantics of the sentence *some students came*, whereas it is not the case for *wenn/falls p, q* as the negative bias does not target the conditional dependence between the antecedent and the consequent. Second, related to the first aspect, the bias by *falls* conveys speaker's assumptions that the hearer can ignore, as it does not target the QUD whereas scalar implicatures can target QUDs, e.g., *How many students came?* in the above example. The third and most straightforward argument against a scalar implicature analysis for the *wenn/falls* contrast is that the Horn scale is based on a proper entailment relation, e.g., *all students came* entails *some students came*, whereas this does not hold for *wenn p, q* and *falls p, q*: semantically, they both convey the same conditional relation between *p* and *q* in that all *p*-cases are *q*-cases. Thus, I take the implicature by *falls* to be different from scalar implicatures.

The naturally occurring examples in (28) show the speaker's awareness of the meaning difference between *wenn* and *falls*. Whereas *wenn* in (28a) can have either a conditional or a temporal reading, (29) is unambiguously meant as a conditional[11].

(28) (a) *Wenn – und nicht falls – wir den nächsten großen*
if – and not if – we the next big
Titel gewinnen – egal ob
title win – regardless whether
Europa- oder Weltmeisterschaft -, bauen wir hier
European or World Cup – build we here
ein großes Zelt auf und lassen es
a big tent up and let it
so richtig krachen![12]
so right crash
'When – and not if – we win the next big title – regardless of whether it is a European or a World Cup – we will put up a big tent here and let it rip!

(b) *. . . falls (ich schreibe bewußt falls und nicht wenn*,
. . . if (I write deliberately if and not if
da das nicht eindeutig aus Deinem
because that not clearly from your
ET[Entrag] hervorgeht) Du nur eine Frau
entry shows you only a woman
akzeptabel findest, die deutlich über deinem eigenen
acceptable find, who well above your own
Level liegt, bist DU das Problem, aber das
level lie, are you the problem, but this
weißt nur Du selbst.[13]
know only you yourself
'. . . if/falls (I deliberately write *falls* and not *wenn*, because that is not clear from your entry) you only find a woman acceptable who is well above your own level, you are the problem, but only you know that yourself.'

I argue that *falls* encodes a weak unlikelihood meaning (i.e., that the speaker does not take the antecedent proposition as likely) instead of a strong unlikelihood meaning (i.e., that the speaker takes a proposition as unlikely). The latter meaning is expressed by, for example, additive particles such as the English *even*. As shown in (29), a strong unlikelihood meaning will be too strong for *falls*.

(29) *Ich rate dir dringend, tu dir selbst einen Gefallen und*
I advise you strongly do you yourself a favor and
kombiniere Deutsch nicht mit Englisch.
combine German not with English.
Falls du es doch tust, wirst du es garantiert bereuen.[14]
if you it still do, will you it guaranteed regret
'I strongly advise you to do yourself a favor and not combine German with English. If you still do so, you're guaranteed to regret it.'

By this analysis, the degradation of *falls* in premise/factive conditionals (Contrast 1) results from the clash between the speaker presupposition and the WUI of *falls*. *Falls* is degraded in counterfactual optatives and arguably in counterfactual conditionals (Contrast 3 and 2) due to the counterfactual presupposition or implicature (i.e., speaker's anti-commitment to the antecedent proposition). This contradicts the meaning of *falls*, which presupposes the absence of bias to start with,

[11] *Wenn* and *falls* can also be used in coordination, such as *wenn und falls, wenn oder falls* indicating that they have common semantic properties and distinctive features at the same time (Breindl et al., 2014).

[12] https://de.fifa.com/who-we-are/news/the-three-lions-new-den-1816565 (accessed on 31.10.2020).

[13] https://community.elitepartner.de/forum/frage/wie-ist-die-wahrscheinlichkeit-mit-mitte-30-eine-nette-frau-zu-finden-und-eine-familie-zu-gruenden.79095/ (accessed on 13.11.2020).

[14] https://www.studis-online.de/Fragen-Brett/read.php?101,330531 (accessed on 01.11.2020).

i.e., the non-veridicality of the antecedent proposition (dubbed as the "Non-veridicality Equilibrium," the default of epistemic possibility, in Giannakidou and Mari, 2021), and implicates the WUI. The reason why *falls* is acceptable in counterfactual conditionals for some speakers might be because the negative bias of *falls* is a conversational implicature (which is lexically triggered but needs contextual support) and thus cancellable or optional (Grice, 1989; Zakkou, 2018), or in general, there might be individual differences in the (quality of the) lexical representations of *wenn/falls.* The latter goes far beyond the scope of this paper and thus will not be addressed here. Another possibility is that subjunctive conditionals do not always presuppose or implicate the falsity of the antecedent proposition. But this is not a plausible explanation, as even though the counterfactuality inference does not always hold, it cannot be canceled without good reason (see Footnote 3 and the cited references therein).

For Contrast 4, I propose an alternative explanation to Zaefferer's (1991) account, namely, the wide-scope *immer* is presuppositional. It presupposes that the event in the antecedent takes place more than once, which clashes with the WUI by *falls.* For illustration, *immer* in (10) presupposes that Steffi wins more than once, but *falls* would convey that it is not likely that Steffi wins, thus their combination is odd. The degradation of *falls* in co-occurrence with factive adverbs (arguably compared to *wenn*, see Contrast 5) can also be attributed to the factivity presupposition of the adverb, in clash with its WUI. All in all, this shows that the proposed difference in terms of speaker commitment for *wenn*/falls is able to account for the listed distributional differences: *falls,* as the more restricted CC in comparison to *wenn*, has the proposed lexical pragmatics, which is cancellable and reinforceable through grammatical devices, as we will see in Section "Experiments." I leave it open for now whether some of the differences can be captured differently, but will discuss several alternative accounts, which I argue are in line with the proposed one.

The present analysis for *wenn/falls* echoes the observation occasionally made in the previous literature, for example, by Breindl et al. (2014, pp. 114, 115): "The difference between *wenn* and *falls* has to do with the probabilities of the occurrence of the antecedent" (translated from German). They use the example in (30) to argue that the speaker, being aware of their differences, uses one CC or the other to indicate implicitly their assessment of the probability of the antecedent proposition (i.e., speaker commitment).

(30) *Rushdie: um es grob zu sagen: falls Gott existiert,*
Rushdie: for it roughly to say if God exists,
wird er sich um die „Satanischen Verse"
will he himself about the Satanic Verses
nicht scheren; wenn er nicht existiert,
not care; if he not exists,
auch nicht. (*Der Spiegel*, 11.05.1992, S. 214)
also not
'Rushdie: To put it roughly, if God exists, he won't care about the "Satanic Verses"; if he does not exist, he won't, either.'

So far, we have seen distributional properties of the *wenn/falls* contrast and I proposed that the two CCs differ in degrees of conveyed speaker commitment. In the following, I will report on experimental work that tested the proposed contrast.

EXPERIMENTS

I conducted three experiments to test the proposed analysis above, as well as a fourth experiment on the interaction between two different SCSs. In addition, I will also report on the related German experiment from Liu (2019).

Experiment 1

The original assumption for Experiment 1 was that if *falls* carries a negative bias in comparison to *wenn*, as illustrated in (31), then it might tend to occur more naturally with events (or, propositions) that are less likely, i.e., more naturally with p3 than with p1 in (32). This assumption turned out to be problematic. I report on this experiment here nevertheless, as it is important to showcase potential pitfalls and necessary "precautions" to take for testing lexical pragmatics in general.

(31) More committed <WENN p, FALLS p> Less committed
(32) More likely <p1, p2, p3> Less likely

Materials and Methods

Experiment 1 was a rating study based on a 2 × 3 within-subjects design, with one factor being CC (*wenn/falls*) and the other factor being the likelihood of the event in the conditional antecedent *p* (in three levels, i.e., very likely/likely/unlikely), see **Table 1**. The starting prediction was that due to the weakened speaker commitment by *falls* in contrast to *wenn,* it might be degraded in combination with very likely and likely events in comparison to the other conditions.

Twenty-four items were used, with one example in (33), as well as 84 additional filler items. The critical stimuli are provided in the **Supplementary Materials** (Test sentences of Experiment 1)[15].

(33) (a) *Wenn die Läden am Montag geöffnet sind, gehen*
if the shops on the Monday opened are, go
Tom und Lisa einkaufen.
Tom and Lisa shop

(b) *##Falls die Läden am Montag geöffnet sind, gehen*
if the shops on the Monday opened are, go

[15] As the fillers involved test materials of other experiments, they are not provided in the **Supplementary Materials**.

TABLE 1 | Factors, conditions and predictions of Experiment 1.

Events	*wenn*	*falls*
Very likely	1: (a)	2: ## (b)
Likely	3: (c)	4: #(d)
Unlikely	5: (e)	6: (f)

Tom und Lisa einkaufen.
Tom and Lisa shop

(c) *Wenn* *die Läden am* *Samstag geöffnet sind, gehen*
if the shops on the Saturday opened are, go
Tom und Lisa einkaufen.
Tom and Lisa shop

(d) *#Falls* *die Läden am* *Samstag geöffnet sind, gehen*
if the shops on the Saturday opened are, go
Tom und Lisa einkaufen.
Tom and Lisa shop

(e) *Wenn* *die Läden am* *Sonntag geöffnet sind,* gehen
if the shops on the Sunday opened are, go
Tom und Lisa einkaufen.
Tom and Lisa shop

(f) *Falls* *die Läden am* *Sonntag geöffnet sind,* gehen
if the shops on the Sunday opened are, go
Tom und Lisa einkaufen.
Tom and Lisa shop

('If the shops are open on Monday/Saturday/Sunday, Tom and Lisa will go shopping.')

As the German CC *wenn* has both a temporal (i.e., *when*) and a conditional (i.e., *if*) reading in present indicative, a temporal adverb was used in the antecedent so that the conditional reading of *wenn* becomes more plausible than the temporal reading. The focus was on whether the naturalness of *falls* decreases by the likelihood increase of *p*. The events were chosen and ordered based on common knowledge in the German context. They were tested informally with 10 subjects (undergraduate students of Osnabrück University). The subjects saw a list of three-event pairs (e.g., the shops are open on Monday/Saturday/Sunday) in different orders and were asked to order the events in each pair by their likelihood. The results show that the scales were all valid[16].

The experiment was conducted online using the SoSci Survey[17]. 36 undergraduate students (25 female, 11 male; 35 between 18 and 29 years old, 1 under 18)[18] of Osnabrück University took part in the study online for course credits. The participants each saw 108 sentences in total, which were presented one by one in the middle of the computer screen, and they rated the naturalness of each sentence (0: unnatural, 1: natural). Our predictions were that (33b/d) would receive lower ratings than (33a/c/e/f), as the negative bias by *falls* might clash with the likely events of the shops being open on Monday or Saturday in the German context.

[16]Please note that the studies reported in the paper were all conducted long before the Covid-19 Pandemic, as today (May 04, 2021 in Germany), it is not unusual that the shops are closed on Monday or Saturday.

[17]https://www.soscisurvey.de/

[18]All the studies reported were conducted between 2016 and 2017, and we only collected binary gender information. We collected age information using ranges for Experiments 1, 3, and 4b. While we also collected dialectal information (i.e., whether the subjects speak any German dialect and if yes, which ones) in the experiments, there was way too little data to report here. We did not control whether subjects were bilingual or not.

Results

All analyses were performed using mixed effects linear regression models. The models were constructed using the *lme4* package (Baayen et al., 2008; Bates et al., 2012) in R (R Core Team, 2018). The reported model is the maximal model that converged (Barr et al., 2013). The model included CC and Event-likelihood (with interaction term) as fixed effects. Furthermore, it included random intercepts for subjects, items and stimuli order, as well as random by-subject and by-item slopes for the effects of CC and Event-likelihood.

The results (see **Table 2** and **Figure 1**) showed neither the interaction nor the main effect for CC. That is, the comparison between *wenn* and *falls* was not significant ($t = 0.15$, $p = 0.88$). For the Event-likelihood, there is a numerical difference in the naturalness rating for likely events when compared to either unlikely or very likely events. However, these contrasts, too, fail to reach significance [Tukey's test for multiple comparisons of means: $t = 1.52$, $p = 0.26$ (very likely vs. likely); $t = 1.63$, $p = 0.215$ (unlikely vs. likely)].

Discussion

Experiment 1 did not confirm the prediction and thus does not provide evidence for the proposed account of the *wenn/falls* contrast. This can mean that the two connectives are interchangeable in the conditional use. Alternatively, the methods used here might be problematic. First, some of the fillers involve unlicensed polarity items, e.g., *jemals* 'ever' in positive sentences, which are ungrammatical (see Ladusaw, 1980; Giannakidou, 2011; Liu et al., 2019). This might have contributed to the ceiling effect for the *wenn/falls* sentences. Second, the attitude by *falls* is a speaker-oriented (i.e., subjective) meaning

TABLE 2 | Descriptive statistics of Experiment 1.

Condition	CC	Event	Rating	*SE*
1	*wenn*	Very likely	0.95	0.02
2	*falls*	Very likely	0.97	0.02
3	*wenn*	Likely	0.88	0.03
4	*falls*	Likely	0.88	0.03
5	*wenn*	Unlikely	0.97	0.01
6	*falls*	Unlikely	0.96	0.02

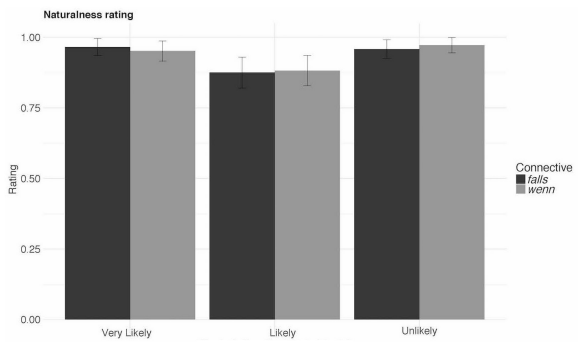

FIGURE 1 | Results of Experiment 1.

at a separate dimension (i.e., non-at-issue), i.e., it can be ignored from the hearer's (i.e., subjects') perspective. In other words, the contrast is context-dependent, i.e., present in some contexts and absent in others (as Experiments 2–4 will show). Thirdly, the naturalness rating studies with the binary scale was maybe not sensitive enough to measure such subtle lexical pragmatics (e.g., due to shallow processing).

Experiment 2

Experiment 2 used the forced lexical choice paradigm. It was based on the assumption that if the *wenn/falls* contrast was real, then, given both as possible lexical items to use, speakers would opt for one or the other in a conditional expression depending on the context, i.e., their degree of commitment or credence in the antecedent. More specifically, in a context where the protagonist is positively biased toward the antecedent proposition, they would use *wenn* more often than *falls;* in a context where the protagonist is negatively biased toward the antecedent proposition, a reverse pattern is to be expected, see **Table 3**.

Materials and Methods

Experiment 2 was based on a one factorial within-subjects design with two levels for the factor CONTEXT, encoded in the sentence preceding the conditional sentence. Bearing in mind that the negative bias of *falls* is subjective meaning, I used a third-person protagonist to keep the doxastic anchoring constant, i.e., to prevent subjects from taking egocentric perspectives. The protagonist either believes *p* or not.

The subjects were asked to choose among *wenn, falls* and a mismatching control item such as *oder* 'or.'

(34) *Susanne hat sich morgen für einen Tag frei genommen.*
Susanne has herself tomorrow for a day free taken
Sie {glaubt/glaubt nicht} dass es morgen regnet und denkt:
She believes/believes not that it tomorrow rains and thinks:
_____*es morgen regnet, bleibe ich zu Hause.*
_____it tomorrow rains, stay I at home
'Susanne has taken tomorrow for a day off. She {believes/doesn't believe} that it will rain tomorrow and is thinking: _____it rains tomorrow, I will stay at home'
(a) wenn (b) falls (c) oder

Twenty-four items such as (34) were used, as well as 48 fillers. The critical stimuli are provided in the **Supplementary Materials** (Test sentences of Experiment 2). The experiment was programmed in Python and conducted in the behavorial lab of the Institute of Cognitive Science of Osnabrück University. 52 undergraduates (29 female, 23 male; mean age = 21.2, *SD* = 1.7) of Osnabrück University participated in the study for course credits.

Results

Wrong answers, i.e., answers with the mismatching lexical items [e.g., *oder* 'or' in (34)], were excluded from the data analysis. The response (see **Table 4**) in this experiment is binary, not numeric (i.e., a binary choice between *falls* and *wenn*), therefore all analyses were performed with mixed logistic regression models. The model included the answer choice (*falls/wenn*) as dependent variable and the CONTEXT (belief/disbelief) as predictor variable. The random effects structure included random intercepts for subjects, items and stimuli order, as well as random by-subject slopes for the fixed effect. The model reported is the maximal model that converged. The model has a good fit (precision = 0.79, recall = 0.78) and performs significantly better than a baseline of guessing a response ($p < 0.0001$). It yields a significant effect of CONTEXT ($t = 6.19$, $p < 0.0001$), that is, subjects are significantly more likely to choose *falls* under the disbelief-condition and *wenn* in the belief-condition.

Discussion

Experiment 2 shows that the doxastic agent chose *falls* over *wenn* if they have a low degree of credence toward the modified antecedent proposition. As the materials use both CCs in the sentence-initial position, I take the results to reflect meaning differences instead of syntactic preferences (in a sentence-final position, as Breindl et al., 2014 point out). That is, they provide positive evidence for the *wenn/falls* contrast in terms of degree of speaker commitment.

Experiment 3

Experiment 1 did not confirm the *wenn/falls* contrast. As mentioned above, this can be due to the context-dependence of the contrast due to, among others, the subjective (speaker-oriented) nature of the bias encoded in *falls*. For this reason, Experiment 3 included an additional relative clause commenting on the conditional antecedent. The rationale was that in this way, the attitudes encoded in the CC and the relative clause had the same anchoring toward the speaker. While this was also controlled in Experiment 2, Experiment 3 was conducted to find out whether the bias is cancellable (by a RC conveying speaker's disbelief in the antecedent proposition) or reinforceable (by a RC conveying speaker's belief in the antecedent proposition). The results can shed further light on the nature of the bias. If the meaning difference is cancellable and reinforceable, we can conclude that it is a pragmatic difference.

Materials and Methods

Experiment 3 was a rating study based on a 2 × 2 factorial within-subjects design, with the factor CC (*wenn/falls*) and RC (relative clause) expressing a likelihood/unlikelihood propositional attitude toward *p*. The method and procedure were

TABLE 3 | Factor, conditions, and predictions of Experiment 2.

Context	*wenn* vs. *falls*
Belief-context	*wenn* preferred to *falls*
Disbelief-context	*falls* preferred to *wenn*

TABLE 4 | Results of Experiment 2.

Context	*wenn*	*falls*
Belief-context	442	185
Disbelief-context	177	436

similar as in Experiment 1. 24 items were used as well as 72 fillers. Half of the items used RCs as in (35) and the other half used RCs as in (36). The critical stimuli are provided in the **Supplementary Materials** (Test sentences of Experiment 3). We did not control the likelihood of the antecedent and used the unlikely conditions due to lack of contrast for this manipulation in Experiment 1. 40 undergraduates (25 female, 15 male; 39 between 18 and 29 years old, 1 between 30 and 39) of Osnabrück University participated in the experiment for course credits.

(35) *Wenn/Falls die Läden am Sonntag*
if the shops on the Sunday
geöffnet sind, was ich für
opened are, which I for
wahrscheinlich/unwahrscheinlich halte,
probable/improbable take,
gehen Tom und Lisa einkaufen.
go Tom and Lisa shop
'Wenn/Falls the shops are open on Sunday,
which I hold as likely/unlikely,
Tom and Lisa will go shopping.'

(36) *Wenn/Falls die Läden am Sonntag geöffnet sind*,
if the shops on the Sunday opened are,
was ich glaube/nicht glaube,
which I believe/not believe
gehen Tom und Lisa einkaufen.
go Tom and Lisa shop
'Wenn/Falls the shops are open on Sunday,
which I believe/do not believe,
Tom and Lisa will go shopping.'

TABLE 5 | Descriptive statistics of Experiment 3.

Condition	CC	RC	Rating	*SE*
1	*wenn*	Belief	0.78	0.05
2	*wenn*	Disbelief	0.73	0.06
3	*falls*	Belief	0.73	0.06
4	*falls*	Disbelief	0.79	0.05
(A) Descriptive statistics of Experiment 3 with the RC *was ich glaube/nicht glaube* 'which I believe/do not believe'.				
1	*wenn*	Belief	0.66	0.04
2	*wenn*	Disbelief	0.66	0.04
3	*falls*	Belief	0.77	0.04
4	*falls*	Disbelief	0.82	0.04
(B) Descriptive statistics of Experiment 3 with the RC *was ich für wahrscheinlich /unwahrscheinlich halte* 'which I deem likely/unlikely'.				
1	*wenn*	Belief	0.91	0.03
2	*wenn*	Disbelief	0.81	0.04
3	*falls*	Belief	0.70	0.04
4	*falls*	Disbelief	0.76	0.04

Results

All analyses of the data (see **Table 5** for descriptive statistics) were performed using mixed effects linear regression models. The models were constructed using the *lme4* package in R (Baayen et al., 2008; Bates et al., 2012; R Core Team, 2018). All contrasts of interest, i.e., CC and RC, were sum coded and included as fixed effects in the models. The reported models are the maximal models that converged.

The first model included CC and RC (with interaction term) as fixed effects. Furthermore, it included random by-subject and by-item intercepts, as well as random by-subject and by-item slopes for the effects of CC and RC (and their interaction). Neither of the two main effects was significant. There was a significant interaction between CC and RC (t = 2.15, p = 0.03): in the belief-condition, *wenn* was rated more natural than *falls*. The reverse pattern was found in the disbelief-condition. Pairwise comparisons between all four conditions, however, showed no significant differences between either of them, indicating that the interaction effect is highly nuanced.

Furthermore, the RCs in the experiment included 12 items with *was ich für wahrscheinlich/unwahrscheinlich halte* ('which I deem likely/unlikely'), and another 12 with *was ich glaube/nicht glaube* ('which I believe/do not believe'). Thus, I did an additional test to see whether and to what extent there were differences between the two types of predicates. First, a model was constructed which included the type of the predicates (henceforth, PREDICATE) as a third fixed effect and included a three-way interaction term between PREDICATE, CC, and RC. The random effects structure was identical to that of the first model. The new model indicates a significant effect for PREDICATE (t = 2.60, p = 0.009), but no other main effects or higher-order interactions between PREDICATE and CC or RC. However, the reader should keep in mind that the factor PREDICATE was not a systematically controlled condition within items. Thus, we cannot exclude the possibility that any effect was due to confounding influences that rendered some items more natural than others. Nevertheless, it may serve as the first indication of an effect, which can be further explored in future studies.

To further explore the effect of the RC, separate models were created for the two sets of items, see **Figure 2** and **Tables 5A,B**. Both models thus used half of the data set and included CC and RC (with interaction term) as fixed effects. Again, the random effects structure included random by-subject and by-item intercepts, as well as random by-subject slopes for the effects of CC and RC.

Firstly, for the items that used an RC with the verb *glauben/nicht glauben* 'believe/not believe,' no significant effects were found (see **Table 5A**). Neither the main effects nor the interaction turned out to be significant, that is, there was no systematic effect of either factor. For the items that used the RC containing *wahrscheinlich/unwahrscheinlich* 'likely/unlikely,' there was a significant interaction between CC and RC (t = 2.05, p = 0.04), see **Table 5B**. However, neither of the main effects turned out to be significant and paired-tests also showed no

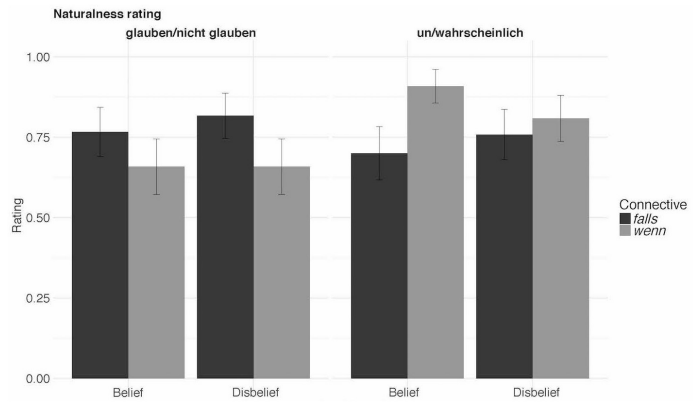

FIGURE 2 | Results of Experiment 3.

significant contrast between any of the comparisons, which indicates a high degree of variation in the data.

Discussion

The naturalness rating results of Experiment 3 are compatible with the lexical choice results of Experiment 2: *Wenn* is preferred over *falls* in the belief-context and vice versa in the disbelief-context. On the other hand, the results in Experiment 3 are also not straightforward to interpret.

As the analysis including PREDICATE shows, the overall interaction effect was mainly driven by the interaction among the items using the RC *was ich für wahrscheinlich/unwahrscheinlich halte* ('which I deem likely/unlikely'). The *wenn*-sentences were rated more natural with these than with *was ich (nicht) glaube* ('which I believe/do not believe'). A possible explanation for the effect of PREDICATE lies in their difference in terms of speaker commitment. As (37) shows, *was ich glaube/nicht glaube* ('which I believe/do not believe') conveys the speaker's full commitment or anti-commitment, whereas *was ich für wahrscheinlich/unwahrscheinlich halte* ('which I deem likely/unlikely') conveys the speaker's weakened commitment or weakened anti-commitment.

(37) More commitment <GLAUBEN, WAHRSCHEINLICH, UNWAHRSCHEINLICH, NICHT GLAUBEN>
Less committed

A full account of these differences presupposes a good understanding of the predicates *believe* and *probable* used in the RC, which goes beyond the scope of this paper. For example, it was pointed out to me (Juliane Schwab, p.c.) that the addition of certain adverbs improves the *believe*-sentences, as in (38). The effect of adding *durchaus* 'quite' weakens the speaker commitment, bearing a similar effect as *wahrscheinlich* 'likely'. The addition of *eigentlich* 'actually' signals that there is a contextual expectation (e.g., of the shops being open) set by the antecedent which the speaker rejects with the use of the RC (see Bergena and Boskerb, 2018).

(38) *Wenn die Läden am Sonntag geöffnet sind,*
if the shops on the Sunday opened are,
was ich durchaus glaube/eigentlich nicht
which I quite believe/actually not
glaube, gehen Tom und Lisa einkaufen.
believe go Tom and Lisa shop
'If the shops are open on Sunday,
which I quite believe/actually do not believe,
Tom and Lisa will go shopping.'

As to *falls,* it has been noted that the negative bias generated by it is not always cancellable, as can be seen in (39) from Liu (2019), attributed to an reviewer.

(39) *??Falls es draußen regnet - und ich bin*
if it outside rains - and I am
mir fast sicher, dass es das tut
me almost certain, that it this does
- bleibt Susanne zu Hause.
- stays Susanne at home
'If it is raining outside - and I am almost certain
that it is - Susanne will stay at home.'

However, the native speakers I checked with have no problems with this sentence. Moreover, the data in **Table 5A** show no difference between the CC or the polarity of the RC. This means that the bias of *falls* is reinforceable (by the negative RC) and cancellable (by the positive RC). With the questions about the predicates (*believe, certain, likely*) left for the future, overall, Experiment 3 shows that the speaker bias encoded in *falls* is a conversational, non-at-issue meaning as is proposed in Section "Non-at-Issue Meanings of *wenn/falls* in German." This is also in line with the results of Experiment 1.

Experiment 4

Experiments 1–3 tested the *wenn/falls* contrast directly with different measures. The results were mixed with no evidence in Experiment 1, with evidence in Experiment 2 and with inconclusive results (weak evidence in the predicted direction) in Experiment 3. In this section, I report two additional experiments addressing the question of how the SCS of CCs interacts with

other SCSs. Experiment 4a on the interaction of the CC scale and the NPI scale (16) refers to the German study reported in Liu (2019), which I summarize here. In comparison, Experiment 4b tested the interaction of the CC scale and the EDAV scale, as shown in (15). The NPI and the EADV scales differ from each other in that the former conveys weakened speaker commitment and the latter high speaker commitment, in line with their distributional requirements (NPIs for negative contexts and EADV for positive contexts). The purpose of these two studies is to reveal the CC contrast by checking their interaction with different SCSs.

Experiment 4a: Summary of Liu (2019)

In Liu (2019), the author reports on a "speaker commitment" rating experiment in German addressing the difference between *wenn/falls,* the effect of NPIs *jemals/überhaupt* 'ever/at all,' and their interaction. Subjects were given scenarios, e.g., (40), consisting of 4 sentences (S1–S4) presented one by one: S1 sets the context; S2 contains a conditional sentence in one of the four combinations, with half of them containing *jemals* and the other half *überhaupt* (e.g., *wenn-überhaupt, wenn+überhaupt, falls-überhaupt, falls+überhaupt*); S3 asks the subjects to rate the degree of the protagonist's commitment to the antecedent on a 5-point Likert scale (1 = certainly not, 5 = certainly yes). S4 is a comprehension question. The results show a significant effect of CC (in that the *falls* conditions received lower ratings than the *wenn* conditionals), a significant effect of NPI (with high ratings in the conditions without NPIs than with NPIs) and a significant interaction. Both scales in (41) and (42) are confirmed, with a significant CC contrast in the absence of NPIs, which disappears in the presence of NPIs.

(40) S1: *Melanie sucht nach einem Sommerkleid.*
'Melanie is looking for a summer dress.'
S2: *Sie denkt:* "*Wenn/Falls ich (*überhaupt*) ein schönes finde, kaufe ich es sofort.*"
'She thinks,' "If I find a nice one (at all), I will buy it immediately."
S3: *Glaubt Melanie, dass sie ein schönes Kleid findet?*
'Does Melanie believe that she will find a nice dress?'
S4: *Möchte Melanie warme Stiefel kaufen?*
'Does Melanie want to buy warm boots?'

(41) NPI SCS: $_{\text{More committed}}$ <~~NPI~~ p, NPI p> $_{\text{Less committed}}$

(42) CC SCS: $_{\text{More committed}}$ <WENN p, FALLS p> $_{\text{Less committed}}$

If Experiment 2 provides indirect evidence for the CC scale, the finding of this study complements it with direct evidence in favor of the *wenn/falls* contrast. But while Liu (2019) focuses on CCs and NPIs from a cross-linguistic perspective by comparing German vs. English, the current paper provides a more detailed descriptions and a set of experiments on *wenn/falls* with both theoretical and methodological implications.

Experiment 4b

Experiment 4b also tested the interaction between the CC SCS with another SCS, namely, by evaluative adverbs (EADV). Liu (2012) presents distributional facts of apparently similar EADVs in German and argues that they differ in terms of factivity. Factive EADVs occur only in veridical contexts, whereas non-factive EADVs are more tolerant, e.g., they can also occur in non-veridical contexts. Without going into detail, their distinction can be illustrated with (43). Both EADVs mean *unfortunately*, but in, for example, questions and conditionals (as non-veridical or entailment-canceling contexts)*, leider* is degraded in comparison to *unglücklicherweise*, which Liu attributes to their difference in degrees of factivity, i.e., speaker commitment.

(43) EADV SCS: $_{\text{More committed}}$ <LEIDER p, UNGLÜCKLICHERWEISE p> $_{\text{Less committed}}$

Experiment 4b, addressing the CC SCS and the EADV SCS, was based on the assumption that the degree of speaker commitment by one expression should be coherent with that of its co-occurring expression. Thus, both CCs should favor non-factive EADVs more than factive ones and factive EADVs should favor *wenn* over *falls* with no difference between the CCs in the case of non-factive EADVs.

Materials and Methods

Experiment 4b used a 3 × 2 factorial within-subjects design, with the factor CONNECTIVE (factive vs. non-factive, i.e., *weil* 'because' vs. *wenn/falls*) and EADV with the levels factive (e.g., *leider*) and non-factive (e.g., *unglücklicherweise*). 36 items such as (44) as well as 72 fillers were used. The critical stimuli are provided in the **Supplementary Materials** (Test sentences of Experiment 4b). The procedure was similar as in Experiment 1, except that the subjects gave naturalness ratings on a 5-point Likert scale (1 = unnatural, 5 = natural). 42 undergraduates (28 females, 14 males; 40 between 18 and 29 years old with 1 under 18 and 1 between 30 and 39) of Osnabrück University participated in the experiment for course credits.

(44) S1: *Katja stellt einen BAföG-Antrag.*
Katja makes a BaföG-application
'Katja applies for BAföG.'

S2: *Weil/Wenn/Falls sie den Abgabetermin*
because/if/if she the deadline
leider/unglücklicherweise verpasst, bittet sie
unfortunately misses, asks she
um eine Fristverlängerung.
for one deadline extension
'Because/If/In case she unfortunately misses the deadline, she will ask for an extension of the deadline.'

Results

All analyses were performed using mixed effects linear regression models. The model was constructed using the

lme4 package in R (Baayen et al., 2008; Bates et al., 2012; R Core Team, 2018). The reported model is the maximal model that converged. The model included CONNECTIVE and EADV (with interaction term) as fixed effects. Furthermore, it included random intercepts for subjects and items, as well as random by-subject and by-item slopes for the effects of CC and EADV.

The results show a highly significant CONNECTIVE × EADV interaction (LRT = 56.92, $p < 0.0001$), see **Table 6** and **Figure 3**. First, *weil*-sentences received significantly higher naturalness ratings overall than either *wenn*- or *falls*-sentences (*weil* vs. *wenn*: $t = 8.19$, $p < 0.0001$; *weil* vs. *falls*: $t = 9.59$, $p < 0.0001$), even though an reviewer pointed out rightly that the *weil*-sentences would have been more natural in present perfect (i.e., *verpasst hat* 'has missed'). Second, for the causal connective *weil* 'because' both factive and non-factive EADVs were rated as equally natural. Third, non-factive EADVs were preferred over factive ones in the case of both CCs (*wenn:* $t = 4.66$, $p = 0.0001$, *falls:* $t = 7.30$, $p < 0.0001$), whereas *wenn* and *falls* did not differ significantly from each other in their ratings.

Discussion

Experiment 4b shows that both CCs are degraded with either type (i.e., factive/non-factive) of EADVs in comparison to the factive causal connective. Concerning CCs, it shows that sentences with either CC are degraded in co-occurrence with factive EADVs in comparison to those with non-factive EADVs. That is, the prediction that factive EADVs should favor *wenn* over *falls* was not borne out. This makes Contrast 5 as in the example of (12) invalid. We can thus conclude that in general, factive EADVs disprefer non-veridical contexts as created by both CCs, compared to non-factive EADVs, which are non-veridical as CCs.

These results have at least the following implications: First, in Karttunen's (1971) term, EADVs are 'semi-factive,' i.e., they lose their factivity in certain contexts including questions, conditionals, and modals (see also Asher, 2000). But earlier work does not make a distinction among EADVs. Experiment 4b shows that EADVs indeed differ in terms of factivity, as argued in Liu (2012). Second, there exists a general constraint on co-occurring expressions with attitudinal meanings, namely, they need to agree with (or at least not clash with) each other. While this constraint needs further qualification and empirical validation, it is probably related to the notions of (in)coherence or (dis)harmony (Lyons, 1977). The factive causal connective is harmonious with both types of EADVs, as EADVs, despite their difference, express high degrees of commitment (i.e., toward full-commitment). Factive EADVs are less harmonious with CCs because the latter are non-factive, which are thus more coherent to combine with non-factive EADVs.

Furthermore, all ratings for sentences in Experiment 4b are very close to the midpoint of the scale, with the exception of sentences with *weil*. In other words, conditionals in general tend to be odd when antecedents are marked by EADVs. While this is an interesting result as far as the German factive and non-factive EADVs are concerned, it suggests that this might not be a useful manipulation for examining the differences between *wenn* and *falls*. If we compare Experiment 4a and 4b, there is one difference in the design in that in Experiment 4a, the two levels of the NPI factor were manipulated via the absence or the presence of NPIs, whereas the two levels of the EADV factor was manipulated via two different kinds of EADVs, not including a third level without EADVs. This difference is crucial in understanding the results: the *wenn/falls* contrast was significant in the NPI-absent conditions but not in the NPI-present conditions of Experiment 4a, whereas there

TABLE 6 | Descriptive statistics of Experiment 4b.

Condition	Connective	EADV	Rating	*SE*
1	*falls*	Factive	2.04	0.07
2	*falls*	Non-factive	2.89	0.08
3	*weil*	Factive	3.61	0.08
4	*weil*	Non-factive	3.60	0.08
5	*wenn*	Factive	2.19	0.07
6	*wenn*	Non-factive	2.73	0.09

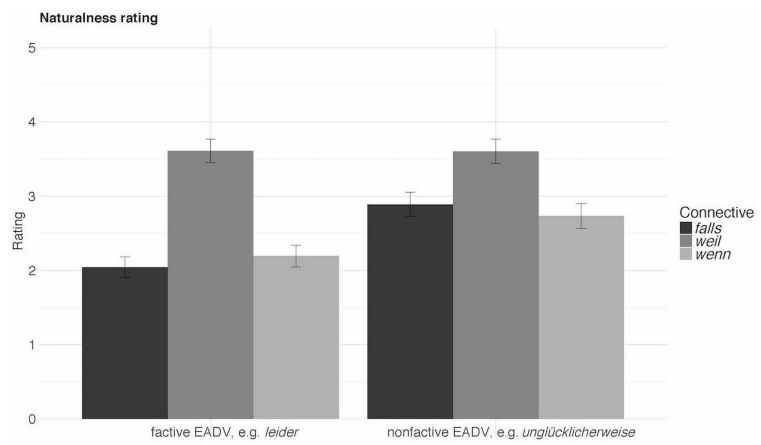

FIGURE 3 | Results of Experiment 4b.

was no difference between *wenn* and *falls* in Experiment 4b, which was potentially due to the lack of a cleaner comparison condition without EADVs.

GENERAL DISCUSSION AND CONCLUSION

In this paper, I provided distributional properties of the two German CCs *wenn* and *falls* and argued that, while they are semantically both non-veridical, they differ in lexical pragmatics in that *falls* conveys a weaker speaker commitment toward the antecedent proposition than *wenn.*

In Experiment 1, subjects rated CCs in combination with events in the antecedent with varying degrees of likelihood by common knowledge. The results did not show any effect and thus were unable to confirm the proposed analysis. This might be due to the conversational nature of the meaning difference or due to the use of the binary scale which was not sensitive enough to measure subtle lexical pragmatics. Experiment 2 used the forced lexical choice task. It showed that *wenn* was preferred over *falls* in contexts where the protagonist had a high degree of credence in the antecedent proposition, and vice versa in contexts where the protagonist had a low degree of credence in the antecedent proposition. In Experiment 3, the conditional sentences were combined with a RC attached to the conditional antecedent, which conveyed the speaker's high or low degree of credence in the antecedent proposition. Overall, it showed an interaction of RC and CC, with *wenn* being rated more natural than *falls* in the belief-condition, and vice versa in the disbelief-condition. This is in line with the proposed *wenn/falls* contrast and the results of Experiment 2. A closer look at the data revealed further differences due to the used RCs (*glauben/nicht glauben* vs. *un/wahrscheinlich*). With the RC containing *(nicht) glauben,* the experiment shows that the speaker can express positive or negative bias toward the antecedent proposition in the RCs, without causing incoherence with *wenn* or *falls.* This means that the lexical contrast between *wenn* and *falls* is part of the pragmatic (i.e., non-at-issue) rather than semantic or conventional meaning.

Experiment 4a (Liu, 2019), as summarized above, provides strong evidence for the proposed lexical pragmatic contrast between *wenn* and *falls.* Combining it with the results of Experiment 2, I argue that the proposed *wenn/falls* contrast is real. Experiment 4b tested connectives and EADVs in co-occurrence, and shows factive EADVs disprefer non-veridical CCs in comparison to non-factive EADVs. But it did not show a difference between the two CCs. In combination with the results of Experiments 1–3, this means that the *wenn/falls* contrast is subject to contextual modulations, i.e., it can be more visible in some and less so in others. In general, Experiment 4b also provides a first step toward understanding the interaction of co-occurring attitudinal expressions.

While CCs are argued to have no conditional meaning in the restrictor analysis, this paper shows that they can differ in meaning. The current study, in particular Experiment 2 and 3 in combination with the results of Liu (2019), provides evidence that the two frequently used German CCs *wenn* and *falls* differ in lexical pragmatics. The non-at-issue meanings of *wenn/falls* are reinforceable and cancellable, indicating their conversational nature and explaining the contextual effects found in the experiments. I relate their difference to the conveyed doxastic assumptions of the speaker, i.e., they express different degrees of speaker commitment. However, alternative analyses are possible.

One alternative is that the higher degree of speaker commitment in the case of *wenn* (in comparison to *falls*) may be due to its ambiguity between conditional and temporal interpretations (and the lack of ambiguity for *falls*), as temporal adverbial clauses are typically presupposed (Levinson, 1983, among others) and therefore factive. While we have seen examples with clearly conditional, non-temporal meaning, such as the biscuit-conditional in (4), the counterfactual conditional in (7), the conditionals with NPIs in (11) and (40), and the conditionals with EADVs in (12) and (41), we cannot rule out the possibility of the interference by the temporal reading of *wenn.* In fact, the results of Experiment 4a and 4b are in line with this possibility: in Experiment 4a, the *wenn/falls* contrast is significant without NPIs, that is, when the temporal interpretation is possible, whereas the difference is not visible with NPIs, i.e., when the temporal interpretation is not possible. In Experiment 4b, as the temporal reading was not possible across all the conditions, we were not able to detect any difference. However, I do not think this contradicts the current proposal for the *wenn/falls* contrast in terms of speaker commitment. In fact, Breindl et al. (2014, p. 265f) have argued that the choice of non-ambiguous *falls* leads to an implicature and that, in order to avoid the implicature, the speaker can consciously choose to use *wenn*, or vice versa, as we see in the examples of (28), possibly with the help of intonation: native speakers confirm that stressed *falls* strengthens the proposed WUI implicature. To sum up, I think the presence and absence of the temporal reading can be seen as a source (possibly out of several) for the *wenn/falls* contrast, with the implicature being the consequence in the choice. Here are two independent examples to illustrate the point.

With (13), we presented the difference between the necessity modal verb and the unmodalized variant in that the former triggers a weakened speaker commitment: $_{\text{More committed}}<$ unmodalized p, MUST p, POSSIBLY p$>_{\text{Less committed}}$. It is unclear, however, how the difference arises. But the lexical choice of MUST can lead to the implicature linked to weakened speaker commitment, just as the choice between *falls* vs. *wenn.* Similarly, in (45), the speaker can use indicative mood or subjunctive I (Konjunktiv I) mood in the verb. Potts (2005, pp. 186, 187) argues that German Konjunctiv I is used to indicate the speaker's wish "to distance himself from the propositional content expressed" or that "the speaker is not publicly committed to the truth of *p*," but "It does not indicate that the speaker is committed to the negation of the propositional content in question."

(45) *Maria sagt, dass sie krank ist/sei.*
Maria says that she sick is/be.
'Maria says that she is sick.'

In other words, the choice between *ist/sei* indicates the degree to which the speaker intends to distance themselves from the given proposition. At the same time, the choice of *sei* over *ist* can give rise to an implicature, just as the choice of *falls* over *wenn* can. In this regard, it is also worth noting that Elder and Jaszczolt (2016) have put forward the notion of "remoteness from reality" in the context of conditionals. While their point is concerned with conditionals in general, e.g., *if* (with remoteness from reality, with regard to the antecedent) in comparison to *since/when* (without remoteness, i.e., alignment with reality), which they attribute to Grice (Elder and Jaszczolt, 2016: 41), the idea may nevertheless be relevant for understanding and modeling the *wenn/falls* contrast. The choice between them is then a choice between remoteness from or alignment with reality; the choice for *falls* over *wenn* can equally lead to an implicature of more focused or increased remoteness. These alternatives provide interesting perspectives that can help us to understand the mechanism behind lexical choices and implicatures, but I also think they are not incompatible with the current proposal.

It is to note, however, that this meaning difference is probably not the only aspect in which the two CCs differ. Consider (46): The sentence can have a conditional reading as in (46b), for which it is fine to replace *falls* with *wenn*. But it also has the reading in (46b), where *falls* can be best translated to "(just) in case" in English. The resulting sentence and its interpretation are different from canonical *falls*-conditionals and it is inappropriate to use *wenn*. Whether the contrast in terms of speaker commitment plays a role here will be left for future research.

(46) *Falls es regnen sollte, nehme ich einen*
if it rain should, take I an
Regenschirm mit.
umbrella with

(a) 'If it should rain, I will take an umbrella with me.'
(b) 'Just in case it should rain, I'm taking an umbrella with me.'

Secondly, *falls* does not always convey negative doxastic bias but sometimes it can convey negative bouletic bias. For example, in (47), the use of *falls* is compatible with the speaker's dispreference but incompatible with their preference for the modified event.

(47) *Kauf kein/#ein Auto. Falls Du unbedingt eins kaufen*
buy no/a car if you definitely one buy
willst, nimm ein Elektroauto.
want, take an electric car
'Don't buy/Buy a car. If you really want to buy one, take an electric car.'

Methodologically speaking, this paper also shows that detection and validation of subtle differences in lexical pragmatics can be methodologically challenging. For example, the forced lexical choice study in Experiment 2 as well as the rating study using the speaker belief judgment task in Experiment 4a (Liu, 2019) show clear positive evidence for the contrast between *wenn/falls*. However, there was no evidence in the rating study in Experiment 1 and only weaker evidence in the rating study in Experiment 3. It is worth noting again that Experiments 1 and 3 used binary rating scales[19], and Experiment 4b used a 5-point Likert scale, whereas Likert scales with less than 7 points are argued to be problematic (see, e.g., Liddell and Kruschke, 2018). Additionally, since the *wenn/falls* difference is of the conversational nature and supposed to be nuanced, a higher number of points (e.g., a 10-point scale) might be needed to make the scale sensitive enough to capture the difference, which I leave for future studies. A final note on the limitation of the current study is that the critical items of different numbers were used in combination with other experimental materials, which should be avoided in future to avoid potential confounds. I report experiments with or without evidence here to hopefully help future studies on testing lexical pragmatic differences.

In general, the results in the case study of *wenn/falls* also call for reconsiderations of Grice's notion of 'implicature' from a probabilistic perspective, e.g., to model lexical pragmatics of near synonyms. We need a more gradable notion of implicature than the conventional and conversational distinction to model lexical semantics and pragmatics. Each case of near-synonyms has its own story in that the distance between them is gradient (cf. experimental evidence) rather than categorical. While this paper does not provide a general integrated theory for this purpose, it showcases the usefulness of speaker commitment scales as a formal tool for modeling lexical pragmatic contrast and the benefits of combining theoretical and experimental perspectives.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committee of Osnabrück University. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

ML conceived and conducted all the reported studies and wrote the manuscript.

[19]We used different scales in Experiments 1, 3, and 4 as they were conducted with different sub-experiments together.

FUNDING

This work has been supported by a grant from the DFG (German Research Foundation) to ML's project "The Semantics and Pragmatics of Conditional Connectives: Cross-linguistic and experimental Perspectives" (project number: 367088975).

ACKNOWLEDGMENTS

I presented part of the manuscript at the following occasions: Regine Eckardt's Anniversary Workshop at Göttingen (2015), the SPIRIT workshop on Negation at the University of Göttingen (2015), the Department of Philosophy at Peking University (2016), the Linguistic Colloquium at the University of Stuttgart (2016), the annual meeting of XPrag.de at the University of Cologne (2017), the Department of Philosophy at Beijing Normal University (2018), the Linguistics and Philosophy Workshop at the University of Chicago (2018), the Berkeley Germanic Linguistics Roundtable at the University of California Berkeley (2018), the German Department at the University of Cologne (2019), and the Workshop on Concessives vs. Adversatives: Opposing Opposition at the 41th annual meeting of the German Linguistics Society, University of Bremen (2019). Among others, I thank Andreas Blümel, Peter Bosch (†), Ellen Brandner, Elena Castroviejo, Regine Eckardt, Lyn Frazier, Itamar Francez, Berit Gehrke, Anastasia Giannakidou, Klaus von Heusinger, Stefan Hinterwimmer, Daniel Hole, Fengkui Ju, Chris Kennedy, Manfred Krifka, Laia Mayol, Janina Radó, Stephanie Rotter, Marvin Schmitt, Ming Xiang, and Alan Yu for their feedback on different parts and pieces of the related work. Furthermore, I owe special thanks to Juliane Schwab and Stephanie Rotter for their help with data acquisition as well as her very useful comments on an earlier version of the manuscript, and Matthias Andreß for proofreading the manuscript. Last but not least, I thank all the four reviewers for their invaluable comments; addressing these has really helped me to improve the previous versions. All mistakes are my own, of course.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.629177/full#supplementary-material

REFERENCES

Anderson, A. R. (1951). A note on subjunctive and counterfactual conditionals. *Analysis* 12, 35–38. doi: 10.1093/analys/12.2.35

Arregui, A. C. (2005). *On the Accessibility of Possible Worlds: the Role of Tense and Aspect*. Ph.D thesis. Amherst, MA: UMass Amherst.

Arregui, A., and Biezma, M. (2016). Discourse rationality and the counterfactuality implicature in backtracking conditionals. *Proc. Sinn Und Bedeutung* 20, 91–108.

Asher, N. (2000). Truth conditional discourse semantics for parentheticals. *J. Semant.* 17, 31–50. doi: 10.1093/jos/17.1.31

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

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. M., Maechler, M., and Bolker, B. (2012). *lme4: Linear Mixed-Effects Models UsingS4 Classes (R Package Version 0.999999-0)*. Available online at: http://cran.r-project.org/web/packages/lme4/index.html (accessed July 6, 2021).

Bergena, G., and Boskerb, H. R. (2018). Linguistic expectation management in online discourse processing: an investigation of Dutch *inderdaad* 'indeed' and *eigenlijk* 'actually'. *J. Mem. Lang.* 103, 191–209. doi: 10.1016/j.jml.2018.08.004

Bhatt, R., and Pancheva, R. (2006). "Conditionals," in *The Blackwell Companion to Syntax*, eds M. Everaert and H. van Riemsdijk (Oxford: Blackwell), 638–687.

Breindl, E., Volodina, A., and Waßner, U. (2014). *Handbuch der Deutschen Konnektoren 2: Semantik der Deutschen Satzverknüpfer* (= "Schriften des Instituts für Deutsche Sprache 13.1-13.2). Berlin: de Gruyter.

Declerck, R., and Reed, S. (2001). *Conditionals: a Comprehensive Empirical Analysis. (Topics in English Linguistics 37.)*. New York, NY: De Gruyter Mouton.

Dostie, G. (1987). Etude semantique de quatre connecteurs conditionnels: à condition que, pourvu que, en autant que et d'abord que. *Fr. Mod.* 55LV(3/4), 174–203.

Elder, C.-H., and Jaszczolt, K. M. (2016). Towards a pragmatic category of conditionals. *J. Pragmat.* 98, 36–53. doi: 10.1016/j.pragma.2016.04.013

Geurts, B. (2019). Communication as commitment sharing: speech acts, implicatures, common ground. *Theor. Linguist.* 45, 1–30. doi: 10.1515/tl-2019-0001

Giannakidou, A. (1998). *Polarity Sensitivity as (Non)veridical Dependency*. Amsterdam-Philadelphia: John Benjamins.

Giannakidou, A. (1999). Affective dependencies. *Linguist. Philos.* 22, 367–421.

Giannakidou, A. (2011). "Positive polarity items and negative polarity items: variation, licensing, and compositionality," in *Semantics: an International Handbook of Natural Language and Meaning*, 2nd Edn, eds C. Maienborn, K. von Heusinger, and P. Portner (Berlin: Mouton de Gruyter), 1515–1538.

Giannakidou, A. (2014). "The prospective as nonveridical: polarity items, speaker commitment, and projected truth," in *The Black Book. Festschrift for Frans Zwarts*, eds D. Gilberts and J. Hoeksema (Groningen: University of Groningen), 101–124.

Giannakidou, A., and Mari, A. (2016). "Epistemic future and epistemic MUST: non-veridicality, evidence and partial knowledge," in *Mood, Aspect and Modality Revisited*, eds J. Blaszczack, A. Giannakidou, D. Klimek-Jankowska and K.Migdalski (Chicago: University of Chicago Press), 75–118.

Giannakidou, A., and Mari, A. (2021). *Truth and Veridicality in Grammar and Thought: Modality, Mood, and Propositional Attitudes*. Chicago, IL: University of Chicago Press.

Grice, H. P. (1975). "Logic and conversation," in *Syntax and Semantics, Speech Acts*, Vol. 3, eds P. Cole and J. L. Morgan (New York, NY: Academic Press), 41–58. doi: 10.1163/9789004368811_003

Grice, H. P. (1989). *Studies in the Way of Words*. Cambridge, MA: Harvard University Press.

Grosz, P. (2012). *On the Grammar of Optative Constructions*. Amsterdam: John Benjamins.

Haegeman, L. (2003). Conditional clauses: external and internal syntax. *Mind Lang.* 18, 317–339. doi: 10.1111/1468-0017.00230

Hoeksema, J. (2012). On the natural history of negative polarity items. *Linguist. Anal.* 38, 3–34.

Iatridou, S. (1991). *Topics in Conditionals*. PhD thesis. Cambridge, MA: MIT.

Iatridou, S. (2000). The grammatical ingredients of counterfactuality. *Linguist. Inq.* 31, 231–270. doi: 10.1162/002438900554352

Ippolito, M., and Su, J. (2014). "Counterfactuals, negation and polarity," in *The Art and Craft of Semantics: a Festschrift for Irene Heim*, eds L. Crnič and U. Sauerland (Cambridge, MA: MITWPL), 225–243.

Karttunen, L. (1971). Some observations on factivity. *Pap. Linguist.* 4, 55–69. doi: 10.1080/08351817109370248

Kratzer, A. (1986). Conditionals. *Chic. Linguist. Soc.* 22, 1–15.

Kratzer, A. (1991). "Conditionals," in *Semantics: an International Handbook of Contemporary Research*, eds A. von Stechow and D. Wunderlich (Berlin: De Gruyter), 651–656.

Krifka, M. (2015). Bias in commitment space semantics: declarative questions, negated questions, and question tags. *Proc. Semantics Linguist. Theory (SALT)* 25, 328–345. doi: 10.3765/salt.v25i0.3078

Krzyżanowska, K. (2019). What is wrong with false-link conditionals? *Linguist. Vanguard* 5:20190006.

Krzyżanowska, K., Collins, P., and Hanh, U. (2017). Between a conditional's antecedent and its consequent: discourse coherence vs. probabilistic relevance. *Cognition* 164, 199–205. doi: 10.1016/j.cognition.2017.03.009

Ladusaw, W. (1980). *Polarity Sensitivey as Inherent Scope Relations*. New York, NY: Garland.

Léard, J.-M. (1987). La syntaxe et la classification des conditionnelles et des concessives. *Fr. Mod.* 55 LV(3/4), 158–173.

Levinson, S. (1983). *Pragmatics*. Cambridge, MA: Cambridge University Press.

Liddell, T. M., and Kruschke, J. K. (2018). Analyzing ordinal data with metric models: what could possibly go wrong? *J. Exp. Soc. Psychol.* 79, 328–348. doi: 10.1016/j.jesp.2018.08.009

Liu, M. (2012). *Multidimensional Semantics of Evaluative Adverbs*. Boston, MA: Brill.

Liu, M. (2019). The elastic nonveridicality property of indicative conditionals. *Linguist. Vanguard* 5:20190007.

Liu, M., and Barthel, M. (2021). *Semantic Processing of Conditional Connectives–German Wenn 'if' vs. Nur Wenn 'Only If'*. Unpublished manuscript, Humboldt University of Berlin.

Liu, M., König, P., and Mueller, J. L. (2019). Novel ERP evidence for processing differences between negative and positive polarity items in German. *Front. Psychol.* 10:376. doi: 10.3389/fpsyg.2019.00376

Liu, M., and Wang, Y. (2021). "Speaker commitment by Mandarin conditional connectives: distributional and experimental evidence," in *Proceedings of the 21th Chinese Lexical Semantics Workshop (CLSW2020). In Springer's Lecture Notes in Computer Science*, (Hong Kong: University of Hong Kong).

Liu, M., Rotter, S., and Giannakidou, A. (2021). *Bias and Modality in Conditionals: Experimental Evidence and Theoretical Implications*. Unpublished manuscript, Humboldt University of Berlin.

Lyons, J. (1977). *Semantics*. Cambridge, MA: Cambridge University Press.

Montoliio, E. (2000). "On affirmative and negative complex conditional connectives," in *Cause, Condition, Concession, Contrast*, eds E. Couper-Kuhlen and B. Kortmenn (New York, NY: Mouton de Gruyter), 143–171. doi: 10.1515/9783110219043-007

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

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. Foundation for Statistical Computing, Vienna, Austria. Available online at: https://www.R-project.org

Reis, M., and Wöllstein, A. (2010). Zur Grammatik (vor allem) konditionaler V1-Gefüge im Deutschen. *Z. Sprachwiss.* 29, 111–179.

Schwenter, S. (2001). Expectations and (in)sufficiency: Spanish como conditionals. *Linguistics* 39, 733–760.

Simons, M., Tonhauser, T., Beaver, D., and Roberts, R. (2010). "What projects and why," in *Proceedings of the Semantics and Linguistic Theory* (SALT), Vol. 21, (Ithaca, NY: CLC Publications), 309–327. doi: 10.3765/salt.v0i20.2584

Skovgaard-Olsen, N., Singmann, H., and Klauer, K. C. (2016). The relevance effect and conditionals. *Cognition* 150, 26–36. doi: 10.1016/j.cognition.2015.12.017

Sode, F., and Sugawara, A. (2019). "On the deliberative use of the German modal sollte," in *New Frontiers in Artificial Intelligence. JSAI-isAI 2018, LNCS 11717*, eds K. Kojima, M. Sakamoto, K. Mineshima, and K. Satoh (Cham: Springer), 341–356. doi: 10.1007/978-3-030-31605-1_25

Starr, W. (2019). *Counterfactuals*. Available online at: https://plato.stanford.edu/entries/counterfactuals/ (accessed July 6, 2021).

Tonhauser, J. (2012). Diagnosing (not-)at-issue content. *Proc. Semant. Represent. Lang. Am.* 6, 239–254.

Visconti, J. (1996). On English and Italian complex conditional connectives: matching features and implicatures in defining semanto-pragmatic equivalence. *Lang. Sci.* 18, 549–573. doi: 10.1016/0388-0001(96)00034-4

Volodina, A. (2006). "Wenn-relationen: schnittstelle Zwischen syntax, semantik und pragmatik," in *Grammatische Untersuchungen, Analysen und Reflexionen. Festschrift für Gisela Zifonun*, eds E. Breindl, L. Gunkel, and B. Strecker (Tübingen: Narr), 359–379.

Volodina, A. (2011). *Konditionalität und Kausalität im Diskurs. Eine korpuslinguistische Studie zum Einfluss von Syntax und Prosodie auf die Interpretation komplexer Äußerungen*. Tübingen: Narr.

von Fintel, K. (2007). *If: The Biggest Little Word*. Available online at: http://mit.edu/fintel/gurt-slides.pdf (accessed July 6, 2021).

von Fintel, K. (1999). NPI licensing, strawson entailment, and context dependency. *J. Semant.* 16, 97–148. doi: 10.1093/jos/16.2.97

von Fintel, K. (2011). "Conditionals," in *Semantics: an International Handbook of Natural Language and Meaning*, eds C. Maienborn, K. von Heusinger, and P. Portner (Berlin: de Gruyter), 1515–1538.

Zaefferer, D. (1991). "Conditionals and unconditionals: cross-linguistic and logical aspects," in *Semantic Universals and Universal Semantics*, ed. D. Zaefferer (Dordrecht: Foris), 210–236. doi: 10.1515/9783110870527-011

Zakkou, J. (2018). The cancellability test for conversational implicatures. *Philos. Compass* 93/3:e12552.

Zimmermann, M. (2004). Zum "Wohl": Diskurspartikeln als Satztypmodifikatoren. *Linguist. Ber.* 199, 1–35.

Conflict of Interest: 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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Liu. 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) and the copyright owner(s) 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: 19 August 2021
doi: 10.3389/fpsyg.2021.623648

Referential Chains Reveal Predictive Processes and Form-to-Function Mapping: An Electroencephalographic Study Using Naturalistic Story Stimuli

Ingmar Brilmayer and Petra B. Schumacher*

Department of German Language and Literature I–Linguistics, University of Cologne, Cologne, Germany

In discourse pragmatics, different referential forms are claimed to be indicative of the cognitive status of a referent in the current discourse. Referential expressions thereby possess a double function: They point back to an (existing) referent (form-to-function mapping), and they are used to derive predictions about a referent's subsequent recurrence in discourse. Existing event-related potential (ERP) research has mainly focused on the form-to-function mapping of referential expression. In the present ERP study, we explore the relationship of form-to-function mapping and prediction derived from the antecedent of referential expressions in naturalistic auditory language comprehension. Specifically, the study investigates the relationship between the form of a referential expression (pronoun vs. noun) and the form of its antecedent (pronoun vs. noun); i.e., it examines the influence of the interplay of predictions derived from an antecedent (forward-looking function) and the form-to-function mapping of an anaphor (backward-looking function) on the ERPs time-locked to anaphoric expressions. The results in the time range of the P300 and N400 allow for a dissociation of these two functions during online language comprehension.

Keywords: prominence, reference, prediction, form-to-function mapping, naturalistic stimuli, P300, N400

OPEN ACCESS

Edited by:
Pietro Perconti,
University of Messina, Italy

Reviewed by:
Paolo Canal,
University Institute of Higher Studies in Pavia, Italy
Sergey Avrutin,
Utrecht University, Netherlands

****Correspondence:***
Ingmar Brilmayer
ingmar.brilmayer@uni-koeln.de

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

Received: *30 October 2020*
Accepted: *12 July 2021*
Published: *19 August 2021*

Citation:
Brilmayer I and Schumacher PB (2021) Referential Chains Reveal Predictive Processes and Form-to-Function Mapping: An Electroencephalographic Study Using Naturalistic Story Stimuli. Front. Psychol. 12:623648. doi: 10.3389/fpsyg.2021.623648

INTRODUCTION

It is a common observation in pragmatic research on discourse structure that different referential forms are indicative of the cognitive status of a referent in the mind of the speaker, as well as of the cognitive status that a speaker assumes in a hearer (e.g., Prince, 1981; Givón, 1983; Ariel, 1990; Gundel et al., 1993). Accordingly, specific referential forms, such as personal pronouns, demonstratives, full noun phrases, or names, can be seen as pointers to the cognitive status of a discourse referent. In the literature, various approaches to discourse structure include a notion of this cognitive status as a key component, such as *salience, attentional focus, accessibility, referential activation*, *givenness*, or *prominence* (Chafe, 1976; Grosz and Sidner, 1986; Ariel, 1990; Gundel et al., 1993; Grosz et al., 1995; Lambrecht, 1996; Chiarcos et al., 2011; Falk, 2014; von Heusinger and Schumacher, 2019). Here, we follow the prominence account to the cognitive status of referents (see Himmelmann and Primus, 2015; von Heusinger and Schumacher, 2019, for details) which aims at a precise characterization of the cognitive

status of discourse referents on the basis of linguistic prominence features (e.g., thematic role, syntactic function, and definiteness). In its discourse-pragmatic formulation (von Heusinger and Schumacher, 2019), the prominence account rests on three basic definitions: (1) Prominence is a relational property that singles out one element from a set of elements of equal rank (e.g., two discourse referents) (2) it shifts in time, e.g., the prominence status of a referent can change, while a discourse unfolds, and (3) prominent referents are structural attractors, i.e., they attract linguistic operations, such as serving as perspectival anchors or licensing more referential variation. In the present event-related potential (ERP) study, we focus on criteria (1) and (3). Specifically, we investigate the relationship between the form of a referential expression (pronoun vs. noun) and the form of its antecedent (pronoun vs. noun), i.e., we examine the contribution of referential chains, i.e., the interplay of antecedent and anaphor during referential processing.

A widely employed indicator for the prominence of referents is the referential form that is used to refer to them. For example, personal pronouns (or other phonetically light forms) with anaphoric function have been claimed to refer to the most prominent entity in the current discourse, while phonetically richer forms, such as full noun phrases, are used to refer to less prominent or newly introduced referents (e.g., Givón, 1983; Ariel, 1990; Gundel et al., 1993; von Heusinger and Schumacher, 2019). In this sense, personal pronouns select the most prominent discourse referent, which is singled-out from all other (less prominent) discourse referents (definition 1). A well-known consequence of this form-to-function mapping of referential expressions is the so-called repeated name penalty (Gordon et al., 1993, 1999; Gordon and Hendrick, 1997; Gelormini-Lezama and Almor, 2011; Almor et al., 2017). Almor et al. (2017), for instance, tested a prominent referent that was introduced by a proper name and was rementioned with the same name again instead of a personal pronoun (e.g., "John went to the store. John/He wished to buy some candy."; Almor et al., 2017, p. 56) and found that this repetition results in processing costs. Importantly, the authors found that a repeated name with a non-prominent referent, for instance a conjoined noun phrase (e.g., "John and Mary went to the store. John/He wished to buy some candy."; Almor et al., 2017, p. 56), did not elicit a repeated name penalty, exemplifying the critical role of prominence information in the establishment of coreference (see also for other non-prominent antecedents, like objects, Gordon et al., 1993; Almor, 1999; Burkhardt and Roehm, 2007; Almor and Eimas, 2008).

Moreover, prominent referents allow for more variability in the referential expressions that can be used to refer to them, i.e., they are *structural attractors* (definition 3). Gundel et al. (1993) already note that a prominent referent (a referent "in focus" in Gundel et al.'s terminology) is preferably referred to by an unstressed personal pronoun or a zero marked expression, yet it might also be referred to by a definite description or a proper name. Yet, less accessible referents can only be referred to by a more limited set of referential expressions. For instance, a newly introduced referent can (usually) not be introduced by a definite description, but must be introduced by an indefinite description. Here, we subsume this line of research under the term *backward-looking function of referential expressions* (cf. Givón, 1983): It focuses on the mapping of the referential form of an anaphor to referents in a discourse model (form-to-function mapping). However, as von Heusinger and Schumacher (2019) argue in accordance with Givón (1983), referential expressions also possess a *forward-looking* or *discourse structuring potential*: Prominent referents have a higher probability to recur in subsequent discourse, preferably with a personal pronoun or other phonetically light expressions (Givón, 1983; for behavioral and electrophysiological evidence, see Brocher and von Heusinger, 2018; Fuchs and Schumacher, 2020). In other words, prominent referents have a stronger influence on the way a discourse unfolds, than non-prominent referents; i.e., they attract linguistic operations (definition 3) at the discourse level.

In the present research, we explore the relationship of the form of a referential expression and the form of its antecedent during online language comprehension using ERP. For this purpose, we analyzed electroencephalographic (EEG) data originally recorded by Brilmayer et al. (2019), who used a German audio book recording of The Little Prince by de Saint-Exupéry (2012) as experimental stimulus. The recording is annotated for a wide range of linguistic features (e.g., syntactic function, thematic role, case, number, part-of-speech, and referential features) and physical properties (e.g., pitch contour and speech envelope) but also for the text-analytic measures proposed by Givón (1983). Here, we contrast referential chains with different referential forms. Based on their particularly strong prominence contrast, we chose to contrast anaphoric nouns and pronouns with noun or pronoun antecedents, resulting in four conditions: pronouns with a pronoun antecedent [pronoun-pronoun chain (1)], pronouns with a noun antecedent [noun-pronoun chain (2)], nouns with a noun antecedent [noun-noun chain (3)], and nouns with a pronoun antecedent [pronoun-noun chain (4)].

1. **She** (the flower) adjusted her petals one by one. **She** did not wish to go out into the world all rumpled, like the field poppies (The Little Prince, chapter 8).
2. But **the flower** was not satisfied to complete the preparations for her beauty in the shelter of her green chamber. **She** chose her colors with the greatest care (The Little Prince, chapter 8).
3. I have serious reason to believe that the planet from which the little prince came is **the asteroid known as B—612. This asteroid** has only once been seen through the telescope (The Little Prince, chapter 4).
4. But he was in Turkish costume, and so nobody would believe what he said. Grown—ups are like that. Fortunately, however, for the reputation of Asteroid B—612, a Turkish dictator made a law that his subjects, under pain of death, should change to European costume. So in 1920 **the astronomer** gave his demonstration all over again, dressed with impressive style and elegance (The Little Prince, chapter 4).

Based on the literature on referential form and prominence (e.g., Givón, 1983; Gundel et al., 1993; Arnold, 1998; Kehler et al., 2008; von Heusinger and Schumacher, 2019), we assume that referents of pronouns with pronoun antecedent are the most prominent referents in our current comparison, because both the antecedent and anaphoric expression clearly mark their referent as prominent. They are followed by pronoun anaphors with noun antecedent, since nouns mark a referent as less prominent than pronouns but having a pronoun anaphor enhances the prominence status of the referent (e.g., Givón, 1983). Regarding noun anaphors, pronoun-noun chains are the most unlikely type of the four present referential chains with regard to the prominence information provided by the anaphor and antecedent: In this case, a referent established as prominent (reference *via* pronoun) is continued with an expression marking it non-prominent, which constitutes a discourse structural mismatch (as long as the referent is still accessible in memory). Note however that across a longer narrative, a pronoun-noun chain is likely in cases where the referent must be reactivated after a longer sequence without any mention. Noun anaphors with noun antecedent, by contrast, are very common, for instance in referential chains consisting of an indefinite noun phrase antecedent and a definite noun anaphor or to avoid referential ambiguity. In summary, we propose the following prominence ranking for the four referential chains under examination: pronoun-pronoun > noun-pronoun > noun-noun > pronoun-noun. Before we move on to the discussion of previous ERP studies and our experimental hypothesis, we would like to elaborate on the neurobiological understanding underlying our assumptions and the interpretation of our results.

In the present manuscript, we base our hypotheses and the discussion of the results on a predictive coding account to ERPs (Friston, 2005), which is formulated in detail for language-related ERPs in Bornkessel-Schlesewsky and Schlesewsky (2019). Overly simplified (see Friston, 2005, for mathematical details), the predictive coding framework rests on the assumption that the (human) brain actively creates explanations for the causes of its own sensory inputs (e.g., Friston, 2005). The brain achieves this *via* an internal, hierarchically organized generative model of the world, thereby constantly mapping (hidden) causes to sensory consequences (predictive coding). This internal model is constantly checked against actual sensory input (hypothesis testing). When there is a mismatch between the internal model and the sensory input, prediction error arises, leading to an instant update of the internal model. Predictive coding and hypothesis testing occur at multiple, hierarchically organized levels, starting with low levels with short timescales at which very precise predictions are generated (in language, e.g., the level of individual phones), to higher-order (conceptual) levels with increasingly imprecise, more general ("conceptual") predictions (e.g., word meaning, sentences, and discourse structure). Within the framework proposed by Bornkessel-Schlesewsky and Schlesewsky (2019), the N400 and other language-related negativities reflect prediction error at different levels of linguistic representation, while positivities, such as the P300/P600, are related to attentional gain control (see Sassenhagen et al., 2014; Sassenhagen and Bornkessel-Schlesewsky, 2015, for further discussion). Here, we focus on the P300 and N400 ERP components because of their relevance in language processing (cf. Roehm et al., 2007; Kutas and Federmeier, 2011).

Event-related potential studies have provided empirical evidence for the relevance of prominence information in linking anaphoric expressions to a referent in discourse during online language comprehension. The most well-researched ERP component in this respect is the N400, a vertex-negative component of the human ERP, peaking at roughly 300–500 ms after the onset of a stimulus with a posterior maximum. Often interpreted as a specific correlate of linguistic meaning processing, the N400 rather reflects activity in widely distributed, heavily interacting neural networks underlying the comprehension of meaning in general (cf. Kutas and Federmeier, 2011). In the discourse literature, N400 effects are, for example, reported in replications of the repeated name penalty using ERPs (Swaab et al., 2004; Camblin et al., 2007; Ledoux et al., 2007; Almor et al., 2017). Camblin et al. (2007), for instance, found increased N400 amplitudes following repeated name anaphors as compared to pronoun anaphors. This effect was absent, when a repeated name referred to a referent that formed part of a conjoined phrase (e.g., "John and Mary went to the store, so that John …"). Results of a study by Schumacher et al. (2015) point into the same direction. They contrasted ERPs following personal pronouns and demonstrative pronouns in German (e.g., "Der Feuerwehrmann will den Jungen retten …, aber er/der hat" and "The firefighter wants to rescue the boy …, but he/Dem has …") and found more pronounced N400 amplitudes following demonstrative pronouns, as compared to personal pronouns. They attribute this effect to differences in the form-to-function mapping of the two types of expressions: While personal pronouns are highly expected and single out the most prominent referent in a discourse model (which is considered the ideal referent), demonstrative pronouns explicitly exclude coreference with the most prominent referent. According to the authors, this additional information ("Exclude the default referent!") is reflected in an increase in the N400 component. Streb et al. (2004) reported an N400 for increased distance (measured in sentences) between anaphor and antecedent. Similar results have been reported with regard to various linguistic prominence features, for instance, givenness (Burkhardt, 2006; Schumacher and Hung, 2012), topicality (Hung and Schumacher, 2012; Wang and Schumacher, 2013), animacy (Nieuwland and van Berkum, 2006; Hung and Schumacher, 2012), or parallel structure/role (Streb et al., 1999). Overall, the N400 in referential processing can be considered to reflect a mismatch with regard to prominence-based predictions.

ERP studies of referential processes frequently also report a late positivity (P600) following the N400 which is usually interpreted as a correlate of revision in a wide sense (van Berkum et al., 1999; Schmitt et al., 2002; Hammer et al., 2005; Burkhardt, 2006; Lamers et al., 2006; van Berkum et al., 2007; Hammer et al., 2008; Schumacher, 2009; Brouwer et al., 2012; Hung and Schumacher, 2012; Schumacher and Hung, 2012; Schumacher et al., 2015). In Schumacher et al. (2015), for instance, the N400 effect for demonstrative pronouns is followed

by a late positivity effect. They interpret the effect as reflecting updating processes associated with the demonstrative pronoun's referential shift potential. As argued elsewhere (Coulson, 1998; Coulson et al., 1998; Sassenhagen et al., 2014; Sassenhagen and Bornkessel-Schlesewsky, 2015; Brilmayer et al., 2017), we view the P600 as a P300 shifted in latency. There is evidence that highly predictable (i.e., preferred) linguistic input leads to an earlier peaking P300. For instance, Roehm et al. (2007) compared the ERP following sentence final words with a cloze probability near one (antonyms, e.g., "white" in "The opposite of black is white.") with related ("The opposite of black is yellow.") and unrelated words ("The opposite of black is nice."). They found a gradient P300 effect (antonym, related, and unrelated) within the time range of the N400, suggesting an overlap of these components. The authors argue that this P300 reflects a prediction match response, or, more precisely, the absence of prediction error following highly predictable linguistic input. In other words, the meaning of the antonym is already predicted and integrated before the word is encountered. If, as it was the case in the related and unrelated condition, there occurs prediction error with regard to linguistic meaning, the same biphasic N400-P600 pattern as in Schumacher et al. (2015) can be observed. As discussed in Alday and Kretzschmar (2019), the N400 and P300 both seem to be sensitive to prediction, yet while the N400 reflects the processing of stimulus related features (e.g., linguistic prominence features) necessary for categorization, the P300 reflects the categorization process itself. Accordingly, in the absence of prediction error, no further information is needed for stimulus classification, hence the early P300 in Roehm et al. (2007), while with linguistic prediction error, and hence, categorization uncertainty, an N400 arises, reflecting the processing of stimulus features relevant for categorization ("evidence accumulation"). The P300 in turn reflects the categorization process itself, thereby linking perception and (cognitive) (re-)action (cf. Verleger et al., 2005). If we transfer this to referential expressions and the establishment of anaphoric relations, prominent referents (e.g., agents/subjects) are predicted to be continuous in discourse and to be referred to by a personal pronoun. When the pronoun is encountered, the referential relation is already established, since it was predicted, similar to the antonyms in the study by Roehm et al. (2007). Hence, we expect a critical involvement of the P300 in the establishment of referential relations in the present study, especially following personal pronoun anaphors.

Moreover, from this perspective, the N400 associated with the repeated name penalty reflects a mismatch between the predicted referential form (pronoun) and the detected referential form (name). Along these lines, the N400 in Schumacher et al. (2015) reflects a mismatch between the predicted referential form (personal pronoun) and the detected form (demonstrative pronouns), while the positivity indicates attentional reorientation toward the non-prominent referent. In other words, the late P300 in referential comprehension reflects the linking of an unpredicted referential form to an unpredicted antecedent in memory (i.e., its categorization) and its potential consequences for discourse (i.e., referential shift). Burkhardt (2006) investigated different degrees of givenness (coreferential expression vs. inferred expression) and also reported a biphasic pattern: The N400 for inferred expressions reflects a mismatch between the predicted entity and the detected entity, and the positivity represents reorientation toward a new referent.

Evidence for an involvement of an "early" P300 in the processing of referential expressions stems from Brilmayer et al. (2019). In this ERP study using an audio book recording of The Little Prince by de Saint-Exupéry (2012), the authors contrasted pronouns of the first, second, and third person singular with reference to the main character (The Little Prince) or his interlocutors. They found an early peaking positivity (200–300 ms) that was sensitive to linguistic person, indicative of attentional processes. First person pronouns thereby elicited the most positive going amplitudes, followed by third person and second person pronouns. Interestingly, the P300 was insensitive to referent identity, suggesting that early processes driven by linguistically definable features already occur in early time windows preceding the N400. Since we use the same data set in the present study, we expect effects in this time range to occur in our analysis. Before we move on to the experimental methods, we would like to discuss several aspects related to naturalistic designs.

In psycho- and neurolinguistic research, a growing interest in speech and language comprehension under naturalistic conditions is observable (e.g., Schmitt et al., 2002; Brennan and Pylkkänen, 2012; Willems, 2015; Alday et al., 2017; Mak and Willems, 2018; Sassenhagen, 2018; Bhattasali et al., 2019; Brennan et al., 2019; Brilmayer et al., 2019; Schilling et al., 2021). Linguistic research thereby follows a more general trend in the cognitive neurosciences toward a more "realistic" picture of brain processes as they occur during real-life events (cf., for instance, Schilbach et al., 2013) (M) EEG higher-order language studies with naturalistic stimuli are still rare, use auditory short stories as the preferred stimulus type, and span a wide variety of topics: predictive sentence comprehension in participants with autism spectrum disorder (Brennan et al., 2019), syntactic structure building (Brennan and Pylkkänen, 2012; Brennan et al., 2016), lexical frequency (Sassenhagen, 2018), pronouns and linguistic person (Brilmayer et al., 2019), thematic role, case, and syntactic function (Alday, 2019), and content versus function words (Schilling et al., 2021). This diversity makes a direct comparison of the results difficult. Yet, there are commonalities all these studies that are compatible with results of controlled experiments: Alday (2019) found an N400 effect (300–500 ms) associated with thematic role, Sassenhagen (2018) and Alday (2019) report an N400 effect associated with lexical frequency, and Brennan and Pylkkänen (2012) as well as Schilling et al. (2021) provide evidence for an involvement of the N400 in naturalistic language comprehension. This suggests that certain generalizations derived from controlled experiments, in particular the ubiquity of the N400, can serve as a useful starting point for hypotheses generation with naturalistic designs.

One of the great challenges of naturalistic stimuli is that ecological validity ("naturalness") and experimental control are two extremes on a continuum, so that a gain in one leads to a loss in the other (cf. Willems, 2015). Audio book stimuli,

such as the present recording, contain a vast amount of variance outside of experimental control. Besides linguistic variables, such as case, syntactic function, or word order, the speech signal itself is a critical source of variance: differences in formant pitch between words, differences in intensity or duration (e.g., in the present audio book, word durations range from 30 to 2.6 s). The traditional (grand-)averaging method cannot adequately model this variance (see Alday, 2019, for discussion). Here, we therefore follow a more adequate approach to the analysis of EEG data from naturalistic experiments that is based on the linear model and allows the consideration of continuous covariates in the statistical analysis (Sassenhagen and Alday, 2016).

The probably most well-known approach of this kind is the linear regression-based approach to event-related potentials (Smith and Kutas, 2015a,b; Ehinger and Dimegen, 2019). Other than traditional averaging, the rERP framework rests on the assumption that every sample of an EEG signal can be described as the linear combination of different factors with different weights (β-coefficients), i.e., by the linear model. In its mass univariate formulation (cf. Hauk et al., 2006), epoched EEG data are modeled *via* separate linear models for every sample point. For instance, for epoched data from −200 to 800 milliseconds time-locked to a critical word with a sampling rate of 500 Hz, 500 linear models would be calculated, resulting in 500 β-coefficients per factor of interest, one for each sample. These coefficients (or the fitted values) can be treated just like traditional ERPs, for instance for (second-order) statistical analyses. One of the big advantages of this method thereby is that continuous covariates and categorical variables of interest can easily be accounted for within a single model. As noted by Smith and Kutas (2015a), in a perfectly controlled design, the rERP and the ERP approach would yield identical results. With naturalistic stimuli, however, there are considerable differences between the results of traditional grand-averaging and regression-based approaches, because of the uncontrolled variance of the stimulus material of interest. Using the mass univariate rERP method, we are able to separate the brain responses to these (linguistic) nuisance variables from those that are related to the variables of interest (the form of anaphor and antecedent). At this point, it is important to note that we do not use the linear deconvolution approach described in Smith and Kutas (2015a,b). Therefore, we have to keep in mind that our results still contain overlapping brain responses to adjacent words, especially in the baseline interval and at latencies of the late components of the rERP (<400 ms). Different variants of the rERP method have already been successfully applied to linguistic experiments in the visual (e.g., Hauk et al., 2006) and auditory domain (e.g., Brennan and Pylkkänen, 2012; Sassenhagen, 2018; Alday, 2019; Röhr et al., 2020; Ventura et al., 2020).

In the following study, we aim at exploring the form-function relation between anaphors and their antecedents as outlined above. Prior to any analysis steps, we chose to analyze the P300 (200–300 ms) and N400 time windows (300–500 ms), because of the sensitivity of the P300 to predictability in language comprehension in general (e.g., Roehm et al., 2007; Sassenhagen et al., 2014; Brilmayer et al., 2017; Alday and Kretzschmar, 2019) and because of the effects in Brilmayer et al. (2019), who recorded the data used for the present analysis. The N400 was chosen because of the ubiquity of N400 effects in discourse research and language research in general (cf. Kutas and Federmeier, 2011). Based on previous findings on the P300 and N400 component in language comprehension, we expect P300 and N400 amplitude to be sensitive to the prominence (i.e., predictability) of a referential form. In particular, we expect the P300 amplitude to increase along the prominence scale provided above (pronoun-pronoun chain > noun-pronoun chain > noun-noun chain > pronoun-noun chain). For the N400, we expect the most pronounced mismatch effect for the pronoun-noun chain relative to the noun-noun chain, because given the form-function mapping, the former combination is the least predicted (pronoun-noun chain > noun-noun chain); as far as pronoun anaphors are concerned, both antecedent types license pronominal coreference, and hence, no prediction error might arise, but alternatively, the pronoun-pronoun chain might represent the most ideal referential chain (noun-pronoun chain ≥ pronoun-pronoun chain).

EXPERIMENT

Materials and Methods

Participants

In the present study, the data of 35 participants were analyzed, all participants were monolingual native speakers of German (23 females; mean age: 25.0 years, range 20–34) with normal hearing and unimpaired vision was analyzed. The data of 25 participants (14 females; mean age: 24.4 years, range 20–29) stem from a study by Brilmayer et al. (2019); 10 additional participants were recorded in our own laboratory (nine females; mean age: 25.6 years, range 21–34). Participants received either course credit or monetary compensation for participation. The data of three participants had to be excluded from further analysis due to heavy artifact contamination.

Experimental Stimuli and Procedure

A German audio book version of The Little Prince by Antoine de Saint-Exupéry (recording by Will Quadflieg, chapters 1–15, excluding chapters 5, 6, and 14) served as experimental stimulus. The book contains non-dialog passages written from the perspective of a third person narrator who is also a protagonist in the story, and dialog passages in which the main protagonist, The Little Prince, interacts with a variety of characters. Dialog passages make up ~40.8% of the story (58.2% narrative passages). The rest of the text consists of free indirect discourse, indirect speech, and direct thought (~1%). The recording was segmented using automatic speech segmentation provided by the Munich Automatic Segmentation (MAUS) Web interface (Schiel, 1999; Kisler et al., 2016), combined with manual corrections.

For the present study, we restricted our analysis to personal pronouns and nouns that were encoded as grammatical subject with a noun or pronoun antecedent, including also pronouns in direct speech and all other types of discourse, resulting in

TABLE 1 | Distribution of referential types in the analyzed sample and the whole audio book recording (in parentheses).

	Noun antecedent	Pronoun antecedent
Noun anaphor	40 (114)	51 (81)
Pronoun anaphor	63 (88)	152 (201)

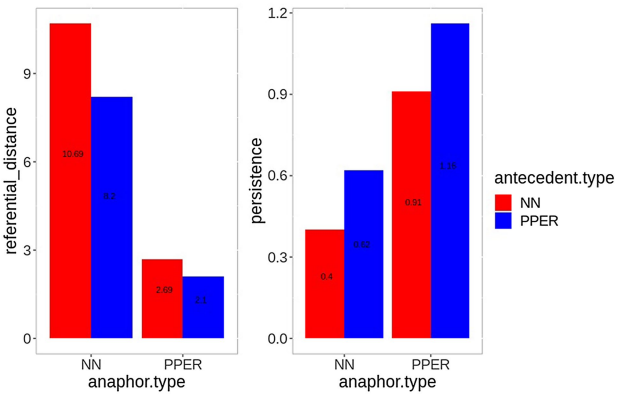

FIGURE 1 | Mean referential distance (left) and persistence (right) for all noun (red) and pronoun (blue) anaphors in the audio book recording by antecedent type.

a total of 215 pronouns [63 with noun antecedent; ich, "I"; du, "you.sg," er, sie, es, "he, she, it," wir, "we," Ihr, "Your" (hon.), Sie, and "you" (hon.)] and 91 nouns (40 with noun antecedent and 29 different nouns, e.g., Geograph, "geographer," Prinz, "prince," Blume, "flower," Planet, and "planet"). We chose these restrictions, in order to reduce the amount of uncontrolled variance in our data and to increase the reliability of our statistical analyses.

In our analysis, we compare ERPs following anaphoric expressions based on their referential type (noun/pronoun; anaphor type) and that of their antecedent (antecedent type). The resulting four referential chains have already been exemplified in (1–4). As argued above, we focus here on the P300 (200–300 ms) and N400 time window (300–500 ms). The distribution of anaphor types and antecedent types in the sample used for the rERP analysis is listed in **Table 1** (values in parentheses represent the occurrences in the entire audio book).

Table 2 summarizes the distribution of anaphors and antecedents with regard to the prominence lending features *identity of the referent* (prince, interlocutors, and other), *syntactic function* (subject, direct object, and other), and *definiteness* (definite, indefinite, and other). In summary, reference to either The Little Prince or interlocutors of The Little Prince made up ~72% of all referents in the current sample. Crucially, although the form (definite, indefinite, and proper name) of the noun anaphors varied, ~82% of them were definite. We find a very similar pattern for the antecedent expressions: About 75% of them were grammatical subjects (nouns: ~67%, pronouns: ~80%), and 90% were definite (nouns: ~74%, pronouns: >99%). That is, about four-fifth of antecedents were definite, grammatical subjects, although we did not formulate any selection criteria regarding their linguistic features. It seems that selecting only subject anaphors already filtered a great amount of linguistic variance among the antecedent expressions.

As mentioned in the introduction, Givón (1983) introduced referential distance and persistence as measures of textual cohesion and we thus annotated the present audio book recording for these quantitative prominence measures. Referential distance counts the number of clauses between a referential expression and the last mention of its antecedent. It ranges from 0 (same clause, e.g., reflexives, such as "John shaved himself.") to a maximum of 20, which is also the ceiling value assigned to newly introduced referents. Persistence determines the number of clauses in which a referent recurs in subsequent discourse. It can take any full number starting with 0 (no recurrence). In the following, we summarize these results, since referential distance and persistence entered the Principal Component Analysis detailed below (see also Torregrossa et al., 2018).

Figure 1 presents the results. As the left panel shows, noun anaphors are generally further away from their antecedent (10.69 and 8.2 clauses for noun and pronoun antecedents, respectively) than pronoun anaphors (2.69 and 2.1 clauses for noun and pronoun antecedents, respectively). In general, this pattern is consistent with the common observation that pronouns are usually closer to their antecedent than nouns (e.g., Givón, 1983; Gundel et al., 1993).

The difference between antecedent types for noun anaphors (+2.76 clauses for noun antecedents) is, thus, more pronounced than for pronoun anaphors (+0.59 clauses). With regard to persistence (right panel), we can observe the opposite pattern. Noun anaphors with noun antecedent have the lowest persistence value (0.4 clauses), followed by nouns with pronoun antecedent (0.62 clauses), pronouns with noun antecedent (0.91 clauses), and pronouns with pronoun antecedent, which have the highest persistence value (1.16). Interestingly, the prominence ranking resulting from the referential distance and persistence values (pronoun-pronoun chains > noun-pronoun chains > pronoun-noun chains > noun-noun chains) is not identical with the prominence ranking based on referential form, as the order of noun-noun and pronoun-noun anaphors is reversed. However, the results of the text analysis demonstrate nicely that discourse structural properties of referential expressions do not only depend on the referential form of the anaphor, but that it interacts with the referential form of the antecedent. It is therefore crucial to consider both in an analysis.

EEG Recording and Analysis

The scalp EEG was recorded using 32 Ag/AgCl electrodes attached according to the international 10–20 system using an elastic EEG cap (EasyCap, EasyCap GmbH, Herrsching, Germany). The EEG was recorded and digitized with a sampling rate of 500 Hz relative to right mastoid reference (BrainAmp DC, Brain Products, Gilching, Germany). Impedances were kept under 3 kΩ. The data were analyzed using a python3 implementation of MNE python (Gramfort et al., 2014) version 0.19. Before any further preprocessing procedures, experimental pauses were manually removed from the raw recordings. Afterward, the data were

TABLE 2 | Distribution of several prominence features of anaphors and antecedents in the analyzed sample (referent identity, definiteness, and syntactic).

Anaphor				Antecedent					
		Noun (91)	Pronoun (215)			N-N (40)	N-P (63)	P-N (51)	P-P (152)
Referent	Prince	25	85	Referent	Prince	9	25	16	60
	Interlocutors	39	68		Interlocutors	13	25	25	43
	Other	27	62		Other	18	13	10	49
Definiteness	Definite	75	213	Definiteness	Definite	28	48	50	151
	Indefinite	8	0		Indefinite	9	12	0	0
	Other	8	2		Other	3	3	1	1
				Syn. function	Subject	25	44	36	127
					Direct object	8	12	10	8
					Other	7	7	5	17

Note that syntactic function is not listed for anaphors, since via our selection criterion, all anaphors were syntactic subjects. N-N = noun antecedents of noun anaphors; N-P = noun antecedents of pronoun anaphors; P-N = pronoun antecedents of noun anaphors; and P-P = pronoun antecedents of pronoun anaphors.

re-referenced to linked mastoids. We then used independent component analysis (ICA) for artifact correction. For ICA, the EEG was filtered with a 1 Hz high-pass filter in order to approach stationarity and a 45 Hz low-pass to remove line noise. ICA was then computed with a decimation factor of 4. Afterward, artifact components (blinks and saccades) were selected and removed from the unfiltered EEG, to which the IC solution was applied. Instead of applying a baseline correction, we chose to filter the EEG with a 0.3 Hz high-pass and a 30 Hz low-pass filter (cf. Friederici et al., 2000; Wolff et al., 2008; Widmann et al., 2015; Maess et al., 2016a,b). For the calculation of the regression-based ERPs (rERP), we re-sampled the data to 250 Hz in order to reduce computational demands.

Principle Component Analysis

As mentioned above, naturalistic stimuli contain huge amounts of uncontrolled variance. Since the available data and thus statistical power are limited (the inclusion of more than three covariates leads to problems with overfitting), we had to decide which of the available covariates to include in the regression model. We thus used principal component analysis to determine which of the available covariates (mean f0-pitch, mean intensity, frequency class, word duration, referential distance, and persistence) explain the most variance in the present sample of the audio book version of The Little Prince. Of the six variables, we chose the three variables with the highest contribution and quality of representation in the first three dimensions (~60% of 72% of total variance): mean f0-pitch, duration, and referential distance. It is important to note here that referential distance was actually outranked by frequency class. The reason why we still chose referential distance over frequency class lies in its distribution: Frequency class almost perfectly divides the data into noun and pronoun anaphors. While 86.5% of pronouns have a frequency class at or below the mean frequency class of about 6.8 (mean: 4.1, classes 3: 40%, 4: 46.5%, and 8: 13.5%), 100% of noun anaphors lie above it (mean: 13.2, range: 7–21). With regard to referential distance, pronouns still have a lower mean than noun anaphors (2.3 vs. 7.8 sentences), yet they both cover the full range from 0 (antecedent in the same sentence) to 20 (20 sentences or more to antecedent, or newly introduced). In addition, duration and frequency class cover almost identical data points ($r = 0.8$). The inclusion of both in one model is thus of low explanatory value.

rERP Calculation

The rERP calculation was performed using the lm() function in R with amplitude scaled to the standard deviation scale as dependent variable and anaphor type and antecedent type as factors with interaction. Duration, referential distance, and mean f0-pitch were added as covariates without interactions. All factors were encoded using deviation coding. We calculated linear models by participant, channel, and sample (= 6526 models per participant). From each of these models, we extracted the fitted values for the interaction of anaphor and antecedent type for the second-order statistical analysis using the function effect() from the package effects (Fox and Weisberg, 2019), thereby disregarding the effects of the covariates. The resulting single-subject rERPs are comparable to traditional single-subject averages and can be used for further analysis in the same way.

Second-Order Statistical Analysis

The second-order statistical analysis was carried out using linear mixed-effect models as implemented in the lme4 package for R (Bates et al., 2014) with N400 (300–500 ms) amplitude as dependent variable. The model included fixed effects for antecedent type (noun/pronoun) and anaphor type (noun/pronoun), as well as two continuous topographic fixed effects based on two-dimensional electrode positions (saggitality/laterality). Contrasts were encoded using deviation coding, so that individual coefficients represent differences from the (grand) mean. Since all our factors have two levels, they are equidistant to the mean, which means that model coefficients can be directly interpreted as differences between conditions. The model was fitted using a backward approach, starting with maximally specified random effects until we arrived at a converging model (cf. Bates et al., 2015). The model included a by-participant

intercept and by-participant random slopes for each fixed factor without interactions. In the following, we will only discuss contrasts that are significant *via* the $|t| \geq 2$ criterion corresponding to traditional $p < 0.05$ (Baayen et al., 2008). To assess pairwise statistical significance, we estimated marginal means using the function emmeans() as implemented in the R library emmeans (Lenth et al., 2019).

Results

rERP

Figure 2 shows the beta coefficients of the critical predictors and their interaction by region-of-interest. Although the coefficients start with a large offset in the baseline interval, anaphor type and its interaction with antecedent type show a zero crossing (reversal of the sign), suggesting that they are strong predictors. Antecedent type has almost no effect (beta coefficient very close to zero), although a small positive effect is visible at ~400 + ms. Anaphor type (red) thereby shows a negative effect over central and posterior electrodes in the time window of the N400. The interaction shows the strongest effect between ~200 and 350 ms. Since the beta coefficient of an interaction is complicated to interpret, the fitted microvolt values are plotted in **Figure 3**. First, the difference between anaphor types becomes obvious: The rERP of pronoun anaphors is characterized by a positivity with posterio-central distribution, while the rERP of nouns is characterized by a posterio-central negativity. Yet, as discussed above, this difference can only be interpreted with caution, since nouns and pronouns differ critically in their phonetic properties and temporal extent (in the current recording, nouns are on average 2.9 times longer than pronouns: 430 ms vs. 150 ms). Therefore, we focus on the effects of antecedent type within anaphor types.

Within noun anaphors (**Figure 3**; red), chains with pronoun antecedent elicit more negative going amplitudes peaking at ~250 ms distributed over the entire scalp with a posterior maximum, and between ~400 and 750 ms (i.e., within the N400 time window) over central and posterior electrodes. Within pronoun anaphors (**Figure 3**; blue), chains with pronoun antecedent elicit more positive going amplitudes at left central and posterior electrodes between ~200 and 400 ms as compared to pronoun anaphors with noun antecedent. This difference is most pronounced at ~250 ms after word onset. In the following sections, we report the results of the time-window analysis.

Second-Order Statistical Analysis

P300

The results of the statistical analysis of the P300 time window are summarized in **Table 3**. We focus here on the effects involving the critical factors antecedent type and anaphor type with $|t| > = 2$. As follows from **Table 3**, there is a significant main effect of *anaphor type*, significant two-way interactions *saggitality*antecedent type*, *laterality*anaphor type*, *saggitality*anaphor type*, and *antecedent type*anaphor type*. In addition, the model includes a significant three-way interaction between *saggitality, antecedent type*, and *anaphor type*. We focus here on the significant contrasts that are not part of a higher-order contrast, i.e., *laterality*anaphor type* and *saggitality*antecedent type*anaphor type*. To assess statistical

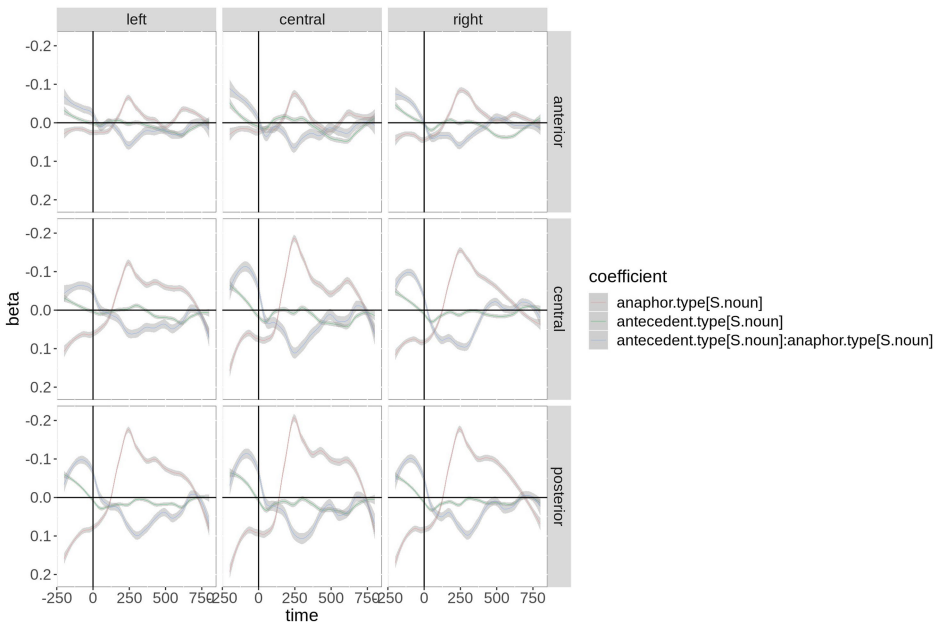

FIGURE 2 | Time course of the beta coefficients of the critical predictors and their interaction by region-of-interest (ROI). For plotting purposes, the continuous topographic variables were grouped into ROIs based on two-dimensional coordinates. Shaded areas represent 83% confidence intervals (an approximation to the traditional 0.05 level of significance for visualization purposes).

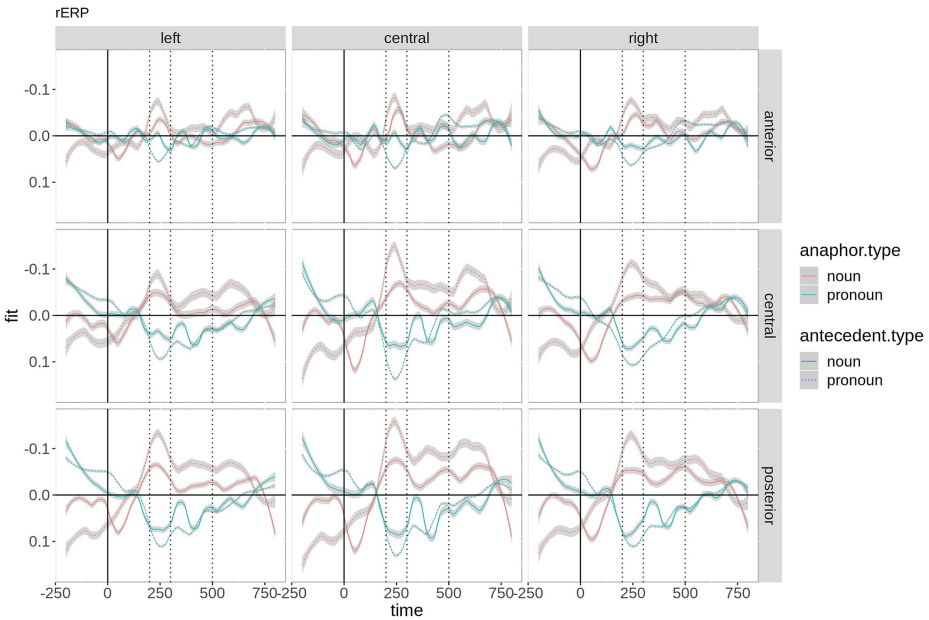

FIGURE 3 | Time course of the beta coefficients of the critical predictors and their interaction by ROI. For plotting purposes, the continuous topographic variables were grouped into ROIs based on two-dimensional coordinates. Shaded areas represent 83% confidence intervals (an approximation to the traditional 0.05 level of significance for visualization purposes).

TABLE 3 | Summary of the statistical model of the P300 time window.

Coefficient	*β*	*SE*	*t*
(Intercept)	−0.002	0.012	−1.79
Laterality	0.003	0.002	1.28
Saggitality	0.007	0.003	2.81
Antecedent type (noun)	−0.005	0.006	0.77
Anaphor type (noun)	−0.068	0.009	−7.16
Laterality:saggitality	0.002	0.005	0.33
Laterality:antecedent type (noun)	0.002	0.002	0.79
Saggitality:antecedent type (noun)	0.011	0.003	4.31
Laterality:anaphor type (noun)	−0.005	0.002	−2.80
Saggitality:anaphor type (noun)	0.049	0.003	−19.8
Antecedent type (noun):anaphor type (noun)	0.020	0.005	4.05
Laterality:saggitality:antecedent type (noun)	−0.002	0.005	−0.38
Laterality:saggitality:anaphor type (noun)	0.009	0.005	1.85
Laterality:antecedent type (noun):anaphor type (noun)	−0.000	0.002	−0.12
Saggitality:antecedent type (noun):anaphor type (noun)	0.006	0.003	2.24
Laterality:saggitality:antecedent type (noun):anaphor type (noun)	0.003	0.005	0.55

significance, we estimated marginal means using the function emmeans() as implemented in the R library emmeans (Lenth et al., 2019). For this purpose, we split the continuous topographic variables laterality and saggitality into three bins: a central bin (lateral and central midline) and two bins based on the mean of their positive (right/anterior) and negative values (left/posterior). According to this analysis, the contrast between noun and pronoun anaphors is significant at left (estimate = −0.125, t = −6.40, p < 0.001), central (estimate = −0.133, t = −6.87, p < 0.001), and right electrodes (estimate = −0.141, t = −7.21, p < 0.001).

The difference between the two anaphoric expressions is estimated to be largest over right hemispheric electrodes. **Figure 4** contains the fitted values of this interaction. Moreover, the contrast between noun and pronoun antecedents is significant following noun anaphors at central and anterior electrodes (estimate = 0.048, t = 2.90, p < 0.007, and estimate = 0.064, t = 2.90, p < 0.001, respectively), while the contrast is significant for pronoun anaphors at central (estimate = −0.029, t = −2.02, p = 0.05) and posterior electrodes (estimate = −0.035, t = 2.34, p = 0.03). The interaction is plotted in **Figure 5**.

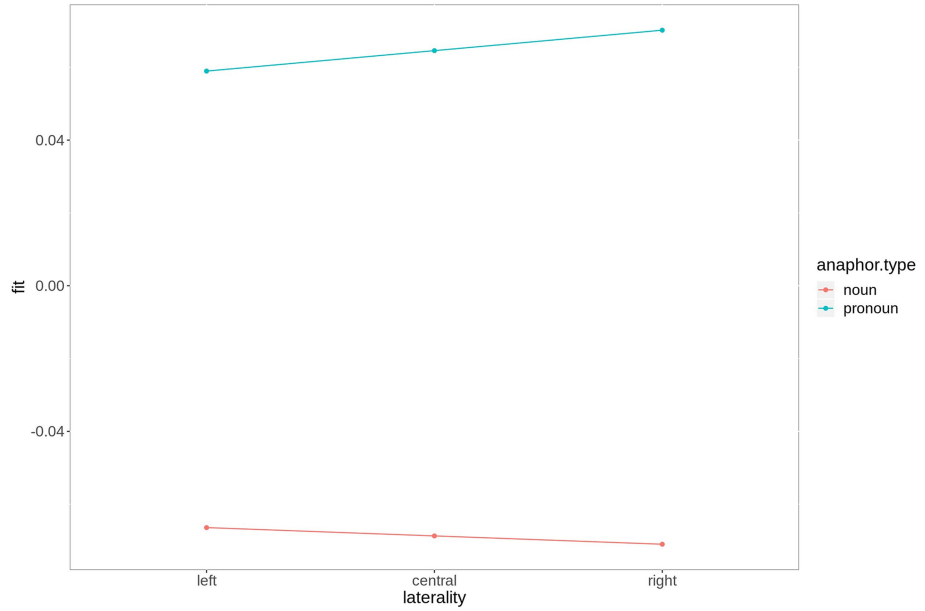

FIGURE 4 | Fitted values for the interaction laterality*anaphor type in the P300 time window. For plotting purposes, the continuous variable laterality was grouped into ROIs based on 2-dimensional coordinates.

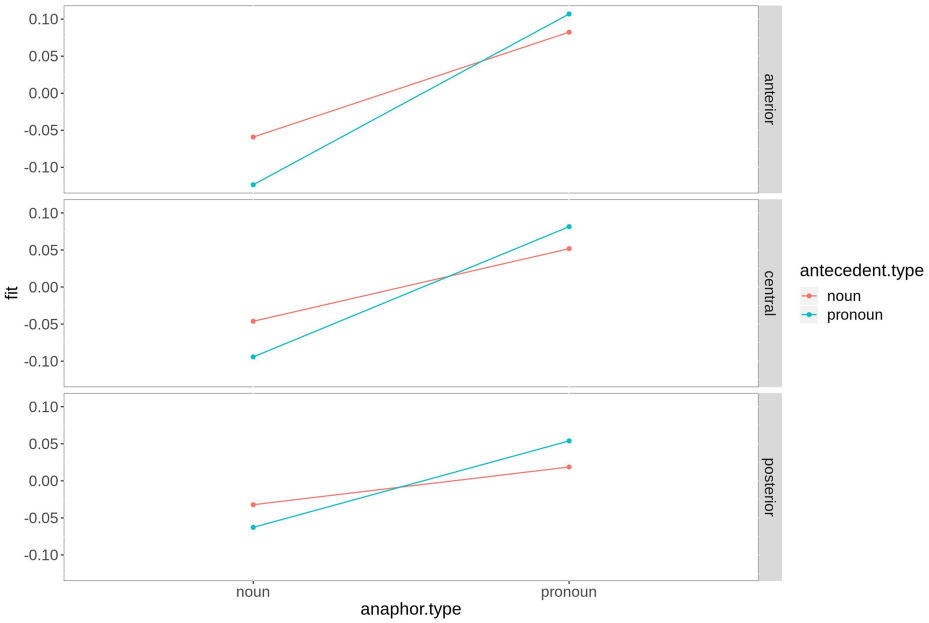

FIGURE 5 | Fitted values for the interaction saggitality*antecedent type*anaphor type in the P300 time window. For plotting purposes, the continuous variable saggitality was grouped into ROIs based on two-dimensional coordinates.

N400

The results of statistical analysis of the N400 time window are summarized in **Table 4**. It revealed a significant main effect of *anaphor type*, significant two-way interactions *laterality*anaphor type*, *saggitality*anaphor type*, *laterality *anaphor type*, *saggitality*anaphor type* and *anaphor type *antecedent type*. In addition, the model predicts a significant three-way interaction *saggitality*anaphor type*antecedent type*. Similar to the analysis of the P300 time window, we resolved the highest-order interactions using *emmeans*(), i.e., the interactions *laterality*anaphor type* and *saggitality*antecedent type*anaphor type*. The analysis revealed a significant effect

TABLE 4 | Summary of the statistical model of the N400 time window.

Coefficient	*β*	*SE*	*t*
(Intercept)	−0.002	0.006	−0.28
Laterality	−0.008	0.002	−4.83
Saggitality	0.000	0.002	0.24
Antecedent type (noun)	0.005	0.005	1.11
Anaphor type (noun)	−0.03	0.009	−3.53
Laterality:saggitality	−0.000	0.003	−0.19
Laterality:antecedent type (noun)	−0.000	0.002	−0.29
Saggitality:antecedent type (noun)	−0.000	0.002	−0.34
Laterality:anaphor type (noun)	0.007	0.002	−3.98
Saggitality:anaphor type (noun)	0.042	0.002	−19.8
Antecedent type (noun):anaphor type (noun)	0.01	0.015	2.04
Laterality:saggitality:antecedent type (noun)	−0.005	0.004	−0.12
Laterality:saggitality:anaphor type (noun)	0.005	0.004	−1.38
Laterality:antecedent type (noun):anaphor type (noun)	−0.003	0.016	−1.71
Saggitality:antecedent type (noun):anaphor type (noun)	0.08	0.002	3.97
Laterality:saggitality:antecedent type (noun):anaphor type (noun)	−0.002	0.003	−0.55

of *anaphor type* at left (estimate = −0.048, $t = -2.75$, $p = 0.009$), central (estimate = −0.058, $t = -3.30$, $p = 0.002$), and right electrodes (estimate = −0.067, $t = -3.81$, $p < 0.001$).

As depicted in **Figure 6**, the effect is most pronounced over right hemispheric electrodes. Regarding the three-way interaction *saggitality*antecedent type*anaphor type*, the analysis revealed a significant effect of *antecedent type* at anterior electrodes for noun anaphors only (estimate = 0.039, $t = 2.32$, $p = 0.03$), while all p-values following pronoun anaphors exceed 0.1. The interaction is plotted in **Figure 7**.

Summary

In the present study, we compared ERP responses to anaphoric nouns or anaphoric pronouns with either a pronoun or a noun antecedent in the P300 (200–300 ms) and N400 time window (300–500 ms) using an audio book version of *The Little Prince* as experimental stimulus. The regression-based ERPs (rERP) reveal large differences in the morphology of the ERPs following noun anaphors and pronoun anaphors. While the rERP of the former is characterized by a large negative potential starting at ~200 ms after word onset, the rERP of pronoun anaphors is characterized by a positivity from ~200 ms onward. Moreover, based on the topography of the rERPs, an earlier (200–300 ms), widespread component with posterior maximum that is negative for noun anaphors and positive for pronoun anaphors can be separated from a later, sustained component only over central and posterior electrodes. Again, this component shows a negative going polarity for noun anaphors and a positive polarity for pronoun anaphors. The results of the statistical analysis in the early time window reveal a P300 amplitude gradient that follows the prominence ranking formulated above. That is, for pronouns and nouns, we found the expected gradient, both numerically, and statistically (pronoun-pronoun chain > noun-pronoun chain > noun-noun chain > pronoun-noun chain). In the N400 time window, noun anaphors elicited larger N400 amplitudes when coreferent with a pronoun than with a noun-pronoun antecedent. The N400 was not sensitive for the antecedent type of pronoun anaphors.

In summary, pronoun anaphors elicit the most positive going P300 and N400 amplitudes, yet only the amplitude of the P300 (200–300 ms) is sensitive to the type of the antecedent expression. Following noun anaphors, we also found a significant gradient that follows the prominence ranking formulated above, in both, the P300 and N400 time window.

DISCUSSION

The present ERP study tested the relationship of the referential form of antecedents and anaphors in referential chains and their influence on the P300 and N400 ERP components in auditory language comprehension using stimuli from a naturalistic audio book. By contrasting noun and pronoun anaphors with noun or pronoun antecedents, we hypothesized that the antecedent form is used as a predictive cue for the form of the anaphor. The results of our study are in favor of this assumption, as they reveal a significant influence of the form of the antecedent expression on the P300 and N400 amplitude following an anaphor. Most interestingly, the effects depend on the referential form of the anaphoric expression, pointing to an interaction of prediction (forward-looking function of the antecedent) and form-to-function mapping (backward-looking function) of referential expressions in the establishment of referential relations. In the following, we argue that this interaction can be explained from a predictive coding perspective on discourse comprehension.

P300

First, let us consider the P300 time window. Recall, that, in line with the literature (e.g., Prince, 1981; Givón, 1983; Ariel, 1990; Gundel et al., 1993; von Heusinger and Schumacher, 2019), we assumed that nouns are preferably used to refer to non-prominent discourse referents, as compared to personal

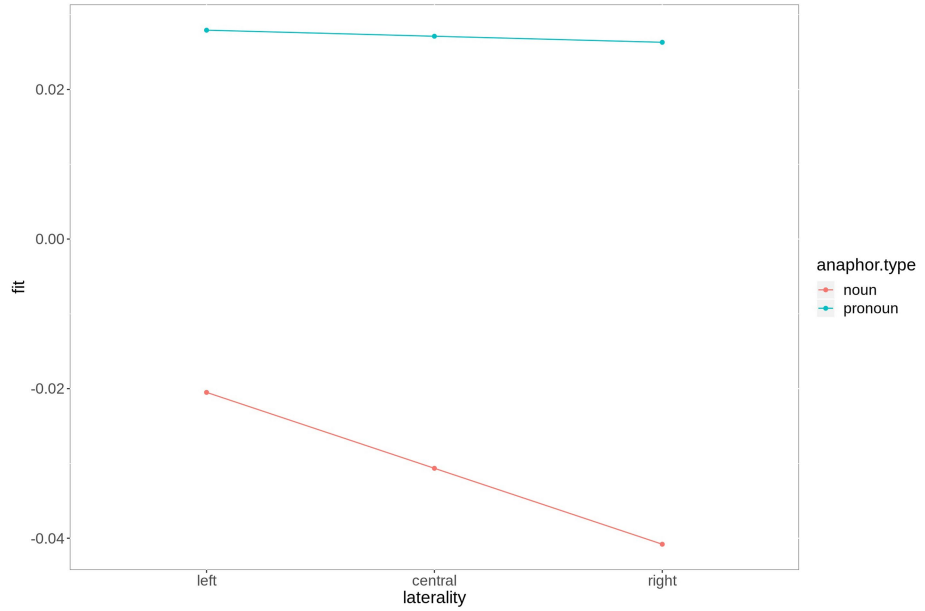

FIGURE 6 | Fitted values for the interaction laterality*anaphor type in the N400 time window. For plotting purposes, the continuous variable laterality was grouped into ROIs based on two-dimensional coordinates.

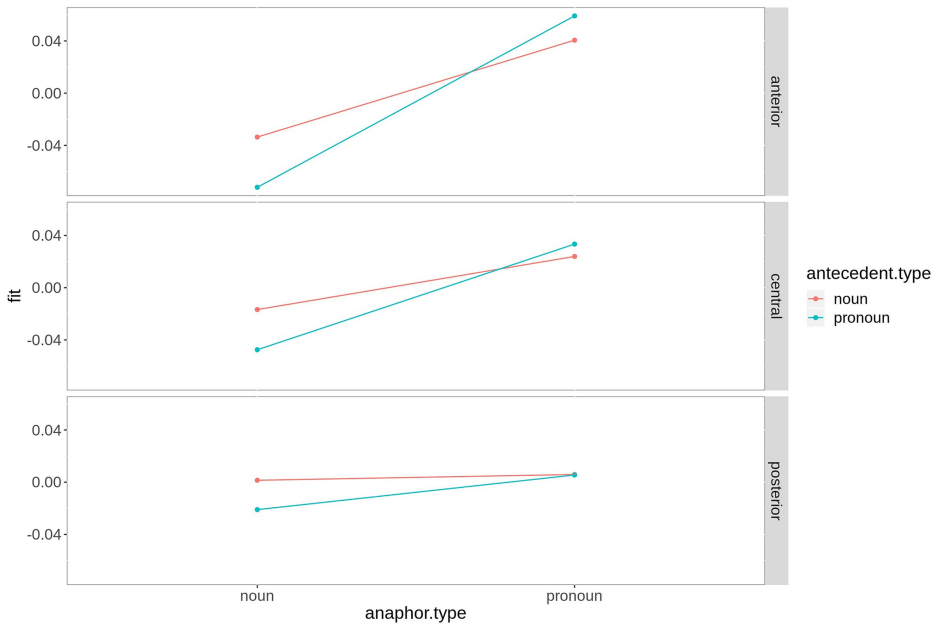

FIGURE 7 | Fitted values for the interaction saggitality*antecedent type*anaphor type in the N400 time window. For plotting purposes, the continuous variable saggitality was grouped into ROIs based on two-dimensional coordinates.

pronouns, which are used to refer to prominent referents in the majority of cases. Crucially, this should result in corresponding prediction match responses (enhanced P300). Following the assumption, noun anaphors with pronoun antecedent exhibit unusual (i.e., unpredicted) referential chains: A pronoun marks a referent as prominent and is used as a predictive cue for the referential form of subsequent mention. Noun anaphors, however, usually refer to non-prominent referents. A pronoun-noun chain, thus, constitutes a mismatch between the predicted form of the anaphor based on the antecedent expression and the preferred antecedent expression based on the form of the anaphor. In other words, the prediction derived from the

forward-looking function of the antecedent (prominent referent > pronoun anaphor preferred) and the mapping to a referent derived from the backward-looking function of the anaphor (non-prominent referent > no pronoun antecedent preferred) contradict each other. We argue that this mismatch is visible in the attenuated P300 amplitudes following anaphors in pronoun-noun chains, reflecting the absence of a highly predictable anaphoric continuation. For noun-noun chains, by contrast (forward-looking) prediction of the antecedent (non-prominent > no pronoun anaphor preferred) and (backward-looking) form-to-function mapping of the anaphor (non-prominent > no pronoun antecedent preferred) converge, mirroring a prediction match response, as visible in amplified P300 amplitudes.

Moving on to pronoun anaphors, we found that the P300 is sensitive to the form of the antecedent expression, with higher P300 amplitudes following pronouns with pronoun antecedent, as compared to pronouns with noun antecedents. This is consistent with our hypothesis that a pronoun antecedent clearly marks a referent as prominent, which makes it a predicted (or preferred) continuous referent that is likely to be rementioned by means of a personal pronoun. The prominence information conveyed by the antecedent and the referential-form prediction derived from it are fully congruent with the prominence information of the anaphor and its preferred antecedent. That is (forward-looking) prediction of the antecedent is satisfied when the anaphor is encountered. Hence, prediction match is achieved, as visible in an increase in P300 amplitude as compared to the less prominent referents with noun antecedent. Following Alday and Kretzschmar (2019), if we wanted to provide a cognitive explanation for the P300 effect, we might say that the P300 reflects the immediate categorization of pronoun anaphors with pronoun antecedents, in the sense that they can be directly linked to a referent in the discourse model without the need for further evidence (e.g., by subsequent context). In other words, as mentioned in the introduction, we argue that pronominal reference to prominent referents is predicted to the extent that the referential relation is anticipated before the anaphoric pronoun is actually detected. With noun antecedents, this linking is more difficult, or, differently speaking, less predicted; hence, no prediction match response arises and the P300 is reduced. We attribute this difficulty to differences in prominence assigned to referents by the referential form of the antecedent expression, with noun antecedents being less prominent than pronoun antecedents.

N400

In the N400 window, only noun anaphors show a graded N400 effect. We found that in noun-noun chains, the N400 following the anaphor was significantly reduced. We take this as evidence for the preference of nouns to corefer with a noun antecedent rather than a pronoun. By contrast, a pronoun-noun chain constitutes an exception with regard to discourse structure: An already prominent referent (realized by a pronoun) is referred to by a referential expression indicating a low level of prominence (noun). Form-to-function mapping of the anaphor (no pronoun antecedent) and the form of the antecedent (pronoun) thus contradict each other, hence the increase in N400 amplitude as a measure of a prediction error. Compared to this, a noun-noun chain is predictable, since nouns can easily be used to refer to a noun antecedent. In fact, this is quite common, for instance in referential chains consisting of an indefinite antecedent and a definite anaphor ("A man entered the room and looked around. The man then walked straight to the counter, when …").

Following pronoun anaphors, we did not find a significant influence of antecedent type on the N400 amplitude, supporting the idea that pronoun anaphors are less dependent on the form of their antecedent. Overall, the patterns for the two time windows are thus distinct, supporting a functional dissociation between processing predicted linguistic content (leading to categorization) and encountering unpredicted linguistic content leading to prediction error. This suggests that the linguistic evidence needed for the establishment of reference might not differ between the two types of pronouns, which is reflected in the absence of an N400 effect, yet the difference in prominence might result in difficulties with respect to the categorization process.

Comparison With Previous Experiments

The present results are compatible with previous experiments in so far as the literature on event-related potentials during referential processing consistently reports increased N400 amplitudes related to unpredicted referential relations based on prominence information (Swaab et al., 2004; Nieuwland and van Berkum, 2006; Camblin et al., 2007; Ledoux et al., 2007; Schumacher and Baumann, 2010; Hung and Schumacher, 2012; Schumacher and Hung, 2012; Wang and Schumacher, 2013; Almor et al., 2017). With the present analysis explicitly contrasting the referential form of anaphoric expressions and of their antecedents, we were able to show that prominence information based on referential form is already relevant for processing between 200 and 300 milliseconds after anaphor onset, and thus, earlier than the N400 time window usually considered crucial for referential processing. This finding is highly compatible with the results of Brilmayer et al. (2019) who provide a different analysis of the present data set. They compared pronouns of the first, second, and third person singular and found a significant P300 gradient (1 > 2 > 3) in the same early time window. As they argue, first person referents are always prominent for a variety of reasons (cf. Comrie, 1989; Dahl, 2008; Frith and Frith, 2010). These results thus corroborate the present finding that linguistic prominence information is already important at comparably early time points during the processing of referential expressions.

Crucially, the present results suggest that not only current linguistic input is reflected in this early component, but also the interaction of current linguistic input with information about the antecedent in memory. That is, stimulus-driven bottom-up information is already influenced by previous context as early as 200 milliseconds after stimulus onset. This strongly supports a predictive coding account to language-related ERPs as argued by (Bornkessel-Schlesewsky and Schlesewsky, 2019; see Bornkessel-Schlesewsky and Schumacher, 2016, for a

discussion of a possible predictive coding framework for discourse comprehension). In predictive coding, top-down information from higher (conceptual) processing levels constantly influences the way in which information is processed at lower (perceptual) levels. Thus, one and the same stimulus (e.g., a personal pronoun) is processed differently based on its own prominence information and prominence information conveyed by referential forms in previous context and long-term experience. Clearly, the future research must consider such early time windows during referential comprehension, given their relevance in referential processing suggested by the present and previous studies.

Overall, our findings provide empirical support for the prominence approach to reference in discourse as proposed by von Heusinger and Schumacher (2019): Referential expressions differ in their form-to-function mapping (related to singling out, definition 1), and in the discourse predictions derived from them (related to structural attraction, definition 3). The interplay of these two functions (forward-looking) prediction and (backward-looking) form-to-function mapping, is reflected in the P300/N400 patterns following anaphoric expressions.

Conclusion

In the present study, we showed that the P300 and N400 component are sensitive to the interaction of prominence information conveyed by an antecedent and an anaphoric expression. We showed that as early as 200 milliseconds after the onset of the anaphoric expression, the referential type of an antecedent has an influence on the ERP of an anaphor. Crucially, the effects were reversed depending on the anaphoric form. While nouns showed a graded negativity in the P300 time window (pronoun-noun chain > noun-noun chain), pronouns showed a graded positivity (pronoun-pronoun chain > noun-pronoun chain). We attribute these effects to the interaction of predictions derived from the antecedent and preferences in the form-to-function mapping of anaphors. The N400, by contrast, was only sensitive to discourse-pragmatic regularities following noun anaphors, suggesting differences in the mapping process between noun and pronoun anaphors.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committee of the German Linguistic Society. The participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

IB is responsible for the analysis of the audio book, the preprocessing and analysis of the EEG data, as well as for statistical analysis. PBS and IB were equally involved in hypothesis generation, interpretation of the data and the writing process. All authors contributed to the article and approved the submitted version.

FUNDING

This research was funded by the German Research Foundation (DFG) as part of the Collaborative Research Center 1252 "Prominence in Language" – Project-ID 281511265 – in the projects C07 "Forward and backward functions of discourse anaphora" and B07 "Agentivity as a key to prominence: Experimental approaches to argument alternations in German" at the University of Cologne, Department of German Language and Literature I, Linguistics.

ACKNOWLEDGMENTS

We would like to thank Brita Rietdorf for her assistance in data recording and Alexandra Werner and Hilde Penner for the many hours they spent annotating the audio book.

REFERENCES

Alday, P. M. (2019). How much baseline correction do we need in ERP research? Extended GLM model can replace baseline correction while lifting its limits. *Psychophysiology* 56:e13451. doi: 10.1111/psyp.13451

Alday, P. M., and Kretzschmar, F. (2019). Speed-accuracy tradeoffs in brain and behavior: testing the independence of P300 and N400 related processes in behavioral responses to sentence categorization. *Front. Hum. Neurosci.* 13:285. doi: 10.3389/fnhum.2019.00285

Alday, P. M., Schlesewsky, M., and Bornkessel-Schlesewsky, I. (2017). Electrophysiology reveals the neural dynamics of naturalistic auditory language processing: event-related potentials reflect continuous model updates. *Eneuro* 4:ENEURO.0311-16.2017. doi: 10.1523/ENEURO.0311-16.2017

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

Almor, A., and Eimas, P. D. (2008). Focus and noun phrase anaphors in spoken language comprehension. *Lang. Cogn. Process.* 23, 201–225. doi: 10.1080/01690960701330936

Almor, A., Nair, V. A., Boiteau, T. W., and Vendemia, J. M. (2017). The N400 in processing repeated name and pronoun anaphors in sentences and discourse. *Brain Lang.* 173, 52–66. doi: 10.1016/j.bandl.2017.06.003

Ariel, M. (1990). *Accessing Noun-Phrase Antecedents. Vol. 96*. Londres: Routlege, Linguistics, 113–118.

Arnold, J. E. (1998). *Reference Form and Discourse Patterns*. dissertation. Stanford University.

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., Kliegl, R., Vasishth, S., and Baayen, H. (2015). Parsimonious mixed models. ArXiv, 1506.04967 [Preprint].

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2014). Fitting linear mixed-effects models using lme4. ArXiv, 1406.5823 [Preprint].

Bhattasali, S., Fabre, M., Luh, W. M., Al Saied, H., Constant, M., Pallier, C., et al. (2019). Localising memory retrieval and syntactic composition: an fMRI study of naturalistic language comprehension. *Lang. Cogn. Neurosci.* 34, 491–510. doi: 10.1080/23273798.2018.1518533

Bornkessel-Schlesewsky, I., and Schlesewsky, M. (2019). Toward a neurobiologically plausible model of language-related, negative event-related potentials. *Front. Psychol.* 10:298. doi: 10.3389/fpsyg.2019.00298

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), 581–598.

Brennan, J. R., Lajiness-O'Neill, R., Bowyer, S., Kovelman, I., and Hale, J. T. (2019). Predictive sentence comprehension during story-listening in autism spectrum disorder. *Lang. Cogn. Neurosci.* 34, 428–439. doi: 10.1080/23273798.2018.1560483

Brennan, J., and Pylkkänen, L. (2012). The time-course and spatial distribution of brain activity associated with sentence processing. *Neuroimage* 60, 1139–1148. doi: 10.1016/j.neuroimage.2012.01.030

Brennan, J. R., Stabler, E. P., Van Wagenen, S. E., Luh, W.-M., and Hale, J. T. (2016). Abstract linguistic structure correlates with temporal activity during naturalistic comprehension. *Brain Lang.* 157-158, 81–94. doi: 10.1016/j.bandl.2016.04.008

Brilmayer, I., Sassenhagen, J., Bornkessel-Schlesewsky, I., and Schlesewsky, M. (2017). Domain-general neural correlates of dependency formation: using complex tones to simulate language. *Cortex* 93, 50–67. doi: 10.1016/j.cortex.2017.05.003

Brilmayer, I., Werner, A., Primus, B., Bornkessel-Schlesewsky, I., and Schlesewsky, M. (2019). The exceptional nature of the first person in natural story processing and the transfer of egocentricity. *Lang. Cogn. Neurosci.* 34, 411–427. doi: 10.1080/23273798.2018.1542501

Brocher, A., and von Heusinger, K. (2018). A dual-process activation model: processing definiteness and information status. *Glossa* 3, 1–34. doi: 10.5334/gjgl.457

Brouwer, H., Fitz, H., and Hoeks, J. (2012). Getting real about semantic illusions: rethinking the functional role of the P600 in language comprehension. *Brain Res.* 1446, 127–143. doi: 10.1016/j.brainres.2012.01.055

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

Burkhardt, P., and Roehm, D. (2007). Differential effects of saliency: an event-related brain potential study. *Neurosci. Lett.* 413, 115–120. doi: 10.1016/j.neulet.2006.11.038

Camblin, C. C., Ledoux, K., Boudewyn, M., Gordon, P. C., and Swaab, T. Y. (2007). Processing new and repeated names: effects of coreference on repetition priming with speech and fast RSVP. *Brain Res.* 1146, 172–184. doi: 10.1016/j.brainres.2006.07.033

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

Chiarcos, C., Claus, B., and Grabski, M. (eds.) (2011). *Salience: Multidisciplinary Perspectives on Its Function in Discourse. Vol. 227*. New York: Walter de Gruyter.

Comrie, B. (1989). *Language Universals and Linguistic Typology: Syntax and Morphology*. Oxford: Blackwell.

Coulson, S. (1998). ERPs and domain specificity: beating a straw horse. *Lang. Cogn. Process.* 13, 653–672. doi: 10.1080/016909698386410

Coulson, J., King, J., and Kutas, M. (1998). Expect the unexpected: event-related brain response to morphosyntactic violations. *Lang. Cogn. Process.* 13, 21–58. doi: 10.1080/016909698386582

Dahl, Ö. (2008). Animacy and egophoricity: grammar, ontology and phylogeny. *Lingua* 118, 141–150. doi: 10.1016/j.lingua.2007.02.008

de Saint-Exupéry, A. (2012). *Der Kleine Prinz*. Düsseldorf: Karl Rauch Verlag GmbH and Co. KG.

Ehinger, B. V., and Dimigen, O. (2019). Unfold: an integrated toolbox for overlap correction, non-linear modeling, and regression-based EEG analysis. *PeerJ* 7:e7838. doi: 10.7717/peerj.7838

Falk, S. (2014). On the notion of salience in spoken discourse-prominence cues shaping discourse structure and comprehension. *TIPA. Travaux interdisciplinaires sur la parole et le langage* 19:1303. doi: 10.4000/tipa.1303

Fox, J., and Weisberg, S. (2019). *An R Companion to Applied Regression. 3rd Edn*. Thousand Oaks, CA: Sage.

Friederici, A. D., Wang, Y., Herrmann, C. S., Maess, B., and Oertel, U. (2000). Localization of early syntactic processes in frontal and temporal cortical areas: a magnetoencephalographic study. *Hum. Brain Mapp.* 11, 1–11. doi: 10.1002/1097-0193(200009)11:1<1::AID-HBM10>3.0.CO;2-B

Friston, K. (2005). A theory of cortical responses. *Philos. Trans. R. Soc. B Biol. Sci.* 360, 815–836. doi: 10.1098/rstb.2005.1622

Frith, U., and Frith, C. (2010). The social brain: allowing humans to boldly go where no otherspecies has been. *Philos. trans. R. Soc. B Biol. sci.* 365, 165–176. doi: 10.1098/rstb.2009.0160

Fuchs, M., and Schumacher, P. B. (2020). "Referential shift potential of demonstrative pronouns – evidence from text continuation," in *Demonstratives in Discourse*. eds. Å. Næss, A. Margetts and Y. Treis (Berlin: Language Science Press), 185–213.

Gelormini-Lezama, C., and Almor, A. (2011). Repeated names, overt pronouns, and null pronouns in Spanish. *Lang. Cogn. Process.* 26, 437–454. doi: 10.1080/01690965.2010.495234

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

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

Gordon, P. C., and Hendrick, R. (1997). Intuitive knowledge of linguistic co-reference. *Cognition* 62, 325–370. doi: 10.1016/S0010-0277(96)00788-3

Gordon, P. C., Hendrick, R., Ledoux, K., and Yang, C. L. (1999). Processing of reference and the structure of language: an analysis of complex noun phrases. *Lang. Cogn. Process.* 14, 353–379. doi: 10.1080/016909699386266

Gramfort, A., Luessi, M., Larson, E., Engemann, D., Strohmeier, D., Brodbeck, C., et al. (2014). MNE software for processing MEG and EEG data. *Neuroimage* 86, 446–460. doi: 10.1016/j.neuroimage.2013.10.027

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

Grosz, B., and Sidner, C. L. (1986). Attention, intentions, and the structure of discourse. *Comput. Linguis.* 12, 175–204.

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

Hammer, A., Jansma, B. M., Lamers, M., and Münte, T. F. (2005). Pronominal reference in sentences about persons or things: an electrophysiological approach. *J. Cogn. Neurosci.* 17, 227–239. doi: 10.1162/0898929053124947

Hammer, A., Jansma, B. M., Lamers, M., and Münte, T. F. (2008). Interplay of meaning, syntax and working memory during pronoun resolution investigated by ERPs. *Brain Res.* 1230, 177–191. doi: 10.1016/j.brainres.2008.07.004

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

Himmelmann, N. P., and Primus, B. (2015). "Prominence beyond prosody: a first approximation," in *pS-prominenceS: Prominences in Linguistics. Proceedings of the International Conference*; December 2013 (Viterbo: Disucom Press), 38–58.

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

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

Kisler, T., Reichel, U., Schiel, F., Draxler, C., Jackl, B., and Pörner, N. (2016). "BAS speech science web services - an update of current developments," in *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)* (Paris, France: European Language Resources Association (ELRA)).

Kutas, M., and Federmeier, K. D. (2011). Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). *Annu. Rev. Psychol.* 62, 621–647. doi: 10.1146/annurev.psych.093008.131123

Lambrecht, K. (1996). *Information Structure and Sentence Form: Topic, Focus, and the Mental Representations of Discourse Referents. Vol. 71*. New York: Cambridge University Press.

Lamers, M. J., Jansma, B. M., Hammer, A., and Münte, T. F. (2006). Neural correlates of semantic and syntactic processes in the comprehension of case marked pronouns: evidence from German and Dutch. *BMC Neurosci.* 7:23. doi: 10.1186/1471-2202-7-23

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

Lenth, R., Singmann, H., Love, J., Buerkner, P., and Herve, M. (2019). Emmeans: estimated marginal means, aka least-squares means (Version 1.3.4).

Maess, B., Schröger, E., and Widmann, A. (2016a). High-pass filters and baseline correction in M/EEG analysis. Commentary on: "how inappropriate high-pass filters can produce artifacts and incorrect conclusions in ERP studies of language and cognition". *J. Neurosci. Methods* 266, 164–165. doi: 10.1016/j.jneumeth.2015.12.003

Maess, B., Schröger, E., and Widmann, A. (2016b). High-pass filters and baseline correction in M/EEG analysis-continued discussion. *J. Neurosci. Methods* 266, 171–172. doi: 10.1016/j.jneumeth.2016.01.016

Mak, M., and Willems, R. M. (2018). Mental simulation during literary reading: individual differences revealed with eye-tracking. *Lang. Cogn. Neurosci.* 34, 511–535. doi: 10.1080/23273798.2018.1552007

Nieuwland, M. S., and van Berkum, J. J. (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

Prince, E. F. (1981). "Toward a taxonomy of given-new information," in *Radical Pragmatics*. ed. P. Cole (New York: Academic Press), 223–255.

Roehm, D., Bornkessel-Schlesewsky, I., Rösler, F., and Schlesewsky, M. (2007). To predict or not to predict: influences of task and strategy on the processing of semantic relations. *J. Cogn. Neurosci.* 19, 1259–1274. doi: 10.1162/jocn.2007.19.8.1259

Röhr, C. T., Brilmayer, I., Baumann, S., Grice, M., and Schumacher, P. B. (2020). Signal-driven and expectation-driven processing of accent types. *Lang. Cogn. Neurosci.* 36, 33–59. doi: 10.1080/23273798.2020.1779324

Sassenhagen, J. (2018). How to analyse electrophysiological responses to naturalistic language with time-resolved multiple regression. *Lang. Cogn. Neurosci.* 34, 474–490. doi: 10.1080/23273798.2018.1502458

Sassenhagen, J., and Alday, P. M. (2016). A common misapplication of statistical inference: nuisance control with null-hypothesis significance tests. *Brain Lang.* 162, 42–45. doi: 10.1016/j.bandl.2016.08.001

Sassenhagen, J., and Bornkessel-Schlesewsky, I. (2015). The P600 as a correlate of ventral attention network reorientation. *Cortex* 66, A3–A20. doi: 10.1016/j.cortex.2014.12.019

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

Schiel, F. (1999). "Automatic phonetic transcription of non-prompted speech," in *Proceedings of the XIVth International Congress of Phonetic Sciences (ICPhS 99)*; August 1–7, 1999; ed. J. J. Ohala (San Francisco), 607–610.

Schilbach, L., Timmermans, B., Reddy, V., Costall, A., Bente, G., Schlicht, T., et al. (2013). Toward a second-person neuroscience. *Behav. Brain Sci.* 36, 393–414. doi: 10.1017/S0140525X12000660

Schilling, A., Tomasello, R., Henningsen-Schomers, M. R., Zankl, A., Surendra, K., Haller, M., et al. (2021). Analysis of continuous neuronal activity evoked by continuous speech with computational corpus linguistics methods. *Lang. Cogn. Neurosci.* 36, 167–186. doi: 10.1080/23273798.2020.1803375

Schmitt, B. M., Lamers, M., and Münte, T. F. (2002). Electrophysiological estimates of biological and syntactic gender violation during pronoun processing. *Cogn. Brain Res.* 14, 333–346. doi: 10.1016/S0926-6410(02)00136-2

Schumacher, P. B. (2009). "Definiteness marking shows late effects during discourse processing: evidence from ERPs," in *Anaphora Processing and Applications. DAARC 2009. Lecture Notes in Computer Science. Vol. 5847*. eds. S. Lalitha Devi, A. Branco and R. Mitkov (Berlin, Heidelberg: Springer), 91–106.

Schumacher, P. B., Backhaus, J., and Dangl, M. (2015). Backward-and forward-looking potential of anaphors. *Front. Psychol.* 6:1746. doi: 10.3389/fpsyg.2015.01746

Schumacher, P. B., and Baumann, S. (2010). Pitch accent type affects the N400 during referential processing. *NeuroReport* 21, 618–622. doi: 10.1097/WNR.0b013e328339874a

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

Smith, N. J., and Kutas, M. (2015a). Regression-based estimation of ERP waveforms: I. The rERP framework. *Psychophysiology* 52, 157–168. doi: 10.1111/psyp.12317

Smith, N. J., and Kutas, M. (2015b). Regression-based estimation of ERP waveforms: II. Nonlinear effects, overlap correction, and practical considerations. *Psychophysiology* 52, 169–181. doi: 10.1111/psyp.12320

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

Streb, J., Rösler, F., and Hennighausen, E. (1999). Event-related responses to pronoun and proper name anaphors in parallel and nonparallel discourse structures. *Brain Lang.* 70, 273–286. doi: 10.1006/brln.1999.2177

Swaab, T. Y., Camblin, C. C., and Gordon, P. C. (2004). Electrophysiological evidence for reversed lexical repetition effects in language processing. *J. Cogn. Neurosci.* 16, 715–726. doi: 10.1162/089892904970744

Torregrossa, J., Bongartz, C., and Tsimpli, I. M. (2019). Bilingual reference production: a cognitive-computational account. *Linguist. Approaches Biling.* 9, 569–599. doi: 10.1075/lab.17026.tor

van Berkum, J. J., Hagoort, P., and Brown, C. M. (1999). Semantic integration in sentences and discourse: evidence from the N400. *J. Cogn. Neurosci.* 11, 657–671. doi: 10.1162/089892999563724

van Berkum, J. J., Koornneef, A. W., Otten, 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

Ventura, C., Grice, M., Savino, M., Kolev, D., Brilmayer, I., and Schumacher, P. B. (2020). Attention allocation in a language with post-focal prominences. *Neuroreport* 31, 624–628. doi: 10.1097/WNR.0000000000001453

Verleger, R., Jaśkowski, P., and Wascher, E. (2005). Evidence for an integrative role of P3b in linking reaction to perception. *J. Psychophysiol.* 19, 165–181. doi: 10.1027/0269-8803.19.3.165

von Heusinger, K., and Schumacher, P. B. (2019). Discourse prominence: definition and application. *J. Pragmat.* 154, 117–127. doi: 10.1016/j.pragma.2019.07.025

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

Widmann, A., Schröger, E., and Maess, B. (2015). Digital filter design for electrophysiological data – a practical approach. *J. Neurosci. Methods* 250, 34–46. doi: 10.1016/j.jneumeth.2014.08.002

Willems, R. M. (ed.) (2015). *Cognitive Neuroscience of Natural Language Use*. Cambridge: Cambridge University Press.

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

Conflict of Interest: 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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Brilmayer and Schumacher. 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) and the copyright owner(s) 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 September 2021
doi: 10.3389/fpsyg.2021.602977

Face Management and Negative Strengthening: The Role of Power Relations, Social Distance, and Gender

Nicole Gotzner[1*†] *and Diana Mazzarella*[2*†]

[1] Cognitive Sciences, Department of Linguistics, University of Potsdam, Potsdam, Germany, [2] Cognitive Science Centre, University of Neuchâtel, Neuchâtel, Switzerland

Negated gradable adjectives often convey an interpretation that is stronger than their literal meaning, which is referred to as 'negative strengthening.' For example, a sentence like 'John is not kind' may give rise to the inference that *John is rather mean*. Crucially, negation is more likely to be pragmatically strengthened in the case of positive adjectives ('not kind' to mean *rather mean*) than negative adjectives ('not mean' to mean *rather kind*). A classical explanation of this polarity asymmetry is based on politeness, specifically on the potential face threat of bare negative adjectives (Horn, 1989; Brown and Levinson, 1987). This paper presents the results of two experiments investigating the role of face management in negative strengthening. We show that negative strengthening of positive and negative adjectives interacts differently with the social variables of power, social distance, and gender.

Keywords: conversational implicature, negation, politeness, social meaning, antonymy, adjectives

OPEN ACCESS

Edited by:
Valentina Cuccio,
University of Messina, Italy

Reviewed by:
Paolo Canal,
University Institute of Higher Studies in Pavia, Italy
Bob van Tiel,
Radboud University Nijmegen, Netherlands

****Correspondence:***
Nicole Gotzner
gotzner@uni-potsdam.de
Diana Mazzarella
diana.mazzarella@unine.ch

†These authors share first authorship

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

Received: *04 September 2020*
Accepted: *26 May 2021*
Published: *27 September 2021*

Citation:
Gotzner N and Mazzarella D (2021) Face Management and Negative Strengthening: The Role of Power Relations, Social Distance, and Gender. Front. Psychol. 12:602977. doi: 10.3389/fpsyg.2021.602977

INTRODUCTION

The last few years have seen a growing interest in the experimental investigation of the role of social context in language comprehension. Social relations among interlocutors as well as social expectations in communication (e.g., politeness) have recently been experimentally manipulated to examine their effect on the interpretation of certain linguistic expressions, such as quantifiers and expressions of uncertainty (for reviews, see Holtgraves and Bonnefon, 2017; Holtgraves, 2019). The seminal paper of Bonnefon et al. (2009) opened up the question of the role of face management in the interpretation of utterances containing scalar terms like 'some.' Based on a series of experimental studies, Bonnefon and colleagues put forward the claim that the scalar term 'some' is less likely to be interpreted as conveying a pragmatically strengthened meaning (*some but not all*) when the utterance represents a threat to the positive social identity or 'face' of the addressee ('Some people hated your poem') than when it does not ('Some people loved your poem'). These findings have been expanded - and debated - in subsequent studies that focused on further scalar expressions such as the connective *or* (Feeney and Bonnefon, 2012) and were investigated with distinct experimental techniques (i.e., reaction-times: Bonnefon et al., 2011; Mazzarella et al., 2018; and electrophysiology: Holtgraves and Kraus, 2018). As of today, though, this emerging experimental literature has not yet addressed the question of the role of face management in the interpretation of other linguistic expressions or constructions, beyond scalar and uncertainty expressions. This paper aims at filling this gap by looking at the interpretation of negated adjectives.

The interpretation of negated adjectives has long received the attention of philosophers, linguists and cognitive scientists due to its intuitive asymmetry (see e.g., Jespersen, 1917; Ducrot, 1973; Hoffmann, 1987; Horn, 1989; Colston, 1999; for more recent contributions see Giora et al., 2005; Krifka, 2007; Ruytenbeek et al., 2017; Gotzner et al., 2018). Consider a pair of antonymic adjectives like 'kind' and 'mean.' Crucially, the negation of the positive adjective ('John is not kind') is more likely to be interpreted as an affirmation of the antonym (*John is rather mean*) than the negation of the negative adjective ('John is not mean' interpreted as *John is rather kind*). This amounts to saying that positive adjectives are more likely to give rise to an inference called 'negative strengthening' compared to negative ones (Horn, 1989).

Interestingly for our purposes, it has been suggested that the use of negated positive adjectives ('John is not kind') to convey a negative interpretation (*John is rather mean*) can be seen as a politeness strategy by which the speaker may reduce the face threat toward the addressee carried by her speech act (Brown and Levinson, 1987; Horn, 1989; Levinson, 2000). 'John is not kind' is thus preferred to 'John is mean' as the former carries a reduced, less open, face threat (on the assumption that *kindness* is a desirable property). The addressee can unravel this reasoning and thus derive a strengthened interpretation of the negation. In the case of negated negative adjectives, or 'double negatives,' however, the affirmative statement often does not carry any potential face threat. For the running example, there is usually no reason relating to face management why speakers cannot directly say that 'John is kind.' Withholding this positive term may instead indicate that the situation does not quite match it (Levinson, 2000, p. 144). Thus, 'John is not mean' is interpreted as a middling term (e.g., *John is neither kind nor mean*) rather than licensing the inference to *John is kind*[1].

The current work experimentally investigates the role of face management in the interpretation of positive and negative antonyms. It does so by looking into multiple sociological variables that calibrate the expected politeness level among interlocutors. In particular, we test the politeness explanation for the polarity asymmetry in two experiments by manipulating the social context in the following ways: (1) by inverting the power relation between the speaker and the hearer and (2) by manipulating their social distance. Based on the politeness explanation of negative strengthening, our main hypothesis is that these sociological variables interact with polarity in that they will mainly play a role in the interpretation of negated positive adjectives. Furthermore, we examine the role of participant gender and speaker gender in negative strengthening. As previous research has emphasized the relationship between face management and gendered communicative practices, we explore whether this relationship carries over to the pragmatic interpretation of negated adjectives.

Our paper is organized as follows. We first describe previous research on negative strengthening and introduce the framework of Politeness Theory by Brown and Levinson (1987). Second, we review key findings in the literature on language and gender and discuss their relevance for the present study. Third, we present our two experiments manipulating adjectival polarity and the sociological variables of power and distance. Experiment 1 focuses on power relations and Experiment 2 on the social distance between the speaker and the hearer. Finally, we discuss the results of our experiments in light of broader face management considerations based on both the speaker's face and the hearer's face and identify open questions for future research.

NEGATED ADJECTIVES AND SOCIAL CONTEXT

The Polarity Asymmetry of Negative Strengthening

The phenomenon of negative strengthening concerns the interpretation of negated antonymic adjectives. According to Horn (1989), negative strengthening arises when "under the right conditions, a formally contradictory negation not-F will convey a contrary assertion G" (Horn, 1989, p. 273). That is, under the right conditions, an utterance of 'John is not kind' ('not-F'), which semantically encodes a meaning spanning from the *zone of indifference* between 'kind' and 'mean' to the contrary 'mean,' can be used to implicate that *John is rather mean* ('G') (for an alternative view, which models the gap between the extension of positive and negative antonyms as a pragmatic effect, see Krifka, 2007). When this is the case, the interpretation of the negated antonym ('not kind') is strengthened to convey *rather mean*. Crucially, according to this view, negative strengthening is an implicature, and, as such, it is a defeasible content. The utterance 'John is not kind' may lead the hearer to derive the implicature that *John is rather mean*, but this implicature can be defeated by a continuation like 'But he is not mean either. Simply, don't expect much support from him.' Furthermore, the defeasibility of the strengthened meaning gives the speaker the possibility, if openly challenged, to retract it and to deny to have had the intention to convey such a meaning (see e.g., Lee and Pinker, 2010 on the deniability of indirect speech acts).

Research on negative strengthening has highlighted the following observation, which we refer to as the *polarity asymmetry of negative strengthening* (Bolinger, 1972; Ducrot, 1973; Brown and Levinson, 1987; Horn, 1989; Blutner and Solstad, 2000; Levinson, 2000): the negation of a positive polarity antonym ('not kind') is more likely to be strengthened than the negation of a negative polarity antonym ('not mean')[2].

This generalization appeals to a notion of *polarity*, which allows us to distinguish between positive and negative antonyms. Polarity is traditionally defined based on the following three criteria (see Cruse, 1986). First, subjective judgments of desirability and undesirability (the so-called 'evaluative polarity'): desirability maps onto positive polarity and undesirability maps onto negative polarity (Boucher and Osgood, 1969;

[1]But see Horn (1989) and Krifka (2007) who point out that certain uses of double negatives do imply the positive form. We return to this point in the General discussion.

[2]The downtoning effect of negation, thanks to which double negatives appear to have a weakened attenuated sense, is known since Jespersen (1917).

Horn, 1989). For instance, 'kind' is desirable thus positive, 'mean' is undesirable thus negative. Second, the relevance of a certain dimension on the associated scale (the so-called 'dimensional polarity'): the relevant dimension maps onto positive polarity. For instance, 'tall' and 'short' are associated with a scale of height, so 'tall' is positive and 'short' is negative (since the positive term is associated with a higher measurement value). Third, markedness based on morphological negation: markedness maps onto negative polarity. 'Unhappy' is morphologically marked by the negative affix un-, thus it is negative, 'happy' is unmarked, thus positive. While these three criteria often converge, there are also possible mismatches (see Cruse, 1986; Ruytenbeek et al., 2017 for a discussion).

The polarity asymmetry of negative strengthening has recently been confirmed by a rigorous experimental study. Ruytenbeek et al. (2017) employed both an acceptability judgment task and an inferential task to test participants' interpretation of negated antonymic adjectives. The first task involved an indirect measure of negative strengthening based on acceptability judgments of sentences of the form 'X is not P. Y is Q too,' where P and Q represent an antonymic pair (e.g., in French *Paul n'est pas grand. Pierre aussi est petit.*). The second task allowed the collection of explicit inferential judgments by asking participants to judge the subject of the sentence on a continuous scale anchored at the antonyms P and Q (e.g., *Paul n'est pas grand* judged on a scale from *grand* to *petit*). Their results confirmed the expected polarity asymmetry: participants were more likely to strengthen the interpretation of negated positive antonyms than the interpretation of negated negative antonyms. They also showed that polarity interacts with morphological markedness, that is, the polarity asymmetry was greater for morphological pairs (containing negative morphemes such as 'happy' and 'unhappy') than for non-morphological pairs (involving lexical antonyms like 'happy' and 'sad'). These results are in line with previous experiments by Colston (1999) and Fraenkel and Schul (2008). However, studies by Giora et al. (2005) and Paradis and Willners (2006) did not find an asymmetric pattern for the interpretation of positive and negative antonyms. Interestingly, these studies revealed that the interpretation of negated terms was dissimilar from their lexical antonyms for both positive and negative adjectives. For instance, the bare negative 'sad' was interpreted as conveying a lower degree of happiness than 'not happy' (as predicted by Krifka, 2007; see also Tessler and Franke, 2018; under review).

Explaining the Polarity Asymmetry in Terms of Politeness

A traditional explanation of the polarity asymmetry of negative strengthening goes back to Horn (1989) and is framed in the context of Politeness theory (Brown and Levinson, 1987). According to Brown and Levinson, the interaction between speakers and hearers is typically regulated by face concerns, where face is defined as "the public self-image that every member [of a society] wants to claim for himself" (Brown and Levinson, 1987, p. 61, building on Goffman, 1967). Crucially, speakers might employ specific linguistic strategies - that Brown and Levinson call "politeness strategies" - to avoid or minimize a potential face loss. The speaker's motivation to opt for a politeness strategy is a function of the level of face threat carried by their act ("weight of the face-threatening act"). Brown and Levinson (1987) identify three sociological variables influencing the calculation of the weight of a face-threatening act (Wx): power (P), distance (D), and ranking of imposition (R).

$$Wx = P(H, S) + D(S, H) + Rx \quad (1)$$

P is the asymmetric social dimension of relative power that the hearer H has over the speaker S. The more powerful H is over S, the greater the weight of the face-threatening act. For instance, an utterance of 'Your publication list is not rich' would be more face-threatening when addressed to the Head of department by a student intern than the other way around. D is the symmetric social dimension of similarity/difference within which S and H stand for the purpose of the act x (typically based on the frequency of interaction and the exchange of social goods). The greater the distance between H and S, the greater the weight of the face-threatening act. Hence, 'Your publication list is not rich' is more face-threatening when addressed to a researcher you have just met at a conference than to your office mate. Finally, R is the ranking of imposition that the act x entails in a certain culture. The greater the imposition, the greater the weight of the face-threatening act. 'Could you please proofread my grant proposal?' is then more face-threatening than 'Could you check my 250-word abstract?'. Overall, as the weight of the face-threatening act increases, speakers are more likely to adopt some politeness strategy. Importantly for our purposes, Brown and Levinson (1987) identify off-record strategies (including understatement) as politeness strategies that allow the speaker to avoid the responsibility of their communicative act by "leav[ing] it open to the addressee to decide how to interpret it" (1987, p. 211).

Drawing upon this framework, the negation of a positive antonym ('not kind') can be seen as a politeness strategy to mitigate the face threat carried by the alternative containing the other member of the antonymic pair ('mean'). The threat might be a threat toward the face of the speaker, who wants to be perceived as benevolent and guarded (as in 'John is not kind'), and/or toward the face of the addressee, who wants to be spared a direct criticism (as in 'You are not kind'). As a result of politeness considerations, hearers may identify this strategy and consequently strengthen the interpretation of the negated positive antonym to convey its contrary. Crucially, though, the negation of a negative adjective ('not mean') does not make politeness a relevant consideration for the addressee. This is because the bare positive ('kind') would not elicit any potential face threat. It follows that politeness considerations facilitate negative strengthening in the case of negated positive adjective, but not in the case of negated negative adjective. Hence, the observed polarity asymmetry: "the relevant strengthening inference will tend to be favored in contexts [.] where there is some plausible reason to mask the speaker's true opinion. These contexts characteristically involve [.] those gradable predications involving desirable properties, those whose denial would reflect

undesirably on the subject, speaker, and or/addressee" (Horn, 1989, p. 334).

This traditional explanation of the polarity asymmetry calls for an experimental investigation to receive empirical support. A previous study from Gotzner et al. (2018) investigated the relationship between negative strengthening and scalar inferences[3]. In the context of this study, the authors collected participants' ratings of the kindness/politeness of statements involving negated adjectives. This study did not find any evidence of a correlation between politeness ratings and degree of negative strengthening. However, as the authors acknowledge, "to discover effects of politeness, test sentences may have to be embedded within a rich conversational context in future studies and politeness may have to be manipulated directly in the experimental setup (Gotzner et al., 2018, p. 11). The current study takes up this challenge and experimentally investigates the role of face management in negative strengthening.

Face Management and Gender

Before turning to our study, it is worth addressing the question of the relevance of gender to face management. Since the seminal work of Lakoff (1973), research on language and gender has focused on identifying specific gendered communicative practices as well as interpreting their significance in interaction. Early accounts argued for the existence of a relationship between linguistic features, such as hedges, tag questions, indirect requests and women's subordinate social status in male-dominated environments. They thus identified power (or lack of) as the driving force of gendered communicative practices (see e.g., Lakoff, 1975). Later research, however, revealed the role of further dimensions, not reducible to status, in accounting for the observed language differences between women and men. For instance, Holmes's (1984) and Cameron et al. (1988) empirical investigations challenged the idea that tag questions unequivocally express a lack of confidence or tentativeness and showed that their use served different functions: while men mainly used tag questions to express uncertainty, women tended to employ them as politeness devices to facilitate conversations or soften criticisms. Further examples of the prominence of solidarity-oriented behaviors in women language were found in the analysis of women's gossip as well as women's feedback to conversational partners (see e.g., Coates, 1988)[4].

This view of women as supportive conversationalists is echoed by much research showing that, in many cultural and conversational contexts, women tend to be more polite than men (for an overview, see Chalupnik et al., 2017). For instance, Holtgraves and Yang (1992) experimentally demonstrate that women produce more polite requests than men (see also Baxter, 1984). Crucially, women are also *expected* to be more polite than men and are judged more severely than men when they fail to meet this expectation. In an experimental study focusing on alignment with the interlocutor's opinion and compliance with their requests, Roberts and Norris (2016) found that a male delay was more tolerated than a female one: a female delay induced participant's lower agreeableness ratings than an equivalent male delay.

These findings suggest that women and men exhibit (or are normatively expected to exhibit) differences in their face management. For the purpose of our study, it is thus relevant to investigate whether the pragmatic phenomenon of negative strengthening reveals any gender differences. For this reason, we included an exploratory analysis of the role of participant gender (Experiment 1 and Experiment 2) as well as of speaker gender (Experiment 2). Crucially, because most literature on gender and face-management focuses on gendered language production rather than comprehension, it is an open question whether similar patterns may emerge in pragmatic interpretation. If comprehension mirrored production, we would expect female participants to be more likely than male participants to interpret the use of a negated positive adjective as a politeness device and thus strengthen its negation. Concerning speaker gender, we would expect higher rates of negative strengthening when the utterance containing the negation of a positive adjective is attributed to female speakers. None of the previous studies on negative strengthening has looked into gender effects. Therefore, it is possible that some discrepancies across studies were caused by gender differences (e.g., the absence of a polarity asymmetry in the studies by Giora et al., 2005; Paradis and Willners, 2006).

THE CURRENT STUDY

Overview of Experiments

The aim of our experiments was two-fold. On the one hand, we aimed to assess the robustness of the polarity asymmetry and replicate the results obtained by Ruytenbeek et al. (2017). On the other hand, we aimed to investigate the role of face management with respect to this asymmetry. In two experiments, we tested the hypothesis that face management considerations affect the interpretation of negated positive adjectives in the following way: the greater the weight of the face-threatening act, the more likely the negation of the adjective will be pragmatically strengthened to convey its negative antonym. According to Brown and Levinson (1987), the weight of the face-threat depends, among other factors, on the power relation between speaker and hearer (P) and their social distance (D). For instance, if the speaker is in a less powerful position than the hearer, the speaker will be more likely to employ a politeness strategy to reduce the face threat carried by the speech act to be performed in a given context. In Experiment 1, we manipulated the power relation between speaker and hearer and in Experiment 2 the social distance between them. In addition, we analyzed participant gender as an exploratory analysis in Experiments 1 and 2, and we further manipulated the gender of the speaker in Experiment 2.

In each experiment, we embedded 20 negated antonym pairs (e.g., 'not kind' and 'not mean') in a context involving two dialog partners. Participants were asked to judge the speaker's intended meaning on a 1-7 point Likert scale with 1 representing the adjective used in the original (negated) statement (e.g., 'kind') and

[3]See also Leffel et al. (2019) who show that relative adjectives like 'tall' but not minimum standard adjectives like 'late' are negatively strengthened (in the 'not very' construction, e.g., 'John is not very tall').

[4]For an extensive overview of the interdisciplinary research on language and gender, see Eckert and McConnell-Ginet (2013).

7 representing its antonym (e.g., 'mean'). We pre-registered the experiments on OSF with the main prediction of an interaction between polarity and our sociological variables (power/distance) (Experiment 1[5]: Experiment 2[6]).

Experiment 1: Power Relations

Goals and Predictions

The first experiment investigates the role of power in negative strengthening for positive and negative adjectives by inverting the power relation between the speaker and the hearer (e.g., the professor talking to a student and vice versa). We use a 2×2 within-subject Latin Square design with polarity (positive, negative) and power (high power speaker, low power speaker) as factors. Our key dependent variable is the degree of negative strengthening. We measure the degree of negative strengthening by using a 7-point Likert scale anchored at the negated adjective (1) and its antonym (7).

Based on Brown and Levinson's (1987) and Horn's (1989) account of negative strengthening, we predict that the negation of a positive adjective is more likely to be pragmatically strengthened in the low power speaker condition than in the high power speaker condition. On the contrary, we expect no effect of power with respect to the pragmatic strengthening of negated negative adjectives. With regard to the comparison between positive and negative antonyms, we predict an interaction between polarity and power. Ruytenbeek et al. (2017) show a strength asymmetry in the negative strengthening of positive versus negative adjectives and attribute it to polarity. We predict that the asymmetry across positive and negative adjectives will be significantly stronger when the context makes politeness consideration relevant (low power speaker) than when it does not (high power speaker).

Methods

Participants

We recruited 60 participants with US IP addresses on Mechanical Turk (30 participants across two experimental lists). Participants were screened for native language and only included in the analysis if their self-reported native language was English. 34 men and 25 women participated in the study (one participant did not provide a response to the gender question). Their mean age was 37.15, with a standard deviation of 12.1 (age range 21 to 72). The experiment was conducted in accordance with the ethics policy of the Deutsche Forschungsgemeinschaft (DFG) under grant Nos. BE 4348/4-1 and GO 3378/1-1. Formal approval from an Ethics Committee is not required for adult studies according to national regulations. Participant's consent was obtained at the start of the survey and their data were fully anonymized. The experiment lasted about 15 min and participants were paid 1 US Dollar in compensation.

Materials

We used the adjectives from Ruytenbeek et al. (2017) that had consistent polarity across different measures (markedness, evaluativity, and dimensionality). The items in the latter study were in French and we verified that English translation equivalent had the same polarity. In total, 20 adjectives pairs were used with their positive and negative antonyms occurring in a negated statement, thus totaling 40 critical items. The statements were embedded in a dialog between a speaker and hearer and preceded by a context sentence. **Table 1** shows an example stimulus. The complete list of stimuli is available in **Supplementary Appendix A**.

TABLE 1 | Example item for the adjective *fair* in Experiment 1 (positive polarity, low power speaker).

Context: At a staff gathering in the factory meeting room, the boss has presented the work-schedule he prepared for that day.
The boss asks an employee: "How do you find the schedule?"
The employee replies: "Your schedule is not fair"
According to the employee, the schedule is:
fair 1 2 3 4 5 6 7 unfair

The speaker who uttered the critical statement was either in high power position and the hearer in low power position (e.g., the boss responding to an employee's question) or vice versa. We relied on the following three power relations: boss-employee, professor-student, editor-intern. The task of the participants was to indicate what the speaker wanted to communicate[7]. For example, in the sample stimulus, participants judged the extent to which - according to the speaker - the schedule is fair/unfair. Judgments were given on a 7-point Likert scale anchored at the negated adjective (1) and its antonym (7). Hence, we measure the degree of negative strengthening as a function of the likelihood with which the antonym of a pair is taken to be conveyed by the speaker's utterance.

Our two factors, polarity and power, were all within-subject but spread across two different item lists in a Latin square design. Each participant saw 20 statements with positive and 20 statements with negative adjectives, rotated over power conditions. Hence, each participant completed 40 critical trials. The resulting overall number of critical observations was 2400. In addition to the critical items, participants were presented with 8 filler statements not involving negation, for example statements like 'John is gorgeous' (where the response scale was anchored the adjectives 'gorgeous' and 'ugly'). The filler sentences also served as attention checks.

The experiment was programmed in HTML and run via Mturk's in-built environment. The pre-registration form of the first experiment is available at the following link: (see text footnote 5)[8].

Procedure

Participants read an instruction explaining the task with an example. They were told to judge what the individuals wanted to communicate. The running example was an adjective not

[5]https://osf.io/d5e6u

[6]https://osf.io/knrdz

[7]In contrast to Ruytenbeek et al.'s (2017) study, our test question involved the explicit attribution of the intended implication to the speaker.

[8]We had originally planned to run linear models as the study by Ruytenbeek et al. (2017) did. Upon closer inspection, we noticed that, in this study, the response variable was re-coded into a binary variable. In the meantime, cumulative link mixed effects models have become the standard to analyze ordinal data. Since such models are more appropriate for ordinal data than linear models, we decided to analyse our data with the former kind.

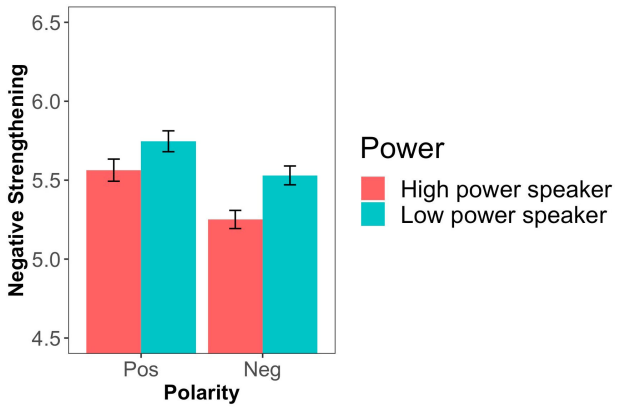

FIGURE 1 | Mean degree of negative strengthening by adjective polarity and power condition (Experiment 1). Error bars represent ± 1 SEM.

used in the stimulus set (John asks Mary: How do I look? and Mary responds: You are not gorgeous). For each stimulus, the 1-7 point scale was anchored to the adjective used in the speaker's statement (1) and its antonym (7). The instructions told participants to judge what the speaker wanted to convey in each dialog. Experimental trials and filler trials were randomized for each participant using an in-built randomization function.

Results

The data were analyzed using R (version 3.6). We excluded four participants based on inconsistent responses in the filler trials (more than 50% responses not in line with the bare adjective used in the filler statements, i.e., a response of 5, 6, or 7). **Figure 1** shows the mean responses by adjective polarity and power condition.

All results were analyzed with cumulative link mixed effects models using the function clmm() in the ordinal package (Christensen, 2018), which are more appropriate for Likert scales than linear mixed models[9]. We included the fixed factors power, polarity, their interactions as well as random intercepts for items and participants. All fixed factors were sum coded. The results of the model showed a main effect of polarity with positive adjectives involving a higher degree of negative strengthening than negative ones ($B = -0.35$, $SE = 0.04$, $z = -8.27$, $p < 0.0001$). This finding replicates the polarity asymmetry discussed in previous work (e.g., Ruytenbeek et al., 2017). In addition, there was a main effect of power with a higher degree of negative strengthening for speakers in a low power position than in a high power position ($B = -0.2$, $SE = 0.04$, $z = -4.76$, $p < 0.0001$). The interaction between polarity and power was not significant ($p = 0.5$). A summary of the model is presented in **Table 2**.

As an exploratory analysis, we computed a model with participant gender as an additional treatment-coded variable. Female participants were chosen as the reference level based

[9]The function clmm() is the more recent variant of clmm2(), allowing for the implementation of multiple random effects. However, at the time of running our experiments and analyzing the data, no random slopes were implemented for ordinal models (see Christensen, 2018).

TABLE 2 | Summary of cumulative link mixed effects model including the sum-coded fixed effects power and polarity (Experiment 1).

	Estimate	*SE*	*z*-value	*p*-value
Polarity	−0.35461	0.04287	−8.272	0.0001
Power	−0.20108	0.04228	−4.756	0.0001
Polarity: power	−0.17623	0.27286	−0.646	0.518

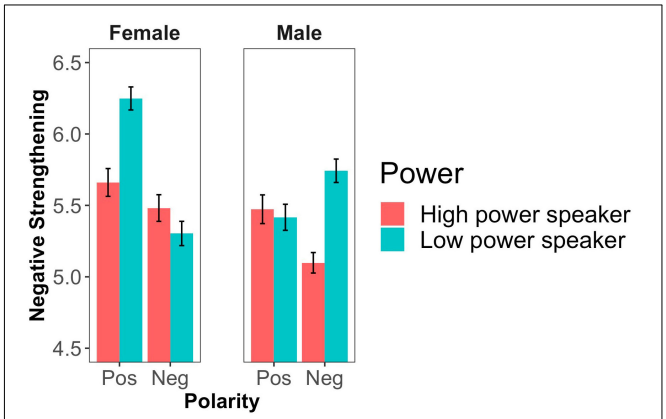

FIGURE 2 | Mean degree of negative strengthening by adjective polarity, power and participant gender (Experiment 1). Error bars represent ± 1 SEM.

TABLE 3 | Summary of cumulative link mixed effects model including the sum-coded fixed effects power, polarity and treatment-coded fixed effect participant gender with females as the reference level (Experiment 1).

	Estimate	*SE*	*z*-value	*p*-value
Polarity	−0.69527	0.06782	−10.252	0.0001
Power	−0.18941	0.06654	−2.847	0.00442
Participant gender male	−0.50794	0.54546	−0.931	0.35175
Polarity: power	0.39707	0.40973	0.969	0.33249
Polarity: participant gender male	0.59673	0.08761	6.812	0.0001
Power: participant gender male	−0.10074	0.08714	−1.156	0.24767
Polarity: power: participant gender male	−0.97942	0.54605	−1.794	0.07287

on the previous literature suggesting that women tend to be more polite than men (for an overview, see Chalupnik et al., 2017). The model again revealed main effects for polarity and power. Further, there was an interaction between participant gender and polarity ($B = 0.59$, $SE = 0.087$, $z = 6.81$, $p < 0.0001$) as well as a tendency toward a three-way interaction of polarity, power and participant gender ($B = -0.098$, $SE = 0.54$, $z = -1.79$, $p = 0.07$). The interaction between participant gender and polarity reveals a larger polarity asymmetry for female participants than male participants. Furthermore, the tendency toward a three-way interaction with power indicates that female participants displayed an effect of power for positive adjectives but not negative adjectives while male participants displayed the opposite pattern. The means across conditions and participant gender are displayed in **Figure 2** and the results of the cumulative link mixed effects model are presented in **Table 3**.

Discussion

Experiment 1 showed an asymmetry of negative strengthening for positive and negative adjectives. That is, negated positive terms were more likely to be strengthened than negated negative terms, replicating the findings of Ruytenbeek et al. (2017)[10]. In addition, we found a main effect of power: negative strengthening was more likely to occur when the speaker was in a low power position as opposed to a high power position. Contrary to the main prediction, we did not find an interaction between polarity and power. That is, independently of the polarity of the adjective, participants were more inclined to interpret low power speakers' utterances as indirect affirmations of the contrary.

As previous research suggests the existence of gender differences in face-management, we conducted an exploratory analysis with gender as a binary factor. We found an interaction between polarity and participant gender with female participants displaying a greater degree of negative strengthening for positive compared to negative adjectives. Furthermore, there was marginal three-way interaction indicating that female participants showed a tendency for a stronger effect of the power manipulation for positive adjectives, which goes in the direction of the predicted interaction. Male participants, in turn, were mainly affected by the power manipulation for negative adjectives. That is, male participants were more likely to strengthen a negated negative statement when the speaker was in the low power position.

Experiment 2: Social Distance

Goals and Predictions

The second experiment manipulated social distance with the speaker and the hearer being either close friends (low social distance) or having just met (high social distance). The professions used in Experiment 1 were replaced with common names. All dialogs were between same-gender names, with half of them including stereotypically female names and the other half stereotypically male names (speaker gender manipulation). Our main prediction, based on Brown and Levinson (1987) and Horn (1989), was again an interaction between polarity and distance. That is, participants should be more likely to strengthen the negation of positive adjectives when the addressee is socially distant than when the addressee is socially close. Since Experiment 1 showed a trend for the predicted interaction across polarity and power for female participants, we included participant and speaker gender in our analysis.

Methods

Participants

We recruited another set of 60 participants with US IP addresses on Mechanical Turk (30 participants across two experimental lists). Participants were screened for native language and only included in the analysis if their self-reported native language was English. One participant's native language was Italian and the data were therefore excluded from further analyses. The remaining 59 participants were 30 men and 29 women with a mean age of 37.18 and a standard deviation of 11.5 (age range: 22 to 65). The experiment was conducted in accordance with the ethics policy of the Deutsche Forschungsgemeinschaft (DFG) under grant Nos. BE 4348/4-1 and GO 3378/1-1. Formal approval from an Ethics Committee is not required for adult studies according to national regulations. Participant's consent was obtained at the start of the survey and their data were fully anonymized. The experiment lasted about 15 min and participants were paid 1 US Dollar in compensation.

TABLE 4 | Example item for the adjective *fair* in Experiment 2 (negative polarity, high social distance, female speakers).

Context: Sue and Mary just started working in the same company. At a staff gathering in the factory meeting room, Mary has presented the work-schedule she prepared for that day.
Mary asks Sue: "How do you find the schedule?"
Sue responds: "Your schedule is not unfair"
According to Sue, the schedule is:
unfair 1 2 3 4 5 6 7 fair

Materials

The materials were the same as in Experiment 1 but we replaced noun phrases with common names (e.g., John and Paul). Social distance was manipulated by describing the two characters as either friends (low social distance) or having just met (high social distance). Dialogs were always between speakers of the same gender and we created items in which either two men or two women interacted (based on stereotypical names). We used the most common American English names for men and women. **Table 4** shows an example stimulus. The complete list of stimuli is available in **Supplementary Appendix B**. The pre-registration form of the second experiment is available at the following link: (see text footnote 6)[11].

Procedure

The procedure was the same as that of Experiment 1.

Results

Four participants were excluded from further analyses for giving inconsistent responses in filler trials (more than 50% responses not in line with the bare adjective used in the filler statements, i.e., a response of 5, 6, or 7). **Figure 3** shows the mean ratings across polarity and social distance conditions. In **Figure 4**, we present the results across participant gender and in **Figure 5** across speaker gender.

We first ran a cumulative link model involving the factors social distance, polarity, and their interaction as well as a random intercept for participants and items. Again, we replicated the polarity effect ($B = -1.05$, $SE = 0.048$, $z = -21.94$, $p < 0.0001$). The main effect of social distance was not significant and neither

[10]Following Ruytenbeek et al. (2017), we also looked at the extent to which morphologically complex pairs (e.g., *happy-unhappy*) triggered a greater degree of negative strengthening compared to simple antonymic pairs (e.g., *happy-sad*). The additional model revealed an interaction between the factors polarity and morphological complexity in the predicted direction ($B = -0.12$, $SE = 0.042$, $z = -2.81$, $p < 0.01$), thus replicating the results of Ruytenbeek et al. (2017).

[11]The pre-registration reports the manipulation of speaker gender. As the main hypotheses derived from the politeness explanation concerned the interaction of polarity and social distance, only these were considered in our pre-registered analysis plan.

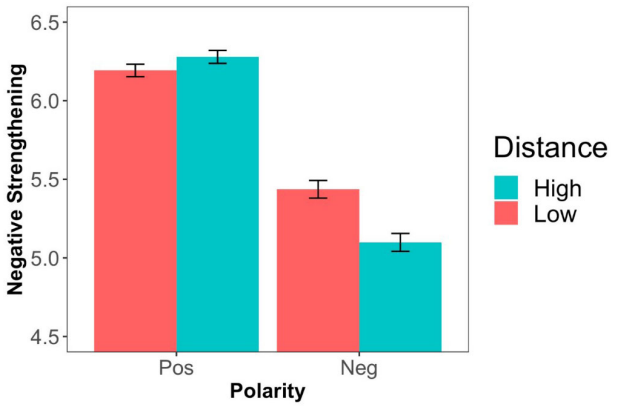

FIGURE 3 | Mean degree of negative strengthening by adjective polarity and social distance (Experiment 2). Error bars represent ± 1 SEM.

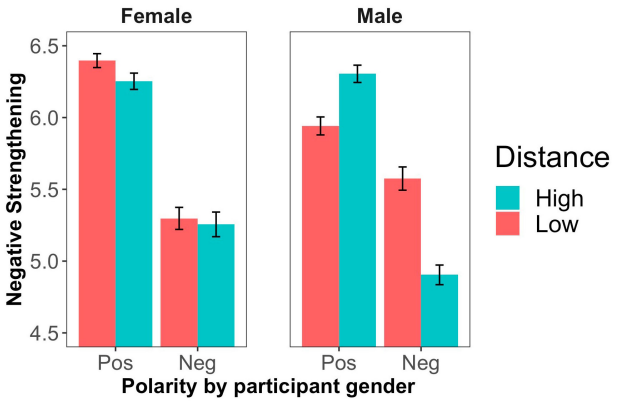

FIGURE 4 | Mean degree of negative strengthening by adjective polarity, social distance and participant gender (Experiment 2). Error bars represent ± 1 SEM.

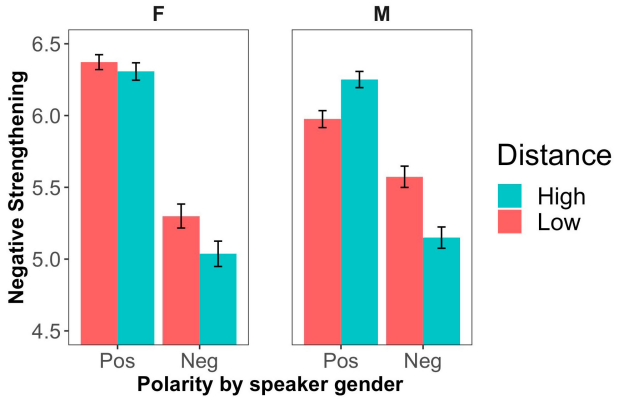

FIGURE 5 | Mean degree of negative strengthening by adjective polarity, social distance and speaker gender as manipulated in Experiment 2 (labeled with 'F' for female and 'M' for male). Error bars represent ± 1 SEM.

TABLE 5 | Summary of cumulative link mixed effects model including the sum-coded fixed effects polarity and social distance (Experiment 2).

	Estimate	*SE*	*z*-value	*p*-value
Polarity	−1.04967	0.04784	−21.94	0.0001
Distance	−0.05998	0.0427	−1.405	0.16
Polarity: distance	−0.29434	0.22804	−1.291	0.197

TABLE 6 | Summary of cumulative link mixed effects model including the sum-coded fixed effects polarity, social distance, and treatment-coded fixed effect participant gender with females as the reference level (Experiment 2).

	Estimate	*SE*	*z*-value	*p*-value
Polarity	−1.15166	0.06447	−17.865	0.0001
Distance	−0.05732	0.06017	−0.953	0.3407
Participant gender male	−0.38879	0.43874	−0.886	0.3755
Polarity: distance	0.12415	0.30197	0.411	0.681
Polarity: participant gender male	0.20435	0.08546	2.391	0.0168
Distance: participant gender male	−0.01142	0.08559	−0.133	0.8939
Polarity: distance: participant gender male	−0.88616	0.43893	−2.019	0.0435

was the interaction between polarity and distance, as presented in **Table 5**[12].

Further, we ran a second model including the additional treatment-coded variable participant gender (female as reference level). In addition to the main of effect of polarity ($B = -1.15$, $SE = 0.06$, $z = -17{,}7$, $p < 0.0001$), this model revealed an interaction between polarity and participant gender ($B = 0.20$, $SE = 0.09$, $z = 2.39$, $p < 0.05$), showing again that female participants displayed a greater polarity asymmetry compared to male participants. Further, there was three-way interaction between polarity, distance and participant gender ($B = -0.09$, $SE = 0.04$, $z = -2.01$, $p < 0.05$). This interaction indicated that female participants were less affected by the social distance manipulation and that male participants differentially strengthened positive and negative terms depending on social distance. The detailed results are presented in **Table 6**.

[12]Additionally, we computed a model to see whether the polarity asymmetry was greater for morphologically complex antonym pairs vs. simple ones and we again found a significant interaction ($B = -0.36$, $SE = 0.043$, $z = -8.46$, $p < 0.0001$).

Finally, we ran a model with polarity, social distance, participant gender and speaker gender (i.e., our manipulated variable of the dialog partner's stereotypical names) as factors. This model showed main effects of polarity ($B = -1.43$, $SE = 0.09$, $z = -15.7$, $p < 0.0001$), social distance ($B = 0.24$, $SE = 0.09$, $z = 2.79$, $p < 0.01$), an interaction between polarity and speaker gender ($B = 0.52$, $SE = 0.12$, $z = 4.3$, $p < 0.0001$), an interaction between social distance and speaker gender ($B = -0.36$, $SE = 0.12$, $z = -2.96$, $p < 0.01$), an interaction between polarity and participant gender ($B = 0.24$, $SE = 0.12$, $z = 1.98$, $p < 0.05$) as well as three-way interactions between polarity, distance and speaker gender ($B = 0.28$, $SE = 0.13$, $z = 2.2$, $p < 0.05$) and polarity, distance and participant gender ($B = 0.96$, $SE = 0.45$, $z = 2.13$, $p < 0.05$) (see **Table 7**). For male names, negative strengthening was more likely for socially close dialog partners when the adjective was negative, but for socially distant dialog

TABLE 7 | Summary of cumulative link mixed effects model including the sum-coded fixed effects polarity, social distance, participant gender, and speaker gender with female participants and female names as the reference level (Experiment 2).

	Estimate	*SE*	*z*-value	*p*-value
Polarity	−1.42778	0.0909	−15.707	0.0001
Distance	0.24481	0.08782	2.787	0.00531
Speaker gender male	−0.1274	0.14441	−0.882	0.37768
Participant gender male	−0.42005	0.454	−0.925	0.35485
Polarity: distance	−0.2582	0.31353	−0.824	0.4102
Polarity: speaker gender male	0.52103	0.12063	4.319	0.0001
Distance: speaker gender male	−0.36401	0.12316	−2.955	0.00312
Polarity: participant gender male	0.24358	0.12295	1.981	0.04757
Distance: speaker gender male	−0.12515	0.12306	−1.017	0.30919
Speaker gender male: participant gender male	0.05489	0.17031	0.322	0.74723
Polarity: Distance: speaker gender male	0.28304	0.12872	2.199	0.02788
Polarity: distance: participant gender male	0.9655	0.45425	2.125	0.03355
Polarity: speaker gender male: participant gender male	−0.07521	0.17027	−0.442	0.65873
Distance: speaker gender male: participant gender male	0.2608	0.17038	1.531	0.12585
Polarity: distance: speaker gender male: participant gender male	−0.14146	0.17036	−0.83	0.40634

partners when the adjective was positive. Dialogs involving female names also showed a greater polarity than those with male names, as evident in the interaction between polarity and speaker gender. Speaker gender and participant gender did not show any interactions.

Discussion

In our second experiment, we replicated the polarity asymmetry of negative strengthening as well as the interaction between polarity and participant gender. As in Experiment 1, negated positive adjectives were more likely to be strengthened than negated negative adjectives and this asymmetry was stronger for female participants than for male participants. Furthermore, we found that social distance had distinct effects for positive and negative adjectives, depending on participant/speaker gender. For male participants/speakers, negative strengthening was more likely to occur when dialog partners were socially close if the adjective was negative ('not mean' to mean *rather kind*). However, when the adjective was positive, in line with our prediction, negative strengthening was more likely to occur ('not kind' to mean *rather mean*) when dialog partners were socially distant. In contrast with this, for female speakers/participants, there was no effect of the distance manipulation for positive adjectives. When the adjective was negative, female speakers (but not female participants) also gave rise to more negative strengthening for socially close dialog partners.

GENERAL DISCUSSION

Across our two experiments, we examined the role of two of the sociological variables identified by Brown and Levinson (1987), that is, power (P) and social distance (D), on the interpretation of negated antonyms. Furthermore, we investigated the existence of possible gender effects by looking both at participant gender and speaker gender.

Across our experiments we found two clear and consistent patterns: (i) the existence of a polarity asymmetry in the interpretation of negated adjectives; (ii) an interaction between adjectival polarity and participant gender. The first finding represents an important replication of Ruytenbeek et al.'s (2017) results by using contextually richer scenarios (as opposed to decontextualized sentences) and confirms the reliability of our adapted paradigm. The second finding reveals that the polarity asymmetry of negative strengthening is modulated by participant gender: female participants display a stronger polarity asymmetry than male participants, as evident in consistent interactions between polarity and participant gender across our two experiments.

Furthermore, we showed that power and social distance both had an effect on the interpretation of negated antonyms. However, in contrast with our main prediction, their effect differed in the following way. With respect to power, Experiment 1 showed that the greater the power of the hearer over the speaker, the stronger was the degree of negative strengthening (with an interesting tendency for an interaction between polarity and participant gender, as revealed by our exploratory analysis). With respect to social distance, Experiment 2 revealed the following interaction with polarity and participant gender. For male participants, the greater the social distance between the speaker and the hearer, the stronger was the degree of negative strengthening for positive adjectives. Furthermore, the smaller the social distance between the speaker and the hearer, the stronger was the degree of negative strengthening for negative adjectives.

Overall, the results do not support a straightforward explanation of the polarity asymmetry of negative strengthening based on politeness considerations, as the one suggested by Brown and Levinson (1987) and Horn (1989). This traditional explanation would have predicted an interaction between polarity and the social variables of power and distance, such that greater negative strengthening for positive adjectives should have occurred when the speaker was in a low power position compared to the addressee (Experiment 1) or socially distant to the addressee (Experiment 2). Crucially, our results reveal a more

complicated picture, one in which the social variable of gender plays an important role. Indeed, the expected effects of power and distance on negative strengthening for positive adjectives selectively appeared only when gender was factored in.

In Experiment 1, female participants – but not male participants – tended to be more likely to strengthen the negation of a positive adjective when the speaker was in a low power position compared to a high power position. That is, when confronted with an utterance like 'Your schedule is not fair,' which represents a potentially face-threatening act, female participants tended to interpret it as a function of the relative power of the hearer over the speaker. They were more likely to strengthen their interpretation toward *Your schedule is unfair* when the speaker was an employee and the addressee the boss, than the other way around.

In Experiment 2, male participants – but not female participants – were more likely to strengthen the negation of a positive adjective when the speaker was in a high distance relationship with the hearer than when they were socially close. That is, when confronted with an utterance like 'Your schedule is not fair,' male participants were more likely to attribute the intention to communicate a strengthened interpretation (*Your schedule is rather unfair*) to socially distant speakers than to socially close ones.

Interestingly, our results suggest that female and male individuals might differ in their attribution of the intention to minimize a face-threat when the interaction involves the expression of evaluations via negated adjectives ('Your schedule is not fair'). First of all, the results indicated that female participants are more likely than male participants to strengthen the negation of a positive adjective ('Your schedule is not fair' to mean *Your schedule is rather unfair*). This suggests that female participants are more prone than male participants to take the negation of a positive adjective as an indirect negative evaluation. Such a gendered interpretative behavior is in line with previous literature suggesting that women are more likely than men to appeal to standard politeness strategies such as indirectness (Holtgraves and Yang, 1992). As suggested by Brown (1980), women are "more sensitive from moment to moment to the face-threatening potential of what they are saying and modify their speech accordingly" (Brown, 1980, p. 131). This parallelism indicates that not only do women rely on polite indirectness more often than men, but they are also more likely to attribute this strategy to speakers in potentially face-threatening situations. Interestingly, data from Experiment 2 highlight that the strength of the polarity asymmetry of negative strengthening also varies as a function of speaker gender: the negation of positive adjectives is more likely to be strengthened when the utterance is attributed to a female speaker. This may indicate that, consistently with findings from Roberts and Norris (2016), participants expected female speakers to be more polite than male speakers, and thus interpreted the use of negation in utterances like 'Your schedule is not fair' as polite indirectness to communicate that *Your schedule is rather unfair*.

Furthermore, our data reveal gendered interpretations of negated antonyms as a function of both power and social distance. For instance, when looking at the negative strengthening of negated positive adjectives, we found that that female participants were more likely to attribute face-saving intentions as a function of power, while male participants as a function of social distance. It is worth noticing that these results suggest different sensitivities to face-threat across genders in relation to the interpersonal nature of the context (for a similar conclusion, see Holtgraves and Yang, 1992). As suggested by Holtgraves and Yang (1992), this may be the result of differences in the *perceptions* of the situation on the power and social distance dimensions, and/or differences in the *weighting* of power and social distance. This suggests that a full appreciation of the formula provided by Brown and Levinson to describe the way in which power and social distance influence the weight of a face-threatening act, $Wx = P(H,S) + D(S,H) + Rx$, cannot overlook some important dimensions of variation, such as cultural patterns that hold for specific groups or social categories. Among these, the gender of an individual appears to be linked to normatively stabilized expectations about the way in which they will perceive or weigh a face-threatening act, thus giving rise to regularities in interactional strategies.

Finally, there is one interesting finding that deserves further attention. Across both experiments, male participants showed greater variability in their interpretation of negated negative adjectives, or double negatives, as a function of the interpersonal context. When confronted with an utterance like "Your schedule is not unfair," male participants derive a strengthened interpretation (*Your schedule is rather fair*) in the following two circumstances: when the speaker is in a relatively low power position and when the speaker is socially close to the hearer. This unexpected result opens up the question of the role of double negatives in interactions and their gendered interpretation. While this is ultimately an empirical question, we suggest that future research might benefit from focusing on considerations about the speaker's face. As Brown and Levinson have argued at length, face-management ordinarily involves considerations about both speaker and hearer face. In his discussion of negative strengthening, Horn mentions that in Western cultures there is sometimes a taboo to state positive emotions directly and to show excessive enthusiasm (Horn, 1989, p. 359). Because of this, speakers may use a weak statement ('Your schedule is not unfair') as a "studied modesty of expression" (Stoffel, 1901, p. 126) in order to safeguard their face, e.g., in order not to appear as overly positive. Indeed, there is some independent evidence that men are expected to temper their positivity to preserve their perceived power (see Sattel, 1983). This kind of face-management concern might have played a role in the pattern of interpretation observed for male participants. Male participants might assume that the speaker will not want to appear overly positive when complimenting a more powerful addressee (hence avoiding being perceived as motivated by the opportunistic desire of pleasing the addressee) or a friend (hence avoiding showing overt admiration). This suggestion fits well with established gender differences in paying compliments (see e.g., Holmes, 1988; Herbert, 1998). Furthermore, it has been noted that certain uses of double negatives convey an interpretation that is stronger than the bare positive as a form of polite understatement (Lyons, 1977; Horn, 1991; Levinson, 2000; Krifka, 2007). An example of this is the use of 'not bad' to mean *very good* (for the role of prosody in eliciting this interpretation see Bolinger, 1972, p. 116). Future work

is needed to distinguish the role of the speaker's face and the hearer's face in negative strengthening, as well as their interaction with gender. Overall, our findings fit well with the theoretical assumptions of recent modeling of polite speech (see e.g., Yoon et al., 2017, to appear), which suggest that politeness emerges from competing social goals. By applying a model comparison approach, these studies show that, over and beyond the informative utility of the communicative act, speakers rely on considerations of pro-social as well as self-presentational utilities when designing their utterances in potentially face-threatening contexts. The role of self-presentational considerations in the adoption of face-management strategies targeting the speaker's face represents a topical issue for future experimental research. In future work, we will extend the current manipulations to the area of language production.

Another productive line of future research concerns the investigation of different notions of polarity. An experiment by Mazzarella and Gotzner (2021) revealed that the standard polarity asymmetry in negative strengthening holds even in contexts in which the face-threatening potential of positive and negative utterances is reversed. These findings indicate a role of adjective polarity that is somewhat independent of face management considerations. However, all previous studies in this area have tested adjective pairs that are consistently positive or negative in terms of evaluative, dimensional polarity and markedness. We propose that investigating polarity mismatches (e.g., 'dirty,' which is evaluatively negative but dimensionally positive) will provide crucial insights into the mechanisms underlying negative strengthening and its polarity asymmetry.

In sum, the results of our study indicate that while face-management considerations have an impact on the interpretation of negated adjectives, but this impact is not limited to the interpretation of positive adjectives. Both positive and negative negated adjectives might undergo a process of negative strengthening as a function of the power relation and social distance among the interlocutors. This suggests that the interplay between face-management and negative strengthening is more complex than previously assumed and it opens up new lines for future research.

CONCLUSION

The present study investigated the role of face management in negative strengthening by manipulating the social context in two different ways: via the manipulation of the power relation between the dialog partners (Experiment 1) and their social distance (Experiment 2). Furthermore, it investigated the presence of gender effects by manipulating the identity of the speaker and examining the interpretative behaviors of female and male participants. The study provided empirical support to the polarity asymmetry of negative strengthening, in line with results from Ruytenbeek et al. (2017). In both experiments, we observed a significant effect of polarity on the interpretation of negated adjectives: positive adjectives were more likely to be negatively strengthened than negative adjectives. Crucially, though, we found that the social context affected the degree of negative strengthening for both positive and negative adjectives. This is in contrast with the prediction based on Brown and Levinson (1987) and Horn (1989) that social context should affect the interpretation of positive adjectives only. While Horn (1989) does consider the role of a taboo to state positive emotions directly in certain cultures, the main explanation of the polarity asymmetry concerned the face-threatening potential of bare negatives. What is more, this standard explanation did not anticipate the complex interactions between different sociological variables. Overall, our results indicate that negative strengthening is the result of wider face management considerations, which might concern both the speaker's intention to mitigate the threat toward the face of the addressee and the speaker's intention to save their own face.

The present study also reveals the existence of gendered expectations about the use and meaning of linguistically conveyed evaluations via negated adjectives. As gender represents an important attribute of an individual's identity, these results confirm the interconnection between face-management and identity, whose importance has been foregrounded by more recent approaches in politeness research (see e.g., Spencer-Oatey, 2009). Furthermore, research on the relationship between language and gender has primarily focused on language production. Our results thus contribute to extend this investigation to the domain of pragmatic interpretation.

Finally, our study broadens the array of pragmatic phenomena that have been investigated with the aim of addressing the question of the interface between politeness and pragmatic inference (see e.g., Bonnefon et al., 2009; Feeney and Bonnefon, 2012; Mazzarella et al., 2018). By focusing on negative strengthening, it enriches our understanding of the way in which language interpretation depends on social context.

DATA AVAILABILITY STATEMENT

The data of Experiment 1 are available at https://osf.io/67he9/ and those of Experiment at https://osf.io/3kqfv/.

ETHICS STATEMENT

The experiment was conducted in accordance with the ethics policy of the Deutsche Forschungsgemeinschaft (DFG) under grant Nos. BE 4348/4-1 and GO 3378/1-1. Formal approval from an Ethics Committee is not required for adult studies according to national regulations. Participant's consent was obtained at the start of the survey and their data were fully anonymized.

AUTHOR CONTRIBUTIONS

NG implemented the experiments, performed the statistical analyses, and wrote the first draft of Section "The Current Study" and parts of Section "Introduction." DM wrote the first draft of Sections "Introduction, Negated Adjectives and Social Context,

General Discussion, and Conclusion". Both authors contributed to the conception and design of the study and edited the manuscript, and read and approved the final submitted version.

FUNDING

This work was supported by the German Research Foundation (DFG) as part of the Xprag.de Initiative (Grant No. BE 4348/4-1) as well as an Emmy Noether grant awarded to NG (Grant No. GO 3378/1-1).

ACKNOWLEDGMENTS

We acknowledge the support from the Open Access Publishing Fund of the University of Potsdam.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.602977/full#supplementary-material

REFERENCES

Baxter, L. A. (1984). An investigation of compliance gaining as politeness. *Hum. Commun. Res.* 10, 427–456. doi: 10.1111/j.1468-2958.1984.tb00026.x

Blutner, R., and Solstad, T. (2000). "Dimensional designation: a case study in lexical pragmatics," in *Studies in Optimality Theory*, eds R. Blutner and G. Jäger (Potsdam: University of Potsdam), 30–40.

Bolinger, D. (1972). *Degree Words*. The Hague: Mouton.

Bonnefon, J.-F., De Neyes, W., and Feeney, A. (2011). "Processing scalar inferences in face-threatening contexts," in *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, eds L. Carlson, C. Hölscher, and T. Shipley (Austin, TX: Cognitive Science Society), 3389–3395.

Bonnefon, J.-F., Feeney, A., and Villejoubert, G. (2009). When some is actually all: scalar inferences in face-threatening contexts. *Cognition* 112, 249–258. doi: 10.1016/j.cognition.2009.05.005

Boucher, J., and Osgood, C. E. (1969). The pollyanna hypothesis. *J. Verbal Learn. Verbal Behav.* 8, 1–8.

Brown, P. (1980). "How and why are women more polite: some evidence from a mayan community," in *Women and Language in Literature and Society*, eds S. McConnell-Ginet, R. Borker, and N. Furman (New York, NY: Praeger), 111–136.

Brown, P., and Levinson, S. C. (1987). *Politeness: Some Universals in Language Usage*. Cambridge: Cambridge University Press.

Cameron, D., McAlinden, F., and O'Leary, K. (1988). "Lakoff in context," in *Women in Their Speech Communities*, eds J. Coates and D. Cameron (London: Longman), 74–93.

Chalupnik, M., Christie, C., and Mullany, L. (2017). "(Im)politeness and gender," in *The Palgrave Handbook of Linguistic (Im)politeness*, eds J. Culpeper, M. Haugh, and D. Kádár (London: Palgrave), 517–537. doi: 10.1057/978-1-137-37508-7_20

Christensen, R. H. B. (2018). Cumulative link models for ordinal regression with the R package Ordinal. *J. Stat. Software*.

Coates, J. (1988). "Gossip revisited: language in all-female groups," in *Women in Their Speech Communities*, eds J. Coates and D. Cameron (London: Longman), 94–122.

Colston, H. L. (1999). "Not good" is "bad," but "not bad" is not "good": an analysis of three accounts of negation asymmetry. *Discourse Process.* 28, 237–256. doi: 10.1080/01638539909545083

Cruse, A. (1986). *Lexical Semantics*. Cambridge: Cambridge University Press.

Ducrot, O. (1973). *Le rôle de la Négation Dans le Langage Ordinaire. La Preuve et le Dire*. Tours: Mame.

Eckert, P., and McConnell-Ginet, S. (2013). *Language and Gender*. New York, NY: Cambridge University Press.

Feeney, A., and Bonnefon, J.-F. (2012). Politeness and honesty contribute additively to the interpretation of scalar expressions. *J. Lang. Soc. Psychol.* 32, 181–190. doi: 10.1177/0261927X12456840

Fraenkel, T., and Schul, Y. (2008). The meaning of negated adjectives. *Intercult. Pragmat.* 5, 517–540.

Giora, R., Balaban, N., Fein, O., and Alkabets, I. (2005). "Negation as positivity in disguise," in *Figurative Language Comprehension: Social and Cultural Influences*, eds H. L. Colston and A. N. Katz (Mahwah NJ: Lawrence Erlbaum Associates), 233–258.

Goffman, E. (1967). *Interaction Rituals: Essay in Face-to-Face Behavior*. Chicago, IL: Aldine Pub. Co.

Gotzner, N., Solt, S., and Benz, A. (2018). Scalar diversity, negative strengthening, and adjectival semantics. *Front. Psychol.* 9:1659. doi: 10.3389/fpsyg.2018.01659

Herbert, R. K. (1998). "Sex-based differences in compliment behavior," in *The Sociolinguistics Reader: Gender and Discourse*, Vol. 2, eds J. Cheshire and P. Trudgill (London: Arnold), 53–75.

Hoffmann, M. (1987). *Negatio Contrarii: A Study of Latin Litotes*. Assen: Van Gorcum.

Holmes, J. (1984). Hedging your bets and sitting on the fence: some evidence for hedges as support structures. *Te Reo* 27, 47–62.

Holmes, J. (1988). Paying compliments: a sex-preferential politeness strategy. *J. Pragmatics* 12, 445–465. doi: 10.1016/0378-2166(88)90005-7

Holtgraves, T. (2019). "Politeness," in *Experimental Semantics and Pragmatics*, eds C. Cummins and N. Katsos (Oxford: Oxford University Press), 512–523.

Holtgraves, T., and Bonnefon, J. F. (2017). "Experimental approaches to linguistic (im) politeness," in *The Palgrave Handbook of Linguistic (Im) politeness*, eds M. Haugh and J. Culpeper (London: Palgrave Macmillan), 381–401. doi: 10.1057/978-1-137-37508-7_15

Holtgraves, T., and Kraus, B. (2018). Processing scalar implicatures in conversational contexts: an ERP study. *J. Neurolinguistics* 46, 93–108. doi: 10.1016/j.jneuroling.2017.12.008

Holtgraves, T., and Yang, J.-N. (1992). Interpersonal underpinnings of request strategies: general principles and differences due to culture and gender. *J. Pers. Soc. Psychol.* 62, 246–256. doi: 10.1037/0022-3514.62.2.246

Horn, L. (1989). *A Natural History Of Negation*. Chicago, IL: University of Chicago Press.

Horn, L. R. (1991). Given as new: when redundant affirmation isn't. *J. Pragmatics* 15, 313–336. doi: 10.1016/0378-2166(91)90034-u

Jespersen, O. (1917). *Negation in English and Other Languages*. Copenhagen: Høst.

Krifka, M. (2007). "Negated antonyms: creating and filling the gap," in *Presupposition and Implicature in Compositional Semantics. Palgrave Studies in Pragmatics, Language and Cognition*, eds U. Sauerland and P. Stateva (Houndmills: Palgrave Macmillan), 163–177. doi: 10.1057/9780230210752_6

Lakoff, R. T. (1973). Language and woman's place. *Lang. Soc.* 2, 45–80.

Lakoff, R. T. (1975). *Language and Woman's Place*. New York, NY: Harper and Row.

Lee, J. J., and Pinker, S. (2010). Rationales for indirect speech: the theory of the strategic speaker. *Psychol. Rev.* 117:785. doi: 10.1037/a0019688

Leffel, T., Cremers, A., Gotzner, N., and Romoli, J. (2019). Vagueness in implicature: the case of modified adjectives. *J. Semant.* 36, 317–348. doi: 10.1093/jos/ffy020

Levinson, S. C. (2000). *Presumptive Meanings: The Theory of Generalized Conversational Implicature*. Cambridge, MA: MIT Press.

Lyons, J. (1977). *Semantics*. Cambridge: Cambridge University Press.

Mazzarella, D., and Gotzner, N. (2021). The polarity asymmetry of negative strengthening: dissociating adjectival polarity from face-threatening potential. *Glossa* 6:47. doi: 10.5334/gjgl.1342

Mazzarella, D., Trouche, E., Mercier, H., and Noveck, I. (2018). Believing what you're told: politeness and scalar inferences. *Front. Psychol.* 9:908. doi: 10.3389/fpsyg.2018.00908

Paradis, C., and Willners, C. (2006). Antonymy and negation—the boundedness hypothesis. *J. Pragmatics* 38, 1051–1080. doi: 10.1016/j.pragma.2005.11.009

Roberts, F., and Norris, A. (2016). Gendered expectations for 'agreeableness' in response to requests and opinions. *Commun. Res. Rep.* 33, 16–23. doi: 10.1080/08824096.2015.1117437

Ruytenbeek, N., Verheyen, S., and Spector, B. (2017). Asymmetric inference towards the antonym: experiments into the polarity and morphology of negated adjectives. *Glossa* 2:92. doi: 10.5334/gjgl.151

Sattel, J. W. (1983). "Men, inexpressiveness, and power," in *Language, Gender, and Society*, eds B. Thorne, C. Kramarae, and N. Henley (Rowley, MA: Newbury House), 118–124.

Spencer-Oatey, H. (2009). "Face, identity and interactional goals," in *Face Communication and Social Interaction*, eds F. Bargiela-Chiappini and M. Haugh (London: Equinox), 137–154.

Stoffel, C. (1901). *Intensives and Down-toners. Anglistische Forschungen, part 1*. Heidelberg: Carl Winter's Universitätsbuchhandlung

Tessler, M. H., and Franke, M. (2018). "Not unreasonable: carving vague dimensions with contraries and contradictions," in *Proceedings of the 41st Annual Meeting of the Cognitive Science Society*, (Stanford: Stanford University).

Yoon, E. J., Tessler, M. H., Goodman, N. D., and Frank, M. C. (2017). ""I won't lie, it wasn't amazing": modeling polite indirect speech," in *Proceedings of the 39th Annual Conference of the Cognitive Science Society*, (Stanford: Stanford University).

Conflict of Interest: 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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Gotzner and Mazzarella. 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) and the copyright owner(s) 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: 28 October 2021
doi: 10.3389/fpsyg.2021.663168

A Study of Zero Anaphora Resolution in Chinese Discourse: From the Perspective of Psycholinguistics

Ning Yang[1†], *Jingyu Zhang*[2†], *Lijun Ma*[3*] *and Zhi Lu*[2,4*]

[1] School of Education, South China Normal University, Guangzhou, China, [2] School of Interpreting and Translation Studies, Guangdong University of Foreign Studies, Guangzhou, China, [3] School of Public Health and Management, Guangzhou University of Chinese Medicine, Guangzhou, China, [4] Center for Translation Studies, Guangdong University of Foreign Studies, Guangzhou, China

Anaphora is the main linguistic means to establish discourse coherence, and anaphora resolution is the psychological process to maintain this coherence. Chinese discourse is characterized with providing multiplicity of linguistic clues to readers by employing various referential apparatuses such as pronoun anaphora, zero anaphora, and so on. As a way of avoiding repeated reference to an object that is mentioned beforehand, zero anaphora is frequently employed in discourse. The production and resolution of zero anaphora largely concerns some constraints underlying psychological mechanisms. We particularly focus on zero anaphora resolution in the present study to try to discover some specific aspects of the underlying mechanism, hoping to find out some factors unique to the resolution process. We designed the first two experiments to probe into the psychological reality when participants were presented with sentences containing either pronoun anaphora or zero anaphora or both under discourse condition with topic continuity in Experiment 1a and topic discontinuity in Experiment 1b. We did not find any significant difference in the reaction time between zero anaphora resolution and pronoun anaphora resolution, indicating that zero anaphora possibly works within the processing mechanism on which pronoun anaphora resolution depends. However, we found significantly longer time in reading the first sentence in any of the discourse, showing that the first-mention effect exists in anaphora resolution. We further explored the time course of zero anaphora resolution by measuring the reaction time during the period when participants read sentences that varied according to the location where zero anaphora occurred under two conditions: topic continuity (Experiment 2a) vs. topic discontinuity (Experiment 2b). The strategies of searching for the referential information were found divergent: the exhaustive searching strategy was adopted when the topics within a discourse were kept continuous and the heuristic searching strategy was employed when the topics were discontinuous. The design of Experiment 5 took the factor of voice type and situation consistency into consideration, investigating in what way do those factors influence the resolution of zero anaphora. The voice type, according to the results, plays a significant role for its exclusively close relationship with the first-mention effect.

Keywords: zero anaphora resolution, psychological reality, time course, topic continuity, situation consistency

OPEN ACCESS

Edited by:
Valentina Cuccio,
University of Messina, Italy

Reviewed by:
Tao Gong,
Educational Testing Service, United States
Hui Zhang,
Nanjing Normal University, China

***Correspondence:**
Lijun Ma
malj@gzucm.edu.cn
Zhi Lu
zhi.lu@foxmail.com

†These authors have contributed equally to this work

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

Received: *02 February 2021*
Accepted: *22 September 2021*
Published: *28 October 2021*

Citation:
Yang N, Zhang J, Ma L and Lu Z (2021) A Study of Zero Anaphora Resolution in Chinese Discourse: From the Perspective of Psycholinguistics. Front. Psychol. 12:663168. doi: 10.3389/fpsyg.2021.663168

INTRODUCTION

Anaphora is interpreted divergently. From the perspective of traditional grammar, it is described as a grammatical relationship among linguistic elements, for example, Crystal (2008) thinks that anaphora is used to "refer to some other sentence constituent." Anaphora is also considered as a grammatical device created for the purpose of avoiding the repetition of a prior mentioned information. Scholars of pragmatics (such as Levinson), however, hold the opinion that anaphora is the identity of entities (e.g., Hirst, 1981). For any entity that is designated as the "correct" one, there is a choice from a set of possible anaphoric expressions during the dynamic course of discourse production. This brings the anaphora into the scope of keeping the topic continuous during discourse comprehension and establishes the theoretical basis on which the anaphora is analyzed from the perspective of pragmatics. Huang (2000) further develops the pragmatic model, known as the neo-Gricean pragmatic approach, by employing some basic pragmatic strategies such as Levinson's Q-, I-, and M-principles, and proposes that those pragmatic principles account for the anaphora referent. Despite some minor differences, the pragmatic models believe that the anaphora plays a role as a substitution of the structural constituent and regard the reference and its referent as co-reference. The key fact is that how the anaphoric or co-referential relationship is established in mind and in what way do people to resolve the anaphora. Indeed, the anaphora resolution is "the process by which an antecedent is assigned to an anaphora" (McDonald and MacWhinney, 1995). On the one hand, from the perspective of anaphoric production, people are instinctive to avoid information redundancy, so they prefer to use reference to achieve this goal. On the other hand, from the vantage point of anaphoric resolution, listeners or readers have to develop the ability of identifying the possible referent and resolve it in order to understand the information. The two perspectives boil down to this: the construction and resolution of anaphora largely depends on anaphora distribution, which is a complex phenomenon involving structural, cognitive, and pragmatic factors that interact with each other (Huang, 2000). The question arises accordingly that how to examine the factors that take effect in exploring anaphora resolution. With the increasing number of the empirical studies, a lot of approaches have been proposed to try to answer the above question, such as the topic continuity or distance-interference model (Givón, 1983), the hierarchy model (Hinds, 1978, 1979; Tai, 1978; Longacre, 1979), and the cognitive model (Gundel et al., 1993; Lambrecht, 1994; Kibrik, 1996). The views are unsystematic and inconsistent because they mainly concern discourse processing. Anaphora is mentioned very briefly and discussed in a rather general way.

In fact, anaphora is essentially a way of reference to entities mentioned earlier in discourse. The discourse comprehension largely relies on resolving references used within the text with the same sense (i.e., referring back to the same entity) (Sukthanker et al., 2020). Anaphora resolution is thought complicated particularly when it appears in a discourse rather than in a sole sentence either in terms of the linear distance between the two mentions of referent or in terms of the number of interfering referents. Additionally, the different anaphoric forms that "reference" is resembled as also differentiate the difficulty of resolution: compared with zero anaphora resolution, pronoun anaphora resolution is supposed much easier since there is a tangible form of pronoun occurring in the place where anaphora is necessary. For example, in the sentence "*Alex (1) ran into her (2) room,*" (2) is the pronominal referent of (1). The referent in this sentence is presented with the antecedent "Alex" with the pronominal form "her," leading the reader to a correct identification of the relationship between the anaphora and its referent. Ariel (1990) believes that different forms of reference in discourse is to refer to different accessibility of referential entities, thus, he divides the degree of accessibility into three categories and points out, for example, in deictics, the distal (such as "that") are less accessible than the proximal (such as "this").

The study of anaphora not only involves solving problems of the anaphoric phenomenon, but also relates to establishing, retaining, and shifting of topic continuity, which, in turn, matters the discourse coherence. When we examine the example further, we find that the topic continuity also contributes to the fast understanding of the sentence by people, because the pronoun "her" serves as a clue for the reader to connect one topic "Alex" with another topic "room" together to keep the topic continuous. Contrary to the tangible form of pronoun anaphora, zero anaphora does not have a visible lexical and phonetic form. For example, in the sentence "*You have two choices (1): to stand still (2) vs. to move forward (3),*" phrases (2) and (3) refer back (are anaphoric) to the same phrase (1). The difficulty of resolving zero anaphora lies on the fact that there is no interfering referent to clearly indicate the anaphora relationship. This arises an important question "how to avoid the arbitrariness in the confirmation of anaphora when it does not appear physically in the discourse." One of the most involved type of encoding a referent is to evaluate with a short distance to its previous mention.

Zero anaphora is rarely seen in English, but frequently adopted as a way to refer back to previously named entities in Chinese. The reasons might be as follows: at first, compared with English, the sentence-oriented and subject-prominent language Chinese is more topic-prominent and discourse-oriented, requiring attention to attaining the topic within a macro discourse level rather than a sentence level. In another word, Chinese is what Song (2003) called the parataxis language, which is characterized with constructing the discourse coherence with analyzing the scattered grammatical components followed by discovering the semantic and internal logic among them. The evidence from Chinese narrative discourse shows that 91.3% of the Chinese zero anaphora have distributed within the same sentence and 93.4% of them are manifested as subjects of the sentence (Hou and Sun, 2005). Li and Thompson (1979) also prove that zero anaphora in Chinese is far more extensively used than that in English. Next, in many cases, there is typically one referent that is established by a given topic but shared by other unrealized topics among the main clause and its subordinates. This type of zero anaphora is called "topic chain," which is frequently occurring in Chinese discourse in order to keep topics

continuous. The topic referent is distinguished from the non-topic referent according to whether the anaphoric form is a reduced anaphoric form or a non or less-reduced anaphoric form. Meanwhile, a short referential distance and no interfering referent also affect the choice of topic referent, particularly in some East Asian languages. Thinking deeply about the distance-interference effect, the process of highlighting topic referent and inhibiting non or less-topic referent is rooted from the cognition of people (e.g., Tomlin and Pu, 1991) or pragmatic principles (e.g., Huang, 1989, 1994). Chinese, a thematic-prominent language, is particularly governed by the principles. On the one hand, the use of zero anaphora largely relies on pragmatic and textual knowledge (Zhang and Cui, 2001; Li, 2004). On the other hand, the repetition of verbs, the gradation of clauses, and the use of other types of anaphora probably cause topic discontinuity. Moreover, the introduction of more than one topic in discourse causes mental competition when people identify the topic referent by ruling out the possibility of the non or less-topic referent.

In Chinese, the referential relationship is not exclusively recognized by zero expression, but by pronouns (like *him*), full-noun phrases (like *Alex* or *the boy*), and reflexives (like *ziji*). Pronoun anaphora and noun anaphora are most commonly discussed in literatures, the notorious reflexive *ziji* (or "self"), however, attracts more and more attention. According to the theoretical linguistic literature, the use of reflexive illustrates the binding relationship (e.g., Huang, 1982; Mohanan, 1982; Wang and Stillings, 1984) and the discourse co-reference with its antecedents (e.g., Yu, 1991; Huang, 1994; Huang and Liu, 2001). The psycholinguistic researchers (e.g., Carroll, 1999; Gao et al., 2005; Liu, 2009) examined the cognitive process of identifying the referent of the Chinese reflexive *ziji* and multiple occurrences of the Chinese reflexive *zijis*. The binding effect (e.g., Carroll, 1999; Liu, 2009), the local-binding effect (e.g., Gao et al., 2005), and the long-distance binding effect (e.g., Shuai et al., 2013) are found. The ERP (Event-Related Potential) is an electrical brain response, recorded via electroencephalography, time-locked to the onset of an event such as a sensory stimulus or a motor act experiment, conducted by Li and Zhou (2010), confirms that the selection of a matrix subject as the long-distance antecedent of *ziji* violates the Principle A during the comprehension of sentences and requires more mental resources. The findings have revealed that some aspects concerning the mechanism above which Chinese reflexive is processed within a range of syntactic distance, the mental resources involvement and the time course. The use of reflexive relates to the understanding of the entity referents by people. Even though it shares similarities with zero anaphora, i.e., forming not only a referential relationship but also a distribution relationship, reflexive anaphora is quite different from zero anaphora because reflexive anaphora, in one way or another, mainly plays a role of intensifying the antecedent (like *Wo ziji*), functioning more as an adverb in grammar rather than the reference of the antecedent. *Z*ero anaphora, however, does not have such a portrait.

Chinese scholars have done extensive studies on zero anaphora either from the perspective of textual analysis (e.g., Lv, 1984; Xu, 1992; Tao and Healy, 2005) or from the pragmatic perspective (e.g., Xiong, 1999). Most of them, if not all, have reached the following consensus: (1) zero anaphora is more frequently used both in written and oral Chinese than it is used in English; (2) due to less strict grammatical rules, the location where zero anaphora occurs is flexible; (3) in some cases (e.g., job interview), zero anaphora is realized without the employment of antecedents; (4) in most cases, the use of zero anaphora is restricted by pragmatic and psychological factors; and (5) certain pronouns and conjunctions supply premise for the use of zero anaphora. These claims are mainly rooted from analyzing the linguistic features that Chinese zero anaphora possesses when it occurs under a particular circumstance on the basis of comparing it with other types of anaphora. However, there are some very important issues the claims do not explain well. For example, in such a case in which anaphors appear within the same clause as their antecedent, there seem to be strong syntactic constraints on the range of possible anaphoric forms. Comparing the two sentences "我喝了柠檬水,Φ很舒服" and "我喝了柠檬水,Φ很酸," we usually refer back to the antecedent of "我" (the subject of the first clause) in the first sentence, but of "柠檬水" (the object of the first clause) in the second sentence. The difference of the accessibility of more than one antecedent in the same discourse indicates that zero anaphora is discrete in space which is not merely governed by syntactic rules. It is assumed that there might be an underlying mechanism beyond the rules that are responsible for the resolution of zero anaphora.

To our knowledge, the majority of studies on Chinese anaphora resolution examined the relationship between pronoun anaphora and discourse comprehension, that is, in what way and in which aspects pronoun anaphora is influenced during the reading comprehension (e.g., Miao, 1994, 1996a,b; Miao and Song, 1995; Wang and Li, 1999; Sun et al., 2001; Zhou et al., 2001; Bai et al., 2005; Jiao and Zhang, 2005; Shen and Yang, 2006; Wang et al., 2006). Instead of paying close attention to anaphora itself, the studies focused on comparing reading strategies and comprehending abilities of processing materials provided in Chinese and a second language. Only a small number of studies painted a mixed picture of Chinese anaphora resolution in discourse processing (e.g., Wang and Mo, 2001; Zhao and Liu, 2006; Zhao and Mo, 2007). In spite of the rich achievements, most of the studies are restricted within the area of Chinese pronoun anaphora and its resolution. Research of zero anaphora was sporadic and unsystematic with little consideration from the psycholinguistics perspective. That is what we mainly concern in the present study: the psychological mechanism with which people can access the antecedent successfully without the indication of obvious referent. Additionally, since Givón (1979, 1995) repeatedly mentioned that topic continuity is the most important textual condition in affecting the use of zero anaphora, we took it into consideration when designing our experiments by addressing the existence and absence of topic continuity in constructing experimental materials.

THE PRESENT RESEARCH

Research Questions

The convergent idea is that, compared with pronoun anaphora and noun anaphora, the referent of zero anaphora is more difficult to identify. For example, when we encounter pronouns or repeated nouns in a discourse, we often automatically start to search the antecedent by referring back to the recent NP. Conversely, we can't follow the same route if we meet a sentence where zero anaphora is contained because there is a lack of tangible form used to encode a referent to its previous mention. In other words, unlike pronoun anaphora and noun anaphora, zero anaphora fails to be supported either by language form (e.g., *it* in pronoun anaphora) or by lexical meaning (e.g., *noun* or *noun phrase* in noun anaphora), or even sometimes by its antecedents (e.g., no antecedent). Accordingly, zero anaphora requires more mental resources than the other two types of anaphora and might be resolved in a way unique to others. We are wondering what mental activities are involved if we resolve zero anaphora during reading comprehension. Previous studies have proved that when anaphora resolution starts, not only an antecedent but also several competing antecedents and concepts co-occurring with them are activated as candidates (Corbett and Chang, 1983; Dell et al., 1983; Gernsbacher et al., 1989; Greene et al., 1992). Only those entities that are first mentioned with a grammatical marker in one way or another are considered closely related to the topic and are highly activated. The activation begins 250 ms after the presentation of the anaphora ends. Dell et al. (1983) used probe word insertion technique to explore to what extent the activation is achieved in noun anaphora resolution and find that both the antecedent of noun anaphora and its co-occurring concepts are activated but only the antecedent keeps being activated as the location of the probe word changes from one place to another. We are wondering in what way the zero anaphora is resolved and when the activation of the antecedent begins during the resolution. We are also curious about the factors related to the resolution of zero anaphora on the basis of existing studies, in which factors concerning pronoun anaphora resolution are summarized into the following categories: (1) the linear distance between the pronoun and its antecedent, (2) the causation implicitly contained in verbs, and (3) the first-mention effect on the accessibility of the pronoun antecedent.

Research Hypotheses

Hypothesis 1

With the same level of activation, the response time to the probe word in zero anaphora resolution is as fast as it is in pronoun anaphora resolution, no matter the topic is continuous or not. However, the reading time of sentences containing pronoun anaphora is longer than those containing zero anaphora and normal sentences for the ambiguous reference in pronoun anaphora examples.

Hypothesis 2

Being that there is a lack of tangible form of zero anaphora, people might use different strategies to cope with the situations in which the probe word is inserted in different positions: either right after the verb or at the end of the sentence. We expect that if the probe word is inserted after the verb, people use a heuristic strategy to search for the antecedent. If the probe word is inserted at the end of the sentence, people use an exhaustive strategy to search.

Hypothesis 3

The factors of situation consistency and voice type are critical to zero anaphora resolution. We expect that response time to the probe word is the fastest when the consistent situation is described in an active voice whereas the response time is the slowest when the inconsistent situation is described in a passive voice.

Experiment 1

This experiment adopted the research method from Gernsbacher (1989) to explore the psychological reality of zero anaphora resolution by measuring response time to the probe word under the condition of topic continuity and topic discontinuity. The materials were classified into three types, each of which contained one of the anaphors: pronoun anaphora, zero anaphora, and both pronoun and zero anaphora.

Experiment 1a: The Psychological Reality of Zero Anaphora Resolution Under the Condition of Topic Continuity

Subjects

A total of 20 sophomores with normal or corrected-to-normal vision took part in this experiment. They were all Chinese native speakers.

Materials and Design

Fifty-four short essays were made up as the experimental materials were made up of fifty-four short essays and fell into 18 groups randomly, each of which involved three types of anaphora named as A for pronoun anaphora, B for zero anaphora, and C for both. Each essay consisted of five sentences[1]. All the essays shared the same topic as well as the same sentence structure. Type C essays were the control materials that had been evaluated with the 5-point Likert Scale by juniors majoring Chinese Literature before the experiment and those rated above 4 were selected. The examples were shown below:

A /中午李明到了学校/接着她开始上课/课后她又到同学那里串门/她坐到快天黑时/她才恋恋不舍地告别回家了/ (pronoun anaphora)

B /中午李明到了学校/接着Φ开始上课//课后Φ又到同学那里串门/Φ坐到快天黑时/Φ才恋恋不舍地告别回家了/ (zero anaphora)

C /中午李明到了学校/接着她开始上课//课后Φ又到同学那里串门/Φ坐到快天黑时/她才恋恋不舍地告别回家了/ (control material)

[1]The five-sentence essays are constructed on the basis of Chen (1987) for the reason that the structure containing anaphora objects and the antecedents in discourse are generally not more than three levels, that is, not more than five sentences.

TABLE 1 | Average reaction time to two types of probe words on the basis of the stimuli varied in different anaphora types under the condition of topic continuity in Experiment 1a (ms).

Probe word type	Anaphora type		
	Pronoun anaphora	Zero anaphora	Control material
True antecedent	886.23	873.6	790.51
False antecedent	914.63	877.63	843.28

TABLE 2 | Average reading time of each clause within sentences that are varied in anaphora types under the condition of topic continuity in Experiment 1a (ms).

Anaphora type	The order of clause				
	1st clause	2nd clause	3rd clause	4th clause	5th clause
Type A	1,869.75	1,240.6	1,069.89	943.83	1,048.01
Type B	1,999.87	1,174.98	992.89	945.81	963.42
Type C	1,964.88	1,069.15	920.86	852.11	875.87

We adopted an offset equilibrium method, similar to the Latin square for the purpose of avoiding the familiarity effect. Specifically, we divided all participants into 6 groups, giving the first group type A material and used true antecedents as probe words (e.g., 李明). The second group was given type B material with false antecedents as probe words (e.g., 王亮) and so on. The experiment was a 3 (type of anaphora: A/B/C) ×2 (probe word: true /false antecedent) within-group design.

Procedure

The experiment was implemented by using E-prime software. Prior to the experiment, the signal "+" was first displayed on the screen for 600 ms, followed by a sentence presented character by character (a moving window technique) in the way in which the first character "中" of the first clause "中午李明到了学校" was presented on the screen for 300 ms and the next character followed and stayed for another 300 ms. They did not disappear until the last character "校" was given for 300 ms. After that, the next clause "接着开始上课" began to appear on the screen in the same way. When the last character of the whole sentence "了" disappeared and the screen was blank for a short period of time, and a probe word, either true to the discourse like "李明" or false like "王玲" appeared and lasted for 3 s. Participants were asked to judge as quickly as possible whether the probe word was previously presented. If participants made wrong judgments, a red sign was presented on the screen. Six hundred millisecond later, they had to answer a question to test their comprehension ability. The question lasted for 800 ms and feedback was given when the answer to the question was incorrect. All participants received a question after one trial to ensure that they had read the essay carefully. The procedure needed practicing beforehand.

Results

We deleted three extremes which were <1% of the total. The average reaction time has been listed in **Table 1**. Analysis of variance was used to reveal the interaction effect between the anaphora type and the probe word type. The difference was found insignificant with $F_{(2,\ 38)} = 0.348$, $p > 0.05$. The main effect of the reaction time to probe word type under three conditions was not significant [$F_{(2,\ 38)} = 1.708$, $p > 0.05$]. The main effect of probe word type was also not significant [$F_{(1,38)} = 1.7758$, $p > 0.05$].

The reading time of each clause was presented in **Table 2**. The ANOVA analysis (two-way repeated measures) showed that there was no interaction effect between sentence order and anaphora type [$F_{(8,112)} = 1.698$, $p > 0.05$]. The main effect of sentence order was significant [$F_{(4,56)} = 78.551$, $p < 0.001$], indicating that the reading time of each sentence varied in accordance with the sentence order: participants took the longest time in reading the first sentence. But the reading time was not significantly different among the rest of the sentences. We did not find the main effect of anaphora type [$F_{(2,\ 28)} = 2.814$, $p > 0.05$] showing that anaphora type did not produce any difference in reading time for participants.

Experiment 1b: Psychological Reality of Zero Anaphora Resolution Under the Condition of Topic Discontinuity

Subjects

Another forty sophomores with normal or corrected-to-normal vision took part in this experiment. They were all Chinese native speakers.

Materials and Design

The experiment materials were composed of 12 groups of essays with each group containing two types of essays. Type A Essay contained zero anaphora only and type B contained both pronoun anaphora and zero anaphora. For each group, type B essays were compiled by matching the topic and the number of sentences with type A essay and being evaluated on the Likert Scale by the same group of students in Experiment 1a. All type B essays were rated above 4 and were used as control materials. The examples were shown below:

A /林冰正在塘边放鸭子/Φ1养的鸭子个个又肥又大/突然Φ1看见几个小孩赤身露体/Φ2一起在池塘追逐打闹/Φ1便大声喊要他们小心/ (zero anaphora)

B /林冰正在塘边放鸭子/她养的鸭子个个又肥又大/突然她看见几个小孩赤身露体/Φ2一起在池塘追逐打闹/她便大声喊要他们小心/ (control material)

We adopted the same method, an offset equilibrium method, in designing this experiment. We allocated all the subjects into four groups and each group only did the experiment under one experimental condition. For example, the first group of participants were presented with type A essay followed by "林冰" (true antecedent as the probe word) and the second group with type B essay followed by "李明" (false antecedent as the probe word). The experiment adopted 2 (type A essay/type B essay) ×2 (true antecedent/false antecedent) within group design.

TABLE 3 | Average reaction time to two types of probe words on the basis of the stimuli varied in different anaphora types under the condition of topic discontinuity in Experiment 1b (ms).

Probe word type	Anaphora type	
	Zero anaphora	Control material
True antecedent	873.6	790.51
False antecedent	877.63	843.28

TABLE 4 | Average reading time of each clause within sentences that are varied in anaphora types under the condition of topic discontinuity in Experiment 1b (ms).

Anaphora type	The order of clause				
	1st clause	2nd clause	3rd clause	4th clause	5th clause
Type A	3386.05	2147.4	2014.57	1585.33	1492.28
Type B	3278.47	1935.74	1887.38	1589.24	1500.11

Procedure

The experiment was implemented following the same procedure in Experiment 1a.

Results

We deleted five extreme data which were <5% of the total and the rest were listed in **Table 3**. We used a two-way repeated measures ANOVA on reaction time to examine the interaction effect of essay type and probe word type. The interaction effect was not significant [$F_{(1,39)} = 0.687$, $p > 0.05$] nor was the main effect of probe word type [$F_{(1,39)} = 0.624$, $p > 0.05$], indicating that the participants did not show any difference in reading the two types of essays. The insignificant main effect of probe word type [$F_{(1,39)} = 0.657$, $p > 0.05$] revealed that regardless of the probe word type, the reaction time was not affected.

The reading time of each clause was presented in **Table 4**. An ANOVA was performed for the design of 2 × 2 two-factor repeated measurement of the reading time in Experiment 1b (**Table 4**). The results showed that the interaction between sentence order and essay type was not significant with $F_{(4,156)} = 2.079$, $p > 0.05$. The main effect of sentence order was significant [$F_{(4,156)} = 101.467$, $p < 0.001$]. The reading time of the first sentence was comparatively longer than that of the rest sentences. The main effect of essay type was not significant [$F_{(1,39)} = 0.739$, $p > 0.05$].

Discussion of Experiment 1a and 1b

Whether the topic was continuous or not, participants showed non-significant difference in reaction time during zero anaphora resolution. This was consistent with our hypothesis. We assumed that this insignificant role topic continuity played in discourse comprehension was due to the same degree of topic availability to readers. By closely examining the two sentences " /中午李明到了学校/……Φ才恋恋不舍地告别回家了/ (zero anaphora)" vs. "/林冰正在塘边放鸭子/……Φ1便大声喊要他们小心/ (zero anaphora)," we found that the topic of the first clause were easily referred back no matter whether they were the true antecedent or not. Even though the topic was not kept continuous, participants would also identify the antecedent to the first topic they encountered. The low demand of topic availability did not produce any significant differences in the reaction time. Wang et al. (2001) also proved that the anaphora could be resolved even when the antecedent was not correct to the anaphora. However, the anaphora type also did not exert any significant difference on reading time. We contributed the reason to the existence of more than one pronoun that misled people and caused ambiguity in resolving anaphora, so that the repetition of the same pronoun interrupted the smooth reading and forced participants to read back to catch up with the meaning and refer back to determine what the pronoun stood for. It is worth mentioning that probe words both in zero anaphora resolution and in pronoun anaphora resolution were activated at the same level, which was reflected in the insignificant reaction time in our experiment. In fact, according to Greene et al. (1992), the resolution of pronoun anaphora in discourse reading occurs automatically without being realized by people. Combined with our research results, we could imply that resolving zero anaphora might share the same underlying mechanism with resolving pronoun anaphora. However, it was contradictory to what Huang, 2001 pointed out: pronoun anaphora contains semantic meaning that is more transparent than zero anaphora does. In other words, zero anaphora was more difficult to resolve than pronoun anaphora.

In fact, Chinese language learners from foreign countries showed their avoidance of using zero anaphora (Xu and Xiao, 2008; Zhou, 2011). In contrast, Chinese native speakers preferred to utilize zero anaphora to indicate the referential relationship, as much as 36% of the total anaphora application (Kim, 2000). The difference might be caused by the strategy that was transferred from resolving pronoun anaphora to resolving zero anaphora within one language or among languages (Tao and Healy, 2005). But our results spoke against the theory proposed by Kintsch et al. (1999), who claimed that the thesis (argument overlap) is a key factor in maintaining coherence where there is the anaphora. The realization of argument overlap relied on the occurrence of noun phrase rendering, which was absent in zero anaphora. At this point, reading time should be increased, but we did not observe it no matter if the anaphora was pronominal or zero.

What we observed was that in spite of the different essay types, it took the participants significantly longer time in reading the first sentence than reading other sentences in both experiments. In fact, the first sentence played a role in building a foundation for comprehending the rest part and applying clues for combing the description of previous sentences and subsequent sentences. It required a lot of cognitive effort. "These initial segments help them establish the basis for the mental representation of larger units, such as sentences, paragraphs, and storylines" (Gernsbacher and Foertsch, 1999). As a result, more time was consumed. Other studies, such as Anderson (1983, 1985), and Anderson and Sanford (1983), also proved that the reading time of the first sentence was systematically and significantly longer than that of the other sentences.

We expected the increased reading time at those places where the topic of the discourse had been transferred, but we failed.

The reason might be due to the materials, which were made up through inserting a sentence intentionally to stop the topic for a while instead of changing the topic into a new one. For example, 王亦东推了自行车进了门, Φ瞧见李贵在刷油漆,他的老伴儿陪在一旁给打扇子, and Φ真是从心里羡慕(c.f. Liu, 2001). Chen (1987) believed that the semantic structure of the sentence has two characteristics: on one hand, the parenthetical sentence is subordinate to either the sentence where the antecedent occurs or that where the anaphora occurs. In our example, "李贵在刷油漆,他的老伴儿陪在一旁给打扇子" was the object clause of "瞧见," which was lower in the level than the main sentence where the antecedent and the anaphoric object occurred. On the other hand, the grammatical structure of the inserted sentence sometimes produced ambiguity in meaning, so it was important to avoid being complicated and redundant in using zero anaphora. As it was indicated in another example, "张宁匆匆忙忙赶到火车站, Φ1庆幸总算赶上火车了, Φ1把车票递给正在检票的列车员, Φ2说票是假的, and Φ1一时觉得有点不知所措," the meaning where "Φ 2" located was ambiguous because either "列车员" or "张宁" grammatically fit into the sentence. Thus, zero anaphora here was not acceptable unless we changed the sentence into "……列车员说票是假的, 张宁一时觉得有点不知所措."

Experiment 2

Previous studies have shown that the resolution of pronoun anaphora is heuristic, showing that the activation of searching for the antecedent begins as soon as the pronoun occurs. However, the resolution of zero anaphora does not possibly fit into the situation. Instead of being introduced by pronouns, there is a lack of the anaphora in zero anaphora, which increases the difficulty of referring back to the antecedent, particularly when the candidate of the antecedent is more than one. Examining the following example "小吴不但接受了李明的帮助, 而且Φ立即表示愿意改正错误(zero anaphora)," we infer that if people begin to search for the antecedent immediately after the verb "表示" shows, they probably resolve the anaphora in a heuristic way. Otherwise, they do in an exhaustive way if the search begins after the last character (e.g., "错误") is given. Consequently, the probe word after the verb "表示" should receive the highest level of activation if people identify the antecedent in a heuristic way and the probe word 错误receives the highest level of activation if the resolution is carried out in an exhaustive way.

Experiment 2a: The Time Course of Zero Anaphora Resolution Under the Condition of Topic Continuity

Subjects

Thirty-two undergraduates with normal or corrected-to-normal vision took part in this experiment. They were all Chinese native speakers.

Materials

The experimental materials were composed of 24 sentences, each of which contained 3 clauses, such as"\终于\爬上\了\山顶, Φ眺望\着\远处\高楼林立\的\城市, Φ心想★[1] 究竟 哪里 才是 自己 的家 啊 ★[2]"The probe words were either true antecedent (e.g., 吴杰) or false antecedent (e.g., 张帆), being inserted either in the position marked ★[1] or ★[2]. The length of the sentence, the word frequency and character strokes, and the familiarity of the name were properly controlled.

Design and Procedure

This experiment was designed as 2 (probe word: true/false antecedent) × 2 (position: ★[1]/★[2]. We examined the period of time consumed by participants in verifying the probe word 250 ms after the characters (e.g., 心想/家啊) were presented on the screen. This experiment was implemented by using E-prime software. In the beginning, "+" was shown on the screen for 600 ms, followed by presenting the materials part by part (each part was identified by "/"). The display time for each part was 300 ms. All the presented parts of the first clause "终于爬上了山顶" disappeared together 300 ms after the last part of this clause "山顶" were given. The next clause followed the same presentation mode. The interval between the disappearance of the verb "心想," and the occurrence of the probe word was 250 ms. Participants were given 3 s to make judgment whether the word occurred or not. As soon as the decision was made, the rest part continued to appear until the last part disappeared. If participants made wrong judgments, a red sign was presented on the screen. Six hundred millisecond later, they had to answer a question to test their comprehension ability. The wrong answer received feedback. A reading comprehension question was attached to each trial to ensure that participants read carefully. The procedure needed practicing beforehand.

Results

We deleted the data that the response time was beyond standard deviations and listed the rest in **Table 5**. We carried out an analysis of ANOVA on the response time. The results showed that the position had no main effect [$F_{(1,76)} = 0.629$, $p > 0.05$] but the probe word type produced the main effect [$F(1,76)=17.402$, $p<.001$]. The two variables had the interaction effect [$F_{(1,76)} = 3.902$, $p = 0.052$ (marginal significance)]. Further *t*-test showed that when the probe word was the false antecedent, the position effect was not significant [$t_{(76)} = 0.525$, $p > 0.05$] whereas when the probe word was the true one, the position effect was significant. The response time was faster if the probe word occurred after the sentence than they did after the verb "心想." Meanwhile, we found that under the condition of topic continuity, the searching for the antecedent did not start until the last character of the essay disappeared. It meant that zero anaphora resolution was possibly carried out in an exhaustive way. We also found that the first-mention effect played a dominant role in zero anaphora resolution.

Experiment 2b: The Time Course of Zero Anaphora Resolution Under the Condition of Topic Discontinuity

Compared with the situation in which zero anaphora resolution was used when the topic was maintained continuously from sentence to sentence within one essay, the topic discontinuity or new topic insertion prohibited the resolution of Chinese zero anaphora. On one hand, the appearance of more than one

TABLE 5 | Average reaction time to probe words in the position either after the verb or at the end of the sentence under the condition of topic continuity in Experiment 2a (ms).

Probe word type	Probe word location	
	After the verb	At the end of the sentence
True antecedent	1157.65	1088.56
False antecedent	1232.11	1252.00

TABLE 6 | Average reaction time to probe words in three positions under the condition of topic discontinuity in Experiment 2b (ms).

Probe word type	Probe word location		
	Position 1	Position 2	Position 3
True antecedent	1032.00	1171.63	1295.73
False antecedent	1154.83	964.60	1125.65

antecedent in front of where zero anaphora occurred usually extracted attention. On the other hand, in order to identify the different entities that were mentioned within two or more zero anaphora, people had to consume the increased cognitive efforts. Previous linguistic studies have given a lot of attention to zero anaphora under the condition of topic continuity, we intended to conduct a psycholinguistic experiment to investigate how people resolve zero anaphora when the topic was discontinuous. Since it was one of the possibilities that the antecedent which zero anaphora referred to belonged to either the first mention or the second mention, we preferred to examine whether there was a difference in response time when the probe word was located in different anaphora positions ($\bigstar^1/\bigstar^2$).

Subjects

Another fifty-two undergraduates with normal or correct-to-normal vision took part in the exam.

Materials

The experiment materials were composed of 24 sentences with each sentence containing 5 clauses. This was the example: "王锋\突然\听到\一声巨响\, \Φ1迅速\冲出\潜伏的\猫耳洞, Φ1\看到\战友\倒在\前面的\山坡\上, Φ2 \浑身是血\艰难地\挪动$\bigstar^1$\着, \Φ1便\毫不犹豫地\冲$\bigstar^2$\了\上去。\$\bigstar^3$," The true antecedents were "王锋" and "战友," and the false one was "李平." The "★" stood for the position where the probe word was inserted. We had matched the length of the sentence, the character frequency and strokes, and the familiarity of the names.

Design and Procedure

This experiment was designed as 2 (probe word: true/false antecedent) × 3 (position: $\bigstar^1/\bigstar^2/\bigstar^3$). We examined the period of time consumed by participants in verifying the probe word 250ms after the characters (e.g., 挪 动/冲/上 去) were presented on the screen. This experiment was implemented by using E-prime software. The procedure was the same as for Experiment 2a except there were three probe words to be judged on their concurrence after "挪 动," "冲," and "上 去" disappeared respectively.

Results

The data beyond 3 standard deviations of the reaction time were deleted (see **Table 6**). The results of variance analysis of the reaction time between groups were analyzed, showing that the interaction effect was significant [$F_{(2,\ 102)} = 10.213$, $p < 0.001$]. Then, the simple effect analysis between position and probe word type was carried out to investigate the position effect produced both by true antecedent and false antecedent. The results showed that when the antecedent wase true, the position effect was significant with $F_{(2,\ 102)} = 3.884$, $p < 0.05$. The t-test further showed that there was a significant difference between $\bigstar^1$ and $\bigstar^2$ with $t_{(51)} = -4.553$, $p < 0.001$. There was also a significant difference between $\bigstar^1$ and $\bigstar^3$. The response time was longer when the probe word appeared at the end of the sentence with $t_{(51)} = -2.321$, $p < 0.05$. However, there was no difference between $\bigstar^2$ and $\bigstar^3$ with $t_{(51)} = -0.528$, $p > 0.05$, indicating that the activation level of probe words in $\bigstar^1$ and $\bigstar^2$ was higher, so people could access the antecedent much easier.

Discussion of the Experiment 2

The results of experiment 2a showed that if the topic was continuous, the activation was at the highest level when the probe word appeared at the end of the sentence. In this situation, zero anaphora resolution was carried out in an exhaustive way. This was consistent with what McDonald and MacWhinney (1995) had proved that the resolution began 250ms rather than immediately after the pronoun appears. Thus, the activation was affected by the position where the probe word was inserted. Meanwhile, the factor of topic continuity was quite important to the anaphoric resolution (e.g., Ariel, 1990, 1994). "When it comes to pronouns, if there is only one entity in the discourse in the focus of attention, processing for pronouns is basically not needed, or the process is automatic" (Greene et al., 1992). The expression "one entity" referred to the topic continuity. It might cause a controversial argument that the topic continuity did not play any significant role in pronoun and zero anaphora resolution in Experiment 1. Why was the case different in Experiment 2? In fact, Experiment 1 and Experiment 2 were designed to explore different aspects of anaphora resolution. Experiment 1 was conducted to require participants to read a sentence and finish a verification judgment after reading. Participants were not interrupted during the time period and they could skip the gap if the topic was not continuous. But the presentation of materials in Experiment 2 was interrupted by the insertion of probe words and the verification task, which forced participants to think about the early information constantly. Therefore, the inconsistent information was very sensitive to them and, therefore, influenced them during the processing. Thus, the topic continuity took different effect. According to Wang and Wu (2020), the change of topic is proven to increase the processing difficulty at some processing stages, but things might be different if anaphora is employed.

Concerning the topic inconsistency, the resolution began when the probe word was present right after the verb. This heuristic way of searching for the antecedent fitted into the structure building model (Gernsbacher, 1990), according to which, the first step to construct comprehension of units larger than a word or image was providing a foundation on which the relations among subsequent events could be laid. Compared with comprehending the topic-maintained discourse, readers had to construct the upcoming information to the current discourse representation and comprehend topic-shifted discourse through building a new substructure for a new topic. People were forced to take different reading strategies in order to achieve a coherent and well-organized representation of the discourse. For Chinese native speakers, as suggested by previous studies, they were used to adopt various reading strategies to cope with the divergent situations: the exhaustive searching strategy for the topic-consistent discourse and the heuristic one for the topic-shifted discourse.

Finally, the insignificant activation of the probe word when they appeared immediately after the verb, which we assumed, did not mean that the resolution was totally inhibited, it was just delayed. According to the investigation of the pronoun anaphora, this tardive resolution was often observed after an interval between the activation of pronoun and the completion of resolution (Sanford and Garrod, 1989; Cristea and Dima, 2001). This was also true to zero anaphora resolution.

Experiment 3: A Preliminary Study of the Factors That Are Related to the Resolution of Zero Anaphora

Summarizing from previous psycholinguistic experimental studies, we concluded the factors influencing zero anaphora resolution into the aspects such as the topic, the referential distance within one sentence, the verb causality, the voice type (the active/passive voice), and the context where zero anaphora is. We have discussed the first two aspects in the Experiment 1a and 1b. Worth mentioning that, not only the linear distance should be considered as a factor of influencing anaphora resolution, but also the psychological distance, the mental gap which is produced when people are trying to refer back to the antecedent from the position where anaphora appears, should also be taken into account. The linear distance is commonly governed by the grammar rules and each word belongs to a designated position, so that people are not supposed to be influenced by the syntactic structure of linear distance. However, the psychological distance largely varies from people to people in terms of world knowledge and the semantic competence they have. Thus, we are going to discuss the factors related to psychological distance and expected that: (1) if sentences are expressed in an active voice (the first mention priority) in the situation in which the information keeps consistent in meaning within the discourse, the response time to the probe words is the shortest. (2) If the sentences are expressed in the passive voice (non-first mention priority) in the situation in which the information is inconsistent within the discourse, the response time to the probe words is the longest, since there is neither the consistent given information nor the existed first mention priority, people need the largest number of energies to process the new information. Lastly, (3) if the sentences are expressed in the passive voice in the situation in which the information is consistent, or the sentences are expressed in the active voice, but the surrounding information is inconsistent, the reaction time to the probe word is in between.

Subjects

The participants were 80 undergraduates with normal or correct-to-normal vision.

Materials

The materials were composed of 30 groups of sentences with each group containing 4 different versions of sentences (a total of 120 sentences). Four versions of sentences were exemplified below and were arranged in accordance with the order: (1) the situation described by the first two clauses in the active voice was consistent (the first mentioned priority, e.g., 在课堂上, 老师狠狠批评了班上最调皮的学生, Φ警告他不要再犯错误); (2) the situation described by the first two clauses in the passive voice was consistent (e.g., 在课堂上, 班上最调皮的学生被老师狠狠批评, Φ警告他不要再犯错误); (3) the situation described by the first two clauses in the active voice was inconsistent (the first mentioned priority, e.g., 在医院, 老师狠狠批评了班上最调皮的学生, Φ警告他不要再犯错误); and (4) the situation described by the first two clauses in the passive voice was inconsistent (e.g., 在医院, 班上最调皮的学生被老师狠狠批评, Φ警告他不要再犯错误).

Design and Procedure

The experiment was designed as 2 × 2 × 3. The first 2 stood for the voice type (the active voice/the passive voice), the second 2 stood for the situation described by the first two clauses (consistent/inconsistent), and the last 3 stood for the types of probe words (true antecedent with the first mention/true antecedent with the second mention/false antecedent). The dependent variable was the time intervals between the presentation of the probe word and the onset of making a judgment on them. All characters in materials were the most commonly used Chinese characters.

E-prime software was used to implement the experiment. Prior to the experiment, participants were invited to practice three groups of sentences. At the beginning of the experiment, the "+" was displayed on the screen for 600 ms, after that, phrases such as "在课堂上" were presented on the screen character by character, 300 ms for each character. Five hundred milliseconds after the phrase "在课堂上" disappeared, the next part was presented in the same way and stayed on the screen for the same period of time. When the last character of the sentence disappeared, the probe words were shown to participants for 500ms. They were asked to make judgments as accurately and quickly as possible on whether the probe words had been presented or not. If yes, participants pressed the "Y" key, otherwise pressed the "N" key.

TABLE 7 | Average reaction time to probe words in situations in which the consistent/ inconsistent situation is described in active/ passive voice in Experiment 3 (ms).

Probe words	Active		Passive	
	consistent	inconsistent	consistent	inconsistent
True antecedent (first-mention)	573	649	615	687
True antecedent (second-mention)	634	694	657	743
False antecedent	714	749	650	735

Results

Table 7 showed the average reaction time to probe words in situations in which the consistent or inconsistent situation was described in active/passive voice in Experiment 3. First of all, we deleted the data of which the response time were beyond standard deviations and conducted ANOVA on the results of response time between groups. The results showed that the interaction effect was not significant [$F_{(1,3)} = 0.329$, $p > 0.05$]. There was a major effect in groups [$F_{(1,3)} = 3.447$, $p < 0.05$], indicating that the response latency of the four groups was significantly different. ANOVA of 2 (voice type: active/passive) × 2 (situation in active voice: consistent/inconsistent) × 2 (probe word type: true antecedent with first-mention/second-mention) was conducted on the experimental data, and the false antecedent was not considered here to further investigate whether there was a difference between active and passive voice and whether there was a significant difference between the consistency and inconsistency of previous information situations.

It was found that all the interactions were not significant, and the interaction between probe word (true antecedent with first-mention/second-mention) and situation (consistent/inconsistent) was not significant [$F_{(1,79)} = 0.055$, $p > 0.05$]. The interaction between probe word and voice type was not significant [$F_{(1,79)} = 0.012$, $p > 0.05$]. The interaction between voice type and situation was not significant [$F_{(1,79)} = 0.854$, $p > 0.05$]. The interaction among three variables (probe word type, situation, and voice type) was also insignificant [$F_{(1,79)} = 0.28$, $p > 0.05$]. However, the main effect of probe word type was significant [$F_{(1,79)} = 39.938$, $p < 0.05$]. The response of subjects to true antecedent with first mention was faster. Compared with the passive voice, people reacted to sentences with active voice significantly faster with $F_{(1,79)} = 5.656$, $p < 0.05$. The main effect of the information situation was not significant with $F_{(1,79)} = 2.388$, $p > 0.05$.

Discussion

What the results have revealed were highly in line with our expectations. If, as the results suggested, the probe words were true antecedents like "老师" (B1) and "学生" (B2), the reaction time under four conditions was systematically different: the shortest reaction time (MB1 = 573 ms, MB2 = 634 ms) and the longest reaction time (MB1 = 687 ms, MB2 = 743 ms) were produced, respectively, when the active voice (the first mention priority) was used in the consistent situation and when the passive voice was used in the inconsistent situation. At the same time, the results under the other two conditions were in between. However, the differences mentioned above were not found when control words were introduced. The effect of voice type was significant, which indicated that the first mention effect (Gernsbacher, 1990) existed under this condition. The effect was also confirmed by Gernsbacher and Hargreaves (1988), in which the advantage of the first mention was found during the processing of the active vs. passive construction in English, despite that the materials did not contain any form of anaphora. It showed that the advantage of first-mention effect is not attributable to the existence of anaphora but the voice type of the language, since it matters that "the foundational role of the earliest participants maintained higher levels of activation than other participants in the same sentence" (MacDonald and MacWhinney, 1990).

The discrepancy in the reaction time was related to the grammatical function the antecedent and zero anaphora played in the whole sentence, i.e., the parallel function. In our experiment, the first type of sentence was processed faster than the second type of sentence, because the antecedent and zero anaphora were the parallel subjects in each clause of the sentence. Similarly, the third type was faster than the fourth type at the processing speed. This was also consistent with Sheldon (1974) and Gernsbacher and Foertsch (1999). Nevertheless, Miao (1996a) found that the parallel function took effect only when the sentence meaning was correct. In another word, if the sentence meaning was correct, the parallel function would not disappear, or vice versa. We cannot examine it very thoroughly, because all materials in our experiment were semantically correct even though the situation they constructed were either consistent or inconsistent. Contrary to our expectations, the effect of situation consistency was insignificant. The first reason, we thought, originated from the failure of the materials in supplying sufficient background information for participants to get ready to refer back to the antecedents and later to resolve the zero anaphora. We carried out the experiment by adopting the research paradigm that Sanford and Garrod (1981) had used, but we revised the words into clauses in our experiment. Theoretically speaking, clauses contained more information than words, but they were not sufficient in assisting participants to identify zero anaphora and refer back to antecedents very well. In other words, the insignificant effect indicated that the resolution of zero anaphora was quite a difficult mental processing. Wang and Yang (2004) admitted that there was a lack of systematic study on the factors related to the underlying psychological mechanism during the pronoun anaphora resolution. This was also true to zero anaphora resolution: the divergent opinions have been formed on the basis of rare but scattered studies and the unsystematic findings of the related factors always vary from one research to another. Compared with pronoun anaphora resolution, zero anaphora resolution encountered more difficulties and required more attention.

GENERAL DISCUSSION

General Discussion of Experiment Materials

The purpose of the current study is to do the preliminary research to explore the psychological reality of Chinese zero anaphora as a basis for further measuring the time course of resolving sentences containing different types of anaphora, and to make a tentative attempt at determining the factors during this process. The hypotheses we have proposed were partially confirmed: we observed the insignificant effect of the reaction time among all the experimental conditions in Experiment 1, which is consistent with our expectation. The effect of reading time is also insignificant, which is beyond our expectations. The results are possibly due to the indiscriminate activation of probe words without being affected by the anaphora type, that is, zero anaphora resolution is performed as fast as pronoun anaphora resolution. For most Chinese native speakers, the language they speak is characterized by a lot of zero anaphora, which forces the language users to build some corresponding processing models in their minds. Even though it is widely accepted that zero anaphora is an intricate language phenomenon involving mental mechanisms such as short-term memory and psychological distance, Chinese native speakers often successfully overcome the difficulty and refer back to the antecedents very quickly. This is, on another side, reflects how zero anaphora is psychologically realized. What's more, with the finding of the first-mention advantage, the implicit causality effect with subject-biasing than object-biasing verbs is easier to detect (Stewart and Gosselin, 2000). This is consistent with our study. The subject and the first-mentioned entity is processed in privilege because it is mentioned first, so that it is impressed a lot in the minds of people and is easily recognized later.

When we compiled the materials of Experiment 1b, we'd like to insert a changed topic into the material to make the topic discontinuous. The purpose is to distract the attention of participants by introducing another antecedent. Generally speaking, the insertion of a new topic might cause the cognitive competition during the resolution in the brain. When people read the sentence, they usually maintain the first-mention information in their mind, i.e., the information related to the first-mentioned topic and prepare to process it. However, the insertion of a new topic results in the interference of the old information and the mixture of information, including the old information, the new information and the intertwining of them, rushes into the mind so that the complexity of resolution is largely increased. The old information plays a role of constructing a primary concept, which needs constant adjusting in order to coincide with an emerged concept constructed by new information. Therefore, this increased processing procedure delayed the access to the antecedents even they are the true ones.

It is, however, easily challenged by people that the materials we used in our experiments are compiled not as naturally as we read in literature and, at the same time, are lacking of ecological validity since the exclusive adoption of only one narrative style without considering other styles. However, in order to obtain the objective results in psychological empirical studies, we have to narrow down the possibilities by choosing one type of material and deal with them in accordance with the research purpose. Making up the experiment materials is also necessary because the unrelated variables can be controlled as strictly as possible with this method. Although the factitious materials are considered less authentic particularly in terms of expressing meaning, they can meet the requirements that experiments demand in aspects such as word frequency, word length, and syntactic structure. From this perspective, compiling materials can be said critical to the experiment. This is also the way people follow in their experiments.

Experiment 2 investigates the time course of Chinese anaphora resolution. Our expectations partially came true, but under different conditions. When the topic was kept consistent, the searching follows the exhaustive way while when the topic was not kept consistent, the searching follows the heuristic way. We assume that the selection of different searching strategies originates from the capacity of short-term memory. Different from pronoun anaphora resolution, which is obviously signified by the appearance of pronoun, the relationship between the antecedent and its anaphora must be identified through going over the given information back and forth within one sentence or even among sentences. It increases the load of the processing capacity of short-term memory and results in two possibilities: as the topic keeps consistent within the given discourse, the information leads the readers to the end of the discourse without being interrupted by the inserted topic. The exhaustive way of searching for the antecedent does not need such extra energy that short-term memory can fulfill the task. Another possibility is produced when the information inconsistency forces the readers to make sure the anaphora relationship from one place to another until the correct one is identified, thus, most of the time, they use the heuristic way of searching because of the limitation of short-term memory capacity. The insertion of new information breaks down the fluency of reading. Consequently, relatively more cognitive energy is consumed.

The third experiment intends to determine whether the two factors produce the influence on the resolution of zero anaphora on the basis that the resolution has been found psychologically resembled in mind with the way of searching information either heuristically or exhaustively. Out of our expectation, unlike the variable of the voice type, situation consistency does not exert any significant difference on the resolution process. It is not very clear how to interpret the indiscriminate between the consistent and inconsistent situations that the zero anaphora appears. It may be, as we claim, that the materials we used in the experiment fail to supply the sufficient background information in facilitating the participants to form the described situation in mind and identify the referent according to the description. If the situations are not described very distinctively, in other words, if participants are not able to tell the difference between the consistent situation and the inconsistent situation, the insignificant difference is supposed to be observed during the process of making judgments. Concerning the voice type, the active voice and the passive voice, people find it easy to tell the difference just because, according to the Chinese grammar, a defining syntactic indicator "被" is required to occur, which

helps people distinguish the voice type even without considering the meaning of the sentences. Consequently, the effect due to the voice type divergence is significantly different, but the situation consistency is not.

Surprisingly enough, by examining the experiment materials further, we discover that when the undergraduates were invited to compile the materials, they preferred to use the verbs like "看(see)," "发现(find)," "遇见(meet)," and "留意到(notice)" as the second antecedents. In fact, these verbs are closely related to the vision and sensation of human beings. In Chinese, this is an effective and a common way used to transfer the topic (e.g., 我看见他来了), because Chinese grammar is featured with paradoxical linkage (Song, 2003) with which the syntactic structure is constructed in the relationship of coordination rather than subordination in English. When we compiled the experiment material, we need to consider the language application habits of using synchronic verbs to show the transition from one topic to another in the described situation. Thus, the verbs mentioned above are chosen, which, in turn, has become one of the factors influencing zero anaphora resolution.

With the analysis of materials from another perspective, the position where zero anaphora should have appeared is still a controversial issue. The experiment materials such as type C in Experiment 1a and type B in Experiment 1b were evaluated by the undergraduates majoring in Chinese Literature with Likert scale to confirm whether zero anaphora appears in the right place. For example, in the sentence "中午李明到了学校, 接着Φ (她)开始上课, 课后Φ又到同学那里串门, Φ坐到快天黑时, Φ (她) 才恋恋不舍地告别回家了," either "Φ" or "她" fits into the place and delivers the same meaning in the position where "Φ (她)" occurs. It shows that zero anaphora and pronoun anaphora can be used alternatively without altering the meaning of the sentence. Gao (2003) investigated the discriminate usage of zero anaphora and pronoun anaphora in written and spoken discourses through pragmatic analysis, finding that the two types of anaphora are used indifferently even the two types of discourse are finished by the same author and in the same literature style. The phenomenon that an anaphoric expression is introduced either by pronoun or zero anaphora brings troubles due to the fact that "return of current discussion to a mention other than the linearly most recent one in the preceding discourse can be done by means of a pronoun or zero anaphora in many languages" (e.g., Huang, 1989, 1994). This is beyond the explanation offered by the recent theory.

In summary, as Greene et al. (1992) argued, the question about pronoun resolution may not be what the pronoun can do for the discourse, but what the discourse can do for the pronoun. As far as zero anaphora resolution is concerned, it is inferred that the question may not be what zero anaphora can do for the discourse, but what the discourse can do for zero anaphora. The anaphoric relationship in discourse is established by nouns, but maintained by pronouns and zero referents (Xiong, 2000).

General Discussion of the Models

There is a large number of anaphora resolution models, such as Kintsch and van Dijk (1978), Kintsch (1988), and Gernsbacher (1990), but they do not fully account for Chinese zero anaphora resolution. According to the theory of argument overlap (Kintsch, 1988), readers can associate information with previously encoded information by means of argument overlap, which is also considered as an essential way of maintaining referential coherence. However, the theory emphasizes that the overlapping of arguments is the premise of processing information smoothly, or else the lack of it among arguments results in the increased time during the discourse comprehension. The argument is easily found overlapped in the pronoun anaphora for the conspicuous existence of *he, she,* etc. Contrary to that, there are no tangible words in zero anaphora to supply the coherence to people. From this perspective, the argument overlap theory is not able to explain zero anaphora resolution during comprehension. Similarly, the accessibility of the antecedents in sentences containing anaphora expression, particularly the pronoun anaphora expression and the repeated nominal anaphora expression, is proposed to be modulated by the cognitive mechanisms, namely, suppression and enhancement: the presence of the anaphora activating the referents to some degree but prohibiting the non-referents. If, as the typical zero anaphora shows, the anaphora is absent, the underlying mechanism might not take effect during the resolution. Myers and O'Brien (1998) believed that three factors are related to the access to the antecedents in the anaphora expressions, but two of them, that are the special distance between the anaphora and the antecedents as well as the elaborate processing amount of the antecedents, are not clearly determined in resolving the zero anaphora.

The models are proposed to explain the pronoun anaphora resolution in English, but they are not powerful in revealing the underlying mechanism when zero anaphora resolution is used in Chinese. Language is a means of communication and is expected to be used to transfer information as accurately and efficiently as possible. Zero anaphora is commonly seen in Chinese and is welcomed by people for its efficiency and accuracy in delivering information. Language reflects the way of thinking and, in turn, is significantly affected by it, so people who speak Chinese have been used to thinking in the way that the language determines and have developed the strategies to deal with the expressions containing zero anaphora. This was proven by Tao and Healy (2005) in their experiment, in which Chinese native speakers were found to show the advantage over English native speakers in coping with zero anaphora even when the materials were presented in English, indicating that Chinese native speakers have transferred such a strategy from one language to another.

What is more, compared with ideal zero anaphora in our experiment, the practical use of zero anaphora is quite complicated. Zero anaphora is just a composition of the "anaphora chain," reflecting the partial rather than the overall meaning that the discourse intends to express. Like the hoops buckling into a chain, each anaphora contains one idea, and the scattering ideas are organized by the chain to produce a comprehensive one and form the coherence of the discourse at the same time. Li and Thompson (1979) use "topic chain" to describe this language phenomenon, pointing out that the topic established in the first clause serves as the referent for the

unrealized topics in the following clauses. Tsao (1990) claims that "a sentence in Chinese can be roughly defined as a topic chain, which, is a stretch of discourse composed of one or more comment clauses sharing a common topic, which heads the chain." Chen (1987) claims that zero anaphora or pronoun anaphora is responsible for encoding the referents that possess high topic continuity in Chinese. However, the question still exists, namely, to what extent and in what aspects the topic chain facilitates the encoding of zero anaphora resolution.

In general, anaphora resolution has received the close attention from disciplines such as psycholinguistics, Chinese linguistics, computational linguistics, pragmatics, and philosophy. As we have mentioned, zero anaphora is more commonly seen in Chinese rather than in Indo-European languages. Studying zero anaphora resolution in depth can enrich our linguistic knowledge about Chinese and supply enlightenment to understand the philosophy among languages and the way of thinking and cultures. At the same time, it helps solve some key problems in machine translation and computer processing of natural language.

It is important to remember the limitations of the current study. First, we intended to test our hypothesis by employing the real-time processing methods, but as the results suggested, not all the hypotheses have been empirically tested. It is for sure that the real-time measurement is a way to reveal the underlying mechanism above which zero anaphora resolution is performed, but it is unable to tell the whole story of reading comprehension. For example, priming and detection technology are specialized at activating, representing, and organizing information, but may disrupt the reading fluency (Tanenhaus, 2004). Considering that anaphora resolution occurs immediately after the trigger is given, the results are less reliable if the detection words are introduced between the antecedents and the zero anaphora. Thus, if possible, more on-line research methods such as ERPs and fMRI should be considered as an alternative way to find out more subtle differences that are not significantly resembled in the present study. Second, the insignificant differences may be due to the design of the experiment. Specifically, although zero anaphora resolution depends on grammatical and syntactical knowledge, it requires background knowledge at the same time. For example, the two sentences "老张生了个儿子, Φ天天哭闹," and "老张生了个儿子, Φ天天炫耀" are constructed identically in grammar and syntactic structure, the antecedents that zero anaphora refers back are quite different with "儿子" in the first sentence and "老张" in the second. It is conceivable that people with background knowledge may indeed exert a great influence on the resolution of zero anaphora in very different ways. Third, further thinking might concern but might not be restricted to the following aspects: (a) The pictographic form of Chinese characters possibly contains more "information load" than Latin words, which is assumed to be highly related to identify zero anaphora and facilitate its resolution. (b) Chinese native speakers usually think in a holistic and systematic way, which potentially cultivates their ability of grasping the main idea even if the sentences are incomplete in grammar. As for zero anaphora, the lack of referent in certain places does not stop them from comprehending the meaning of the sentence. According to Dopkins et al. (1992), anaphora resolution, in most cases, is either a bottom-up or top-down procedure, we assume that Chinese native speakers tend to construct a situational model in discourse comprehension. As a result, zero anaphora resolution is a top-down procedure. (c) The adding of some Chinese adverbs in experiment materials such as "也(also)" "只好(have to)" in the sentences improves the logical relationship among clauses where zero anaphora is contained. We just wonder whether the fluent resolution of zero anaphora is also determined using adverbs. If so, in what way does the use of adverbs affect the resolution.

CONCLUSION

Initiated by the idea that Chinese zero anaphora resolution might involve a unique way of processing information that is quite different from the way of processing other types of anaphora, we carried out five experiments to test three hypotheses from the perspectives of the psychological reality of the resolution, the time courses of the resolution and the factors related to the resolution. Based on the results, we conclude that the first-mention effect exists during zero anaphora resolution, which follows the same processing mechanism with pronoun resolution. During the resolution, the exhaustive way of searching for the antecedents within one sentence happens when the topic is in consistency, but the heuristic way of searching is carried out when the topic is out of consistency. Meanwhile, the first-mention effect is also found when the discourse is organized in the active voice no matter the information within the discourse is consistent or not.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Review Board of South China Normal University. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

NY and ZL designed the study. NY, JZ, and LM acquired the data. NY and LM analyzed the data. JZ wrote the manuscript. ZL polished it. All authors contributed equally to reviewing the manuscript.

FUNDING

This work was supported by the Key Program of National Fund of Philosophy and Social Science of China (19AYY014), Major Program of Guangdong Provincial Fund of Philosophy and

Social Sciences (GD20ZDWY02), Bidding Project of 2021 of the Key Lab for Social Sciences of Guangdong Province, Bilingual Cognition and Development Lab (BCD202103), Project of 13th Five-Year Plan of Philosophy and Social Science Development in Guangzhou (2019GZYB55 and 2020GZGJ277), Project of 13th Five-Year Plan of Philosophy and Social Science Development in Guangdong (GD20XWY05 and GD20XXL01). Major Cultivating Project of Guangdong University of Foreign Studies (20ZD01); Distinguished Innovation Project of Guangdong University of Foreign Studies (19SS03).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.663168/full#supplementary-material

REFERENCES

Anderson, A. G., and Sanford, A. J. (1983). The accessibility of pronominal antecedents as a function of episode shifts in narrative text. *Q. J. Exp. Psychol.* 35, 427–440. doi: 10.1080/14640748308402480

Anderson, J. (1983). *The Architecture of Cognition*. Cambridge, MA: Harvard University Press.

Anderson, J. (1985). *Cognitive Psychology and Its Implications. 2nd edn*. New York, NY: Freeman.

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

Ariel, M. (1994). Interpreting anaphoric expressions: a cognitive versus a pragmatic approach. *J. Linguist.* 1, 3–42. doi: 10.1017/S0022226700016170

Bai, X.-J., Zhang, X.-L., and Yan, G.-L. (2005). Eye movement research on the time course of verb's implicit causality in pronoun resolution. *Psychol. Explor.* 3, 19–23. doi: 10.3969/j.issn.1003-5184.2005.03.005

Carroll, D. W. (1999). *Psychology of Language, 3rd Edn.* Pacific Grove, CA: Brooks Cole Publishing Company, 129–153.

Chen, P. (1987). Discourse analysis of Chinese zero anaphora. *Stud. Chin. Lang.* 5, 363–378.

Corbett, A. T., and Chang,. F. R. (1983). Pronoun disambiguation: accessing potential antecedents. *Mem. Cogn.* 11, 283–294. doi: 10.3758/BF03196975

Cristea, D., and Dima, G. E. (2001). An integrating framework for anaphora resolution. *Inform. Sci. and Technol.* Romanian Academy Publishing House, Bucharest, 4:273–291. Available online at: https://profs.info.uaic.ro/~dcristea/papers/AcademieCristea&Dima.pdf

Crystal, D. (2008). *A Dictionary of Linguistics and Phonetics*. Chichester; West Sussex: Wiley-Blackwell. doi: 10.1002/9781444302776

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

Dopkins, S., Morris, R. K., and Rayner, K. (1992). Lexical ambiguity and eye fixations in reading: a test of competing models of lexical ambiguity resolution. J. Mem. Lang. 31:461–476.

Gao, L., Liu, Z., and Huang, Y. (2005). Who is *ziji*? An experimental research on binding principle. *Ling. Sci.* 2,39–50. doi: 10.1037/a0015251

Gao, Y. (2003). *A Study of Anaphora From the Perspective of Cognition*. Beijing: Foreign Language Teaching and Research Press.

Gernsbacher, M. A. (1989). Mechanism that improve referential access. *Cognition* 32, 99–156. doi: 10.1016/0010-0277(89)90001-2

Gernsbacher, M. A. (1990). *Language Comprehension as Structure Building.* Hillsdale, NJ: Erlbaum. doi: 10.21236/ADA221854

Gernsbacher, M. A., and Foertsch, J. (1999). "Three models of discourse comprehension," in *Human Language Processing*, eds S. Garrod, M. J. Pickering (East Sussex: Psychology Press), 283–299.

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., Hargreaves, D. J., and Beeman, M. (1989). Building and accessing clausal representations: the advantage of first mention versus the advantage of clause recency. *J. Mem. Lang.* 28, 735–755. doi: 10.1016/0749-596X(89)90006-5

Givón, T. (1979). "From discourse to syntax: grammar as a processing strategy," in *Syntax and Semantics: Discourse and Syntax*, ed T. Givón (New York, NY: Academic Press), 81–114. doi: 10.1163/9789004368897_005

Givón, T. (1983). "Topic continuity in discourse: a quantitative cross-language study," in *Topic Continuity in Discourse: An Introduction,* ed T. Givón (Amsterdam: John Benjamins). doi: 10.1075/tsl.3

Givón, T. (1995). *Functionalism and Grammar*. Amsterdam: John Benjamins. doi: 10.1075/z.74

Greene, S. B., McKoon, G., and Ratliff, R. (1992). Pronoun resolution and discourse models. *J. Exp. Psychol. Learn. Mem. Cogn.* 18, 266–283. doi: 10.1037/0278-7393.18.2.266

Gundel, J., Nancy, H., and Ron, Z. (1993). Cognitive status and the form of referring expressions in discourse. *Language* 69, 274–307. doi: 10.2307/416535

Hinds, J. (1978). *Anaphora in Discourse*. Edmonton, AB: Linguistic Research Inc.

Hinds, J. (1979). "Organizational patterns in discourse," in *Syntax and Semantics: Discourse and Syntax*, ed T. Givón (New York, NY: Academic Press), 135–157. doi: 10.1163/9789004368897_007

Hirst, G. J. (1981). Discourse-oriented anaphora resolution in natural language understanding: a review. *Am. J. Comput. Linguist.* 7, 85–98.

Hou, M., and Sun, J.-J. (2005). Zero anaphora in Chinese and how to process it in Chinese-English MT. *J. Chin. Inform. Proc.* 1, 14–20. doi: 10.3969/j.issn.1003-0077.2005.01.003

Huang, C.-T. J. (1982). Logical relations in Chinese and the theory of grammar (Ph.D. thesis). MIT, Cambridge, MA, United States.

Huang, C.-T. J. (2001). *Logical Relations in Chinese and the Theory of Grammar*. (Ph.D. thesis), MIT, Cambridge, MA (USA).

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

Huang, Y. (1989). *Anaphora in Chinese: Toward a pragmatic analysis*. (Ph.D. dissertation), Cambridge, United Kingdgom: University of Cambridge.

Huang, Y. (1994). *The Syntax and Pragmatic of Anaphora: A Study With Special Reference to Chinese.* Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511554292

Huang, Y. (2000). Discourse anaphora: four theoretical models. *J. Pragmat.* 32, 151–176. doi: 10.1016/S0378-2166(99)00041-7

Jiao, J.-T., and Zhang, B.-Y. (2005). The effect of the implicit causality of Chinese verbs on pronoun resolution. *J. Psychol. Sci.* 5, 1082–1085. doi: 10.3969/j.issn.1671-6981.2005.05.014

Kibrik, A. A. (1996). "Anaphora in Russian narrative prose: a cognitive calculative account," in *Studies in Anaphora*, ed B. Fox (Amsterdam: John Benjiamins), 255–303. doi: 10.1075/tsl.33.09kib

Kim, Y.-J. (2000). Subject/object drop in the acquisition of Korean: a cross-linguistic comparison. *J. East Asian Ling.* 4, 325–351. doi: 10.1023/A:1008304903779

Kintsch, W. (1988). The role of knowledge in discourse comprehension: a constructive integration model. *Psychol. Rev.* 95, 163–182. doi: 10.1037/0033-295X.95.2.163

Kintsch, W., Patel, V., and Ericsson, K. A. (1999). The role of long-term working memory in text comprehension. *Psychol. Rev*. 42, 186–198.

Kintsch, W., and van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychol. Rev.* 85, 363–394. doi: 10.1037/0033-295X.85.5.363

Lambrecht, K. (1994). *Information Structure and Sentence Form: A Theory of Topic, Focus and the Mental Representations of Discourse Referents*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511620607

Li, C.-N., and Thompson, S. A. (1979). "Third-person pronouns and zero-anaphora in Chinese discourse," in *Syntax and Semantics: Discourse and Syntax*, ed T. Givón (New York, NY: Academic Press), 311–336. doi: 10.1163/9789004368897_014

Li, W.-D. (2004). Topic chains in Chinese discourse. *Discourse Process.* 37, 25–45. doi: 10.1207/s15326950dp3701_2

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

Liu, L.-H. (2001). Analysis of zero anaphora in Japanese. *J. Jpn. Lang. Study Res.* 4, 10–15. doi: 10.3969/j.issn.1002-4395.2001.04.003

Liu, Z. J. (2009). The cognitive process of Chinese reflexive processing. *J. Chin. Ling.* 37, 1–26.

Longacre, R. E. (1979). "The paragraph as a grammatical unit," in *Syntax and Semantics: Discourse and Syntax*, ed T. Givón (New York, NY: Academic Press), 115–134.

Lv, S.-X. (1984). *Thesis Collection on Chinese Grammar*. Beijing: The Commercial Press.

MacDonald, M. C., and MacWhinney, B. (1990). Measuring inhibition and facilitation from pronouns. *J. Mem. Lang.* 29, 469–492. doi: 10.1016/0749-596X(90)90067-A

McDonald, J. L., and MacWhinney, B. (1995). The time course of anaphora resolution: effects of implicit verb causality and gender. *J. Mem. Lang.* 34, 543–566. doi: 10.1006/jmla.1995.1025

Miao, X.-C. (1994). A study of pronoun processing in sentence and paragraph understanding. *J. Psychol. Sci.* 3, 159–163.

Miao, X.-C. (1996a). The influence of sentence semantics, the distance between pronouns and antecedents on pronoun processing. *J. Psychol. Sci.* 2, 71–78.

Miao, X.-C. (1996b). A study of semantic and grammatical factors influencing pronoun processing. *Acta Psychol. Sin.* 28, 352–357.

Miao, X.-C., and Song, Z.-G. (1995). The influence of word meaning and sentence grammar on pronoun processing. *J. Psychol. Sci.* 18, 197–200.

Mohanan, K. P. (1982). Grammatical relations and anaphora in Malayalam. *MIT Work. Pap. Ling.* 4, 159–163.

Myers, J. L., and O'Brien, E. J. (1998). Accessing the discourse representation during reading. *Discourse Process.* 26, 131–157. doi: 10.1080/01638539809545042

Sanford, A. J., and Garrod, S. C. (1981). *Understanding Written Language: Explorations of Comprehension beyond the Sentence*. Chichester: John Wiley and Sons.

Sanford, A. J., and Garrod, S. C. (1989). What, when and how: questions of immediacy in anaphoric reference resolution. *Lang. Cogn. Process.* 4, 3–4. doi: 10.1080/01690968908406369

Sheldon, A. (1974). The role of parallel function in the acquisition of relative clauses in English. *J. Verbal Learn. Verbal Behav.* 13, 272–281. doi: 10.1016/S0022-5371(74)80064-2

Shen, M., and Yang, Y.-F. (2006). The effects of implicit verb causality and accentuation on pronoun processing. *Acta Psychol. Sin.* 38, 497–450. doi: 10.1149/1.3610221

Shuai, L., Gong, T., and Wu, Y.-C. (2013). Who is who? Interpretation of multiple occurrences of the Chinese reflexive: evidence from real-time sentence processing. *PLoS ONE.* 8:e0073226. doi: 10.1371/journal.pone.0073226

Song, Z.-P. (2003). Aspects of parataxis vs. hypotaxis between English and Chinese. *J. Northeast Norm. Univer.* 92–98. doi: 10.3969/j.issn.1001-6201.2003.02.014

Stewart, A. J., and Gosselin, F. (2000). "A simple categorization model of anaphor resolution," in *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society*, eds L. R. Gleitman, A. K. Joshi (Hillsdale, NJ: Erlbaum), 930–935.

Sukthanker, R., Poria, S., Cambria, E., and Thirunavukarasu, R. (2020). Anaphora and coreference resolution: a review. *Inform. Fus.* 59, 139–162. doi: 10.1016/j.inffus.2020.01.010

Sun, Y., Shu, H., Zhou, X.-L., and Zheng, X.-J. (2001). The effect of implicit verb causality on pronoun processing. *J. Psychol. Sci.* 24, 39–41. doi: 10.3969/j.issn.1671-6981.2001.01.011

Tai, J. H. Y. (1978). "Anaphoric constraints in mandarin Chinese narrative discourse," in *Anaphora in Discourse*, ed J. Hinds (Edmonton, AB: Linguistic Research).

Tanenhaus, M. K. (2004). "On-line sentence processing: past, present, and future," in *On-Line Sentence Processing: ERPs, Eye Movements and Beyond*, eds M. Carreiras, C. Clifton. Jr (Brighton: Psychology Press).

Tao, L., and Healy, A. F. (2005). Zero anaphora: transfer of reference tracking strategies from Chinese to English. *J. Psycholinguist. Res.* 34, 99–131. doi: 10.1007/s10936-005-3634-5

Tomlin, R. S., and Pu, M.-M. (1991). The management of reference in Mandarin discourse. *Cogn. Linguist.* 2, 65–95. doi: 10.1515/cogl.1991.2.1.65

Tsao, F. F. (1990). *A Functional Study of Topic in Chinese: The First Step Towards Discourse*. Taipei: Student Book Co.

Wang, D., and Yang, Y.-F. (2004). Progress in pronominal expression in discourse. *J. Psychol. Sci.* 27, 1428–1431. doi: 10.3969/j.issn.1671-6981.2004.06.039

Wang, F., and Wu, F.-Y. (2020). Psycholinguistic studies on Chinese topic structure: a critical review. *For. Lang. Res.* 6, 9–16. doi: 10.16263/j.cnki.23-1071/h.2020.06.002

Wang, H.-B., and Li, B.-S. (1999). Analysis of the acquisition process of zero anaphora in Chinese. *Yantai Teach. Univer. J.* 2:64–68.

Wang, J.-L., and Stillings, J. (1984). "Chinese reflexives," in *Proceeding of the 1st Harbin Conference on Generative Grammar*, eds X. Y. Li, et al (Harbin: Heilongjiang University Press), 100–109.

Wang, S.-P., and Mo, L. (2001). Accessing of backgrounded information in discourse comprehension. *Acta Psychol. Sin.* 33, 312–319.

Wang, S.-P., Mo, L., and Xiao, X. (2001). The influence of features of antecedent information on its access during discourse comprehension. Acta Psychol. Sin. 33:509–517.

Wang, W.-J., Yan, G.-L., and Bai, X.-J. (2006). Antecedent frequency effects during processing of pronouns on eye movement in reading Chinese. *Psychol. Explorat.* 26, 30–34.

Xiong, X.-L. (1999). *A Comparative Study on the Anaphoric Phenomenon*. Shanghai: Fudan University Press.

Xiong, X.-L. (2000). Relevance in cognition, communication and logic. *Mod. Fore. Lang.* 23, 13–23.

Xu, K.-Y., and Xiao, X.-Q. (2008). Study on acquisition of anaphora in Chinese essays written by CFL learners. *Appl. Ling.* 4,118–125. doi: 10.16499/j.cnki.1003-5397.2008.04.013

Xu, Y.-L. (1992). *Introduction to Contrastive Linguistics*. Shanghai: Shanghai Foreign Language Education Press.

Yu, W. (1991). "Logophoricity in Chinese," in *The 3rd North-American Conference on Chinese Linguistics.* Ithaca, NY.

Zhang, W.-X., and Cui, X.-J. (2001). Rethinking the use of zero anaphora in spoken Chinese dialogue. *J. Tianjin For. Stud. Univer.* 3, 69–73. doi: 10.3969/j.issn.1008-665X.2001.03.017

Zhao, D.-M., and Liu, Z.-Y. (2006). The effected factors of anaphoric inference during reading process. *J. Psychol. Sci.* 29, 1144–1148. doi: 10.3969/j.issn.1671-6981.2006.05.028

Zhao, D.-M., and Mo, L. (2007). The integrative factors of anaphoric inference during reading process. *Acta Psychol. Sin.* 39, 225–234.

Zhou, X.-F. (2011). The acquisition of anaphora in narratives of English-speaking Chinese learners. *Chin. Teach. World* 25, 422–432. doi: 10.13724/j.cnki.ctiw.2011.03.015

Zhou, Y.-J., Chen, Y.-M., Cui, Y., Yang, L.-X., and Chen, X.-Z. (2001). The effects of anaphora on accessibility of antecedents in Chinese. *Acta Psychol. Sin.* 28, 198–203.

Conflict of Interest: 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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Yang, Zhang, Ma and Lu. 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) and the copyright owner(s) 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: 29 November 2021
doi: 10.3389/fpsyg.2021.648478

Correction by Focus: Cleft Constructions and the Cross-Linguistic Variation in Phonological Form

Markus Greif[1] *and Stavros Skopeteas*[1,2]*

[1] Linguistics Department, Bielefeld University, Bielefeld, Germany, [2] Linguistics Department, University of Göttingen, Göttingen, Germany

A challenging issue of cross-linguistic variation is that the same syntactic construction may appear in different arrays of contexts depending on language. For instance, cleft constructions appear with contrastive focus in English, but in a larger array of contexts in French. A part of the cross-linguistic variation may be due to prosodic differences, since prosodic possibilities determine the array of focus structures that can be mapped onto one and the same syntactic configuration. In the present study, we compare languages with flexible nuclear-accent placement (English, German), with languages that do not use this prosodic strategy (French, Mandarin Chinese). In a speech production experiment, we examine the prosodic realization of contrastive focus and identify prosodic reflexes of focus in all languages. The presence of different phonetic reflexes of focus suggests that – anything else being equal – the same syntactic constructions should be possible in the same array of contexts. In an acceptability study with written questionnaires, we examined the felicity of cleft constructions in contexts licensing a focus within the cleft clause. This focus structure is orthogonal to the preferred focus structure of cleft constructions and can appear in cases of second-occurrence foci (in contexts of correction). The obtained judgments reveal a distinction between languages with flexible nuclear-accent placement (English, German) and languages with other types of reflexes of focus (French, Chinese): languages of the former type have an advantage in using cleft constructions with a focus within the cleft clause, which shows that the array of contexts of using clefts in English and German is not a proper subset of the array of contexts applying to the same constructions in French and Chinese. The obtained differences can be explained by the role of prosodic devices and corroborate the view that prosodic reflexes of focus have different semantic-pragmatic import: it is easier to establish a focus structure that is orthogonal to the syntax in a language with flexible nuclear-accent placement (English, German); this does not hold for prosodic correlates of focus that reinforce the articulation of prosodic constituents (French) or the articulation of lexical tones (Chinese).

Keywords: focus, correction, pitch accent, tonal compression, second occurrence focus, cleft constructions, deaccenting

OPEN ACCESS

Edited by:
Yan Huang,
The University of Auckland,
New Zealand

Reviewed by:
Gilbert Ambrazaitis,
Linnaeus University, Sweden
Anja Arnhold,
University of Alberta, Canada

***Correspondence:**
Stavros Skopeteas
stavros.skopeteas@uni-goettingen.de

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

Received: *31 December 2020*
Accepted: *27 October 2021*
Published: *29 November 2021*

Citation:
Greif M and Skopeteas S (2021) Correction by Focus: Cleft Constructions and the Cross-Linguistic Variation in Phonological Form. Front. Psychol. 12:648478. doi: 10.3389/fpsyg.2021.648478

INTRODUCTION

Discourse notions such as topic and focus are reflected in different grammatical layers, notably in syntax and prosody. The idea that these layers are complementary has been fruitfully used in order to account for the fact that similar syntactic constructions appear in different arrays of contexts depending on language. Vallduví and Engdahl (1996: 497) explain the differences in the use of syntactic movement in Catalan and English in terms of prosodic plasticity. 'Plastic' languages, such as English, shift the nuclear stress signaling that the focus is part of the stressed constituent; in 'non-plastic' languages, such as Catalan, the nuclear stress appears in a fixed position within the linearization (in case of Catalan, it is the rightmost constituent); syntactic operations are employed such that the focus appears in the position that bears the nuclear stress.[1] In the same vein, Samek-Lodovici (2005) accounts for the choice of alternative strategies to express focus in English, Italian, and Bantu languages by means of alternating rankings of constraints that sanction deviations from syntactic and prosodic principles. Zubizarreta (1998: 21–22) observes that languages differ with respect to the expression of prosodic prominence of focus. In English, German, and French, clause-initial non-contrastive foci are realized with prosodic prominence followed by deaccenting. In contrast to these languages, Spanish and Italian have a default prosodic prominence on the rightmost prosodic constituent that is not modulated by focus; in order to maintain this prosodic pattern, these languages employ deviations from the canonical word order such that non-contrastive foci surface rightmost in the clause. These approaches share the reasoning that syntactic movement is a last resort, employed for discourse functions that cannot be expressed by prosodic means in the language at issue. The distinction between two classes of languages may be oversimplified, as various instrumental phonetic studies on prosody show (see, e.g., effects of focus on the pitch range of tonal events in Chinese; Xu, 1999). Finally, it is cross-linguistically possible to increase the articulatory effort in order to draw the attention of the hearer to salient parts of the utterance (see effort code in Gussenhoven, 2004: 85–89). However, we know that the exact semantic-pragmatic value of similar prosodic devices can vary between languages (see Vander Klok et al., 2018 for differences in the prosodic means expressing variation in prominence between English and French). Thus, the core question is how different prosodic means of expressing prominence (e.g., nuclear-accent placement in English, pitch range expansion of tonal events in Chinese) can account for the possibility of using the same construction in different contexts depending on language.

Within this line of thought, the present study examines cleft constructions, which are informative for the general question at issue since these constructions are associated with a particular information structure.[2] In the typical instances of cleft constructions in English, the 'pivot,' that is the constituent in the matrix clause, is contrastively focused; this construction asserts that the proposition is true for the pivot to the exclusion of some alternatives that are relevant in discourse (see 'cleft-focus principle,' Rochemont, 1986: 133). The 'cleft clause,' that is the constituent that surfaces as a relative clause, contains the background information. Example (1) illustrates a context in which the contextual conditions for a felicitous use of the cleft construction are met. In this realization of the cleft, the nuclear stress is aligned with the pivot, as indicated by the small capitals.

(1) A: *Did Mary buy the bicycle?*
B: *No, it's* JOHN *that bought the bicycle.*

Beyond cleft constructions with a focus in the pivot, as seen in (1), earlier research in English has shown that cleft constructions appear in a variety of contexts such that the focus domain of the utterance is (a part of) the cleft clause (e.g., 'informative presupposition clefts' in Prince, 1978; 'topic-comment clefts' in Hedberg, 1990, 2013; detailed classification in Delin, 1992; discussion of various classes of examples in Hartmann (2015): 252–270). The information structure of these examples is reflected in prosody: the nuclear accent in informative-presupposition clefts is realized within the cleft clause (see discussion in Delin, 1992, 1995; Hedberg, 2013), while the pivot is not completely deaccented (Hartmann, 2015: 214).

In the present study, we examined a particular type of context that enforces a focus within the cleft clause, namely cases of correction, as introduced in (2). Assume a context containing a cleft construction such that the pivot of the cleft (John) is focused as in (2A). In this context, it is possible to use a cleft construction as in (2B), correcting a part of the utterance in (2A). Correction establishes a relation between an 'antecedent statement,' that is available in the discourse, and a 'corrective statement,' that is a denial of (a part of) the antecedent statement. The corrective statement contains a replacement that is interpreted as incompatible with the antecedent statement and which is contrastively focused (Steube, 2001; Van Leusen, 2004; Repp, 2010). An important aspect of correction is the *structural parallelism* between the corrective statement and the antecedent statement, which is an instruction to the addressee to identify the relevant statement in discourse (Van Leusen, 2004: 437; Clifton and Frazier, 2016). The effects of structural parallelism are shown in (2): assuming an antecedent statement that contains a cleft construction (for reasons that depend on the contextual conditions of A and are not crucial for our purposes), it is possible to utter a corrective statement as in B, that is structurally parallel to the antecedent claim and involves a contrastive focus within the cleft clause. This configuration deviates from the expectation that the pivot of a cleft construction is the main focus of the utterance.

[1]The concept 'nuclear stress' refers to the maximal prosodic prominence within an intonation phrase and is underspecified with respect to the phonetic reflexes of prominence. The concept 'nuclear accent' refers to the pitch accent that realizes the nuclear stress (in languages expressing prosodic prominence by means of pitch accents).

[2]Cleft constructions are analyzed as the result of syntactic movement (see Kiss, 1998), but the exact syntactic analysis of cleft constructions is not crucial for the argumentation of the present article. The crucial issue is that cleft constructions convey the same propositional content as the corresponding canonical constructions with a different information structure.

(2) A: . . . *It's* [JOHN]$_{FOC}$ *that bought the car.*
B: *No, it's* [*John*]$_{FOC2}$ *that bought the* BICYCLE$_{FOC1}$.

The corrective statement in (2B) contains a complex focus structure, involving a primary focus (FOC1) and a secondary focus (FOC2). The primary focus is the focus of the corrective assertion that is expressed by the nuclear accent. The focus on 'bicycle' excludes the alternative in the antecedent statement: 'it's John that bought the bicycle' is contrasted to 'it's John that bought the car.' Additionally, this utterance has a second-occurrence focus[3], FOC_2, which is inherited from the context utterance. If the cleft construction in (2A) identifies 'John' in contrast to further relevant alternatives (e.g., 'Peter' or 'George'), this information is presupposed by the corrective statement in (2B). The second-occurrence focus is expressed by the cleft construction in this case and may have some secondary prosodic prominence (Féry and Ishihara, 2006; Beaver et al., 2007; Howell, 2011; Büring, 2015; Baumann and Ishihara, 2016). The asserted and presupposed information of (2B) can be paraphrased as: 'it's John (in contrast to 'Peter' or 'George') that bought the bicycle (not the car).'

The cleft constructions in (1) and (2) share the interpretation that some contextually relevant alternatives to the pivot are excluded (which applies to further contextual instances of cleft constructions, as shown by Hartmann, 2015: 253). These constructions differ with respect to the partitioning of the utterance in asserted and presupposed information, which is expressed by the nuclear stress placement, as summarized in (3).

(3) Cleft constructions and focus structure
The pivot of a cleft construction excludes alternatives that are relevant in the context.

(a) If the nuclear stress falls within the pivot, the exclusion of alternatives is the asserted information (focus).
(b) If the nuclear stress falls within the cleft clause, the asserted information is in the cleft clause (focus), while the exclusion of alternatives is part of the presupposed information (second-occurrence focus).

The crucial issue is that the variation in the focus structure of cleft constructions requires the possibility of variable nuclear stress placement, as stated in (3). The predictions of (3) are straightforward for languages such as English and German that realize the nuclear stress by means of pitch accents. Our first question is how this contrast can be expressed in languages that do not rely on pitch accents for signaling focus, such as French and Chinese. In order to establish the corresponding prosodic means in these languages, we conducted a cross-linguistic study on speech production (comparing English, German, French, and Chinese), which is reported in Section 2. The results of this study show that reflexes of prosodic prominence appear in all examined languages, but these reflexes are different in nature.

With this background, we examined whether a cleft construction with a focus in the cleft clause is equally felicitous in these languages (Section 3). Judgments of contextual felicity revealed a typological distinction between languages with flexible nuclear-accent placement (English and German) and languages that do not rely on this strategy (French and Chinese). Hence, these findings are in line with the idea that various classes of prosodic events have distinct semantic-pragmatic import: precisely, using cleft constructions with a focus in the cleft clause has an advantage in languages in which nuclear-accent placement unambiguously identifies the intonational nucleus (English and German); see discussion in Section 4.

[3]Second-occurrence foci refer to expressions that contain the focus domain of a relevant statement within the background partition of the utterance (see Krifka, 1997; Rooth, 1996, 2009; Büring, 2015).

PROSODIC REFLEXES OF FOCUS

Aims

The present experiment examines whether canonical and cleft constructions can be realized with different prosodic patterns depending on focus in typologically different languages: languages allowing for flexible placement of nuclear accents (English, German), and languages that do not employ this prosodic strategy (French, Chinese).

Method

Participants

Sixteen native speakers of each language participated in this study. They were explained that their participation was voluntary and that the data will be used in anonymized form for research purposes. Written consent (translated into the native language of the participants) was acquired; participants were paid for their contribution to the experiment. Sex was controlled in the samples in order to outbalance the influence of sex on pitch: English (n = 16, female = 8, age range = 18–29, average = 22.1; collected in London), German (n = 16, female = 8, age range = 19–34, average = 23.4; collected in Bielefeld), French (n = 16, female = 8, age range: 18–44 = average 25.9; collected in Lyon), and Chinese (n = 16, female = 8, age range = 18–24, average = 20.8; collected in Beijing).

Factorial Design

The trials of this study presented short dialogical interactions. The instructor introduced a context, as in (4A). The participant produced a target utterance (4B) containing a corrective statement, whose antecedent was the last sentence of the context.

(4) A: Everyone brought something to the potluck today. Peter brought the bread.
B: No, [Layla]$_F$ brought the bread today.

In order to assess the impact of contrastive focus on the prosodic realization of canonical and cleft constructions, we designed an experiment with the factors FOCUS and CONSTRUCTION of the target utterance; see (5). The factor FOCUS refers to the focus domain of the utterance, which depends on the relation of the target utterance to the last sentence of the context, and contains two levels: subject focus and object focus. The factor CONSTRUCTION relates to the syntactic construction of the target utterance: either 'canonical constructions' or 'cleft constructions.' The target utterance has always the same structure as the antecedent statement, maintaining the structural

parallelism of correction as introduced in (2): canonical and cleft constructions in the target utterance always relate to canonical and cleft constructions respectively in the context utterance.

(5) Factorial design of the speech production study

(a) FOCUS: subject, CONSTRUCTION: canonical
A: *Everyone brought something to the potluck today. Peter brought the bread.*
B: *No,* [*Layla*]$_F$ *brought the bread today.*

(b) FOCUS: subject, CONSTRUCTION: cleft
A: *Everyone brought something to the potluck today. It's Peter that brought the bread.*
B: *No, it's* [*Layla*]$_F$ *that brought the bread today.*

(c) FOCUS: object, CONSTRUCTION: canonical
A: *Everyone brought something to the potluck today. Layla brought the salad.*
B: *No, Layla brought the* [*bread*]$_F$ *today.*

(d) FOCUS: object, CONSTRUCTION: cleft
A: *Everyone brought something to the potluck today. It's Layla that brought the salad.*
B: *No, it's Layla that brought the* [*bread*]$_F$ *today.*

Material

The experimental conditions were implemented in four items involving different lexicalizations of simple transitive clauses. All lexicalizations had the same syntactic constituents, the same number of syllables and the same word stress pattern (English, German) or tonal structure (Chinese); voiceless obstruents were avoided whenever possible in order to reduce missing values in the f_o measurements;[4] see full listing of the items in **Supplementary Material**, Section 2. The number of items is arguably low. Beyond limitations in developing lexicalizations with the present phonological requirements (same syllabic structure, word stress, tonal structure, avoidance of voiced consonants), the main motivation for this decision is to obtain minimal pairs of prosodic realizations of the same lexicalization and by the same speaker under different treatments. Hence, we created four different lexicalizations in order to obtain four repeated observations with each speaker. The drawback of the limited sample of items is that the findings cannot claim generalizability for the population of possible lexicalizations.

The objects were not final within the utterance, such that tonal events that are associated with object focus do not clash with the final lowering at the right edge of the utterance. Therefore, we used a clause-final temporal adverb in those languages in which the object would otherwise be the rightmost constituent (English and French). These items were recorded in all conditions with all participants, which renders a total of 4 items × 16 participants = 64 tokens per experimental condition (à four conditions: 256 utterances per language). Experimental items were mixed with fillers in a proportion 1 (target): 3 (fillers), whereby a part of the fillers (1:3) were items of a further experiment and the remaining fillers (2:3) were distractors. All trials (targets and fillers) were performed with the same instruction and had the same dialogical structure, as illustrated in (6).

The same types of constructions (canonical constructions vs. cleft constructions) were examined in all languages at issue. German declarative main clauses have a verb-second order, as seen in (6a). Cleft constructions as in (6b) are possible in German but occur less frequently and in restricted contexts compared to English (Dufter, 2009: 168; Fischer, 2009: 90). Narrow focus is usually expressed by prosodic means and/or syntactic movement in German. It is possible to use German cleft constructions with a focus within the cleft clause (Fischer, 2009: 168; Hartmann, 2015: 271), as discussed in Section 1 for English ('informative presupposition clefts' in terms of Prince, 1978). Experimental results show that the exhaustive interpretation (i.e., the interpretation that the pivot is the only alternative for which the presupposition of the cleft clause holds true) is not part of the truth-conditional meaning of German clefts (Drenhaus et al., 2011), which differs from English clefts that are exhaustively interpreted (Kiss, 1998: 268; Destruel and De Veaugh-Geiss, 2018).

(6) German[5]

(a) Canonical construction

Nein,	*Leni*	*hat*	*die*	*Bluse*	*getragen.*
NEG	Leni	have:3SG	DEF	blouse	wear:PTCP

'No, Leni wore the blouse.'

(b) Cleft construction

Nein,	*es*	*war*	*Leni,*	*die*	*die*
NEG	3SG	be:PST:3SG	Leni	REL	DEF
Bluse	*getragen*	*hat.*			
blouse	wear:PTCP	have:3SG			

'No, it was Leni that wore the blouse.'

French *c'est* clefts, as in (7), occur in a larger array of contexts than English *it*-clefts. While English clefts are licensed by contrastive focus, French clefts also appear in answers to *wh*-questions (Skopeteas and Fanselow, 2010). Furthermore, French *c'est* clefts with a subject as pivot do not only occur when the subject is in narrow focus, but also whenever the subject is part of a larger focus domain (Lambrecht, 2001; corpus findings in Karssenberg and Lahousse, 2018). While English clefts come with an exhaustive interpretation, this is not necessarily the case for French clefts (Destruel and De Veaugh-Geiss, 2018).

(7) French

(a) Canonical construction

Non,	*Lilou*	*a*	*porté*	*le*
NEG	Lilou	have:3SG	bring:PTCP	DEF
gilet	*hier.*			
waistcoat	yesterday			

'No, Lilou wore the waistcoat yesterday.'

[4]For the fundamental frequency we use the notation f_o, whereby "o" stands for oscillation (Titze et al., 2015).

[5]Abbreviations for glosses: 3, 3rd person; ACC, accusative; COP, copula; DEF, definite; EXPL, expletive pronoun; NEG, negation; NOM, nominative; OBJ, object; PFV, perfective; PST, past; PTCP, participle; REL, relative pronoun; SG, singular.

(b) Cleft construction

Non,	*c'*	*est*	*Lilou*	*qui*
NEG	EXPL	be:3SG	Lilou	REL
a	*porté*	*le*	*gilet*	*hier*.
have:3SG	bring:PTCP	DEF	waistcoat	yesterday

'No, it's Lilou that wore the waistcoat yesterday.'

In Chinese, the canonical order with finite verbs is SVO; see (8a) (see discussion in Huang et al., 2009: 199–202). The 'bare *shi*' construction in (8b) (with *shi4* preceding the subject) is a cleft construction, typically expressing contrastive focus on the subject. Similarly as with French, the same construction occurs in sentence focus (Cheng, 2008: 255; Paul and Whitman, 2008: 426; Von Prince, 2012: 342; Paul, 2015: 216; see discussion of the tonal properties in Section 2.4).

(8) Chinese

(a) Canonical construction

bu4	*dui4*,	*Niu2*	*Meng2*	*mai3*	*niu2 rou4*	*le*.
NEG	correct	Niu	Meng	buy	beef	PFV

'No, Niumeng bought the beef.'

(b) Cleft construction

bu4	*dui4*,	*shi4*	*Niu2*	*Meng2*	*mai3*	*niu2 rou4*
NEG	correct	COP	Niu	Meng	buy	beef

le.
PFV

'No, it's Niumeng that bought the beef.'

Procedure

Recordings took place in quiet rooms in the four places of data collection (London, Bielefeld, Lyon, Beijing). The data was recorded with an Olympus digital recorder (LS-13) with in-built microphones and saved in .wav files at a sampling frequency of 44.1 kHz. The participants were presented with the material in a power point presentation. Each trial was presented in two slides: in a first slide, they read a context-target pair as in (5) and were instructed to look carefully at the dialogue and to memorize the target sentence. In a second slide, only the context was presented, while a native speaker/instructor performed it orally (instructors were advised to perform the context sentences as natural contributions in a dialogue and to avoid a non-expressive style like repeating sentences from a list). The participants were instructed to perform the memorized target utterance in a way that naturally fits to the context (the purpose of this manipulation was to avoid effects of read speech). The participants were allowed to repeat the trial if they thought that their performance was not natural enough (without further guidance by the instructor).

Data Analysis

The recordings were processed in praat (Boersma and Weenink, 2020). The data set contained 64 utterances per condition/language; a few tokens had to be removed due to speech disfluencies or errors (two tokens in German and five tokens in Chinese). TextGrid objects were created for the valid data, with intervals corresponding to the syllables of the target utterances. All sound files and TextGrid objects are available at zenodo (Greif and Skopeteas, 2021).

A praat script written by the authors extracted the timing of the onset and the offset of each syllable, as well as the mean f_0 of five equal time bins per syllable. The extracted measurements were processed in *R* (R Core Team, 2020). The f_0 values in Hz were converted into semitones with a reference value of 50 Hz, with the formula $f_{0\ (\text{semitones})} = 12(\log_2 . f_{0\ (\text{Hz})}/50)$ (Nolan, 2003; Grice et al., 2007; Wang and Xu, 2011).

The f_0 values in semitones were averaged per experimental condition in order to detect the impact of the factors at issue on the f_0 excursion in visualizations. Statistic evaluation was conducted on the non-averaged data.

Linear mixed-effects models were fitted on the (semitone transformed) f_0 measurements in each area of interest (subject or object, see details in Section 2.3) separately (using package *lme4* in R; Bates et al., 2015). We examined f_0 excursions as time series, with the f_0 mean of time bins as dependent variable. The fixed effects were the experimental factors FOCUS (level 0 = object; level 1 = subject) and CONSTRUCTION (level 0 = canonical; level 1 = cleft), and the continuous variable of TIME (levels: 1–5), whose levels refer to the corresponding time bin within the syllable. Including TIME to the model offers the possibility to examine the impact of the fixed effects on the f_0 excursion as a function of time: the interaction effects with TIME reflect the impact of the corresponding fixed factor on the f_0 slope within the area of interest (Barr, 2008). Starting with a random-effects structure with intercepts for PARTICIPANTS and ITEMS as well as by-PARTICIPANTS and by-ITEMS random slopes of FOCUS and CONSTRUCTION, we identified the maximal random-effects structure that converges in all languages for the analyses in a certain area of interest.[6] Keeping the maximal converging random-effects structure constant (as suggested by Barr et al., 2013), we reduced the fixed-effects structure (FOCUS × CONSTRUCTION × TIME) with a backward-elimination procedure of non-significant effects (performed automatically by the function *step* of the package *lmerTest* in R; Kuznetsova et al., 2017). The fixed effects that were not nested in a higher interaction were additionally tested with Likelihood Ratio Tests (Bates et al., 2015: 35); for the significance of fixed effects that were nested in higher interactions, we can only rely on the *t*-values (ratio of the estimate to its standard error).

Predictions

The experimental material contains two areas of interest: the f_0 excursion of the subject and f_0 excursion of the object; in languages with stress, either lexical (German, English) or postlexical (French), the area of interest is the corresponding stressed syllable. In the area of the subject, we expect a contrast between nuclear accents (if the subject is focused) and prenuclear accents (if the focus falls on the object); in the area of the object, we expect a contrast between nuclear accents (if the

[6] A model fails to "converge" if the procedure estimating the model parameters does not find a solution within the defined number of iterations. A common treatment of this problem is to simplify the random-effects structure, since the likelihood of convergence failure increases with the complexity of model parameters (Barr et al., 2013: 261).

object is focused) and deaccenting (if the focus falls on the subject). In Chinese, we expect that the f_o excursion of non-focused constituents will be tonally compressed compared to the f_o excursion of focused constituents (in either area). The type of accent depends on language and will be introduced with the presentation of the results in Section 2.4. In all cases, the expected contrasts imply a difference in the f_o slope, while the direction of the difference is language-specific (it depends on the prosodic events at issue).

The predictions of this study will be examined by testing for an interaction of the fixed factors with the variable of TIME within the areas of interest (i.e., the syllables in which phonological considerations predict reflexes of focus). Effects of TIME are evidence for a difference in the f_o slope, reflecting tonal events aligned with the area of interest (Grabe et al., 2007; Isaacs and Watson, 2010). Hence, an interaction FOCUS × TIME or an interaction CONSTRUCTION × TIME indicates that the corresponding fixed factor has an impact on the change of f_o within the area of interest. Effects that are independent of the time variable, such as a main effect of FOCUS, are evidence for a difference of the f_o level (see Barr, 2008 concerning the relevance of 'rate effects' in time series).

With this background, the major question in cross-linguistic perspective is whether FOCUS × TIME effects appear in all languages. The distinction between plastic (English, German) and non-plastic (French, Chinese) languages predicts that the effects of FOCUS will appear only in the former language type. However, earlier studies have shown that various phonetic reflexes of focus, such as a pitch range expansion or reflexes of demarcation of focused constituents, are found in non-plastic languages as well (see Xu, 1999; Chen and Gussenhoven, 2008 on Chinese and German and D'Imperio, 2010; Delais-Roussarie et al., 2015 on French), which predicts an effect on the f_o slope in all languages.

An interaction CONSTRUCTION × TIME may appear if certain constructions are associated with prosodic events that are independent of focus. Precisely, cleft constructions differ with respect to prosodic phrasing, such that the cleft clause forms an intonation phrase on its own (Féry, 2013: 699 on French); edge tones that delimit intonation phrases may appear around the boundary between the pivot and the cleft clause.

A threefold interaction FOCUS × CONSTRUCTION × TIME indicates that the effect of FOCUS on the f_o slope is modulated by CONSTRUCTION. Since cleft constructions with a focus in the cleft clause bear a second-occurrence focus as seen in (2), subject constituents may be not completely deaccented, which predicts a threefold interaction within the area of interest of the subject. In cross-linguistic perspective, effects of second-occurrence focus entail effects of focus. That is, a threefold interaction may appear in a subset of the languages that have a FOCUS × TIME interaction. Our predictions are summarized in (9).

(9) Predicted effects on the f_o slope

   (a) FOCUS × TIME: focus influences the f_o slope (language-specific effects).
   (b) CONSTRUCTION × TIME: canonical and clefts constructions differ with respect to p-phrasing.
   (c) FOCUS × CONSTRUCTION × TIME: second-occurrence focus in cleft constructions predicts that the effect of focus on the f_o slope will be modulated by construction.

Results

The f_o excursions in **Figure 1** illustrate the basic contrast between early and late foci in British English. Annotations indicate the tonal events that are relevant for our discussion on the prosodic reflexes of focus, assuming the ToBI conventions (Veilleux et al., 2006). When the subject is focused (**Figure 1A**) it is realized with a bitonal accent L + H*, which stands for a substantial rising pitch movement that reaches a high target within the stressed syllable; this realization is characteristic of contrastive foci in English (Ladd, 2008: 96; Watson et al., 2008; Gotzner, 2015: 130–136). The realization of a subject preceding the focus in **Figure 1B** also has a rising f_o excursion, starting from a low target within the stressed syllable and rising toward a high target that may be reached after the stress (L* + H).[7] The prosodic realization of the objects is different in both figures. When the object is focused, it is realized with a rising contour (**Figure 1B**), similarly as with the focused subject in **Figure 1A**. When the object follows the focus, it is deaccented (**Figure 1A**), which means that it does not contain any significant prosodic events (Ladd, 2008: 231–236) and ends up with a final low target as expected for declaratives, which is phonologically represented by the sequence of a phrase tone (L−) and a boundary tone (L%).

The average f_o excursions of British English (**Figure 2**)[8] show a major distinction between early focus (on the subject, blue line) and late focus (on the object, red line), which applies to canonical and cleft constructions. The f_o rise in the stressed syllable of focused subjects (gray cell) has a greater slope with focus on the subject (blue line) than with focus on the object (red line). The realization of the objects show a rising contour when the object is focused (red line) and is deaccented when the object is given (blue line). These properties apply to canonical and cleft constructions, which means that prosodic marking of focus is not compensated by marking the focus in syntax (see the same effect for Canadian English in Arnhold, 2021).

The German data shows a similar pattern in canonical and cleft constructions (**Figure 3**). Focused subjects (blue lines) are realized with an f_o excursion rising up to a H target that is close to the right edge of the stressed syllable, reflecting the fact that German has a bi-tonal accent L + H* for contrastive assertions (Grice et al., 2005: 65, 71; see Alter et al., 2001 on contrast). Non-focused subjects (red lines) optionally have prenuclear accents, reaching an f_o maximum after the right edge of the stress, which reflects the fact that the H-target of prenuclear accents (L* + H) may follow the stressed syllable (Féry and Kügler, 2008; Baumann and Riester, 2013: 20; Féry, 2017: 154). The impact of focus on

[7] Prenuclear accents may make various contributions to the meaning of sentences (Baumann and Winter, 2018; Cole et al., 2019), but they are not necessarily associated with information structure since they may just reflect the rhythmical organization of the utterance (Baumann et al., 2021). The relevant aspect for our data is that prenuclear accents may be particularly weak (in terms of f_o scaling) if they precede narrow foci as in our data (Calhoun, 2010: 28).

[8] Abbreviations for Figures: ADV, adverb; AUX: auxiliary; C, copula; E, expletive; NEG, negation; OBJ, object; T, tense; REL, relative; SBJ, subject; V, lexical verb.

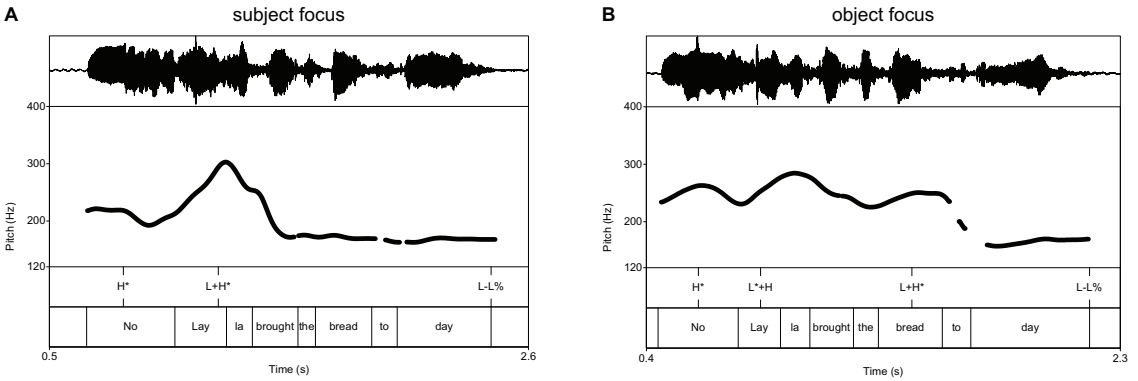

FIGURE 1 | Illustrative examples of **(A)** subject and **(B)** object focus in British English.

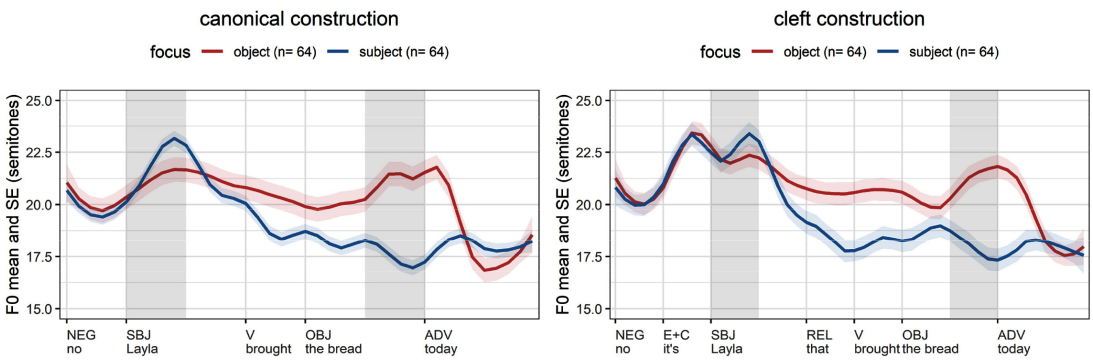

FIGURE 2 | Average f_o measurements in British English (time normalization based on five equal intervals per syllable; vertical lines: word edges; gray cells: areas of interest, stressed syllable of subject and object).

object constituents is similar: a rise within the stressed syllable (L + H*) when the object is focused (red lines) viz. deaccented objects with a flat contour when the object is given (blue lines).

In French, the rightmost full (i.e., non-schwa) syllable is characterized by metrical prominence, which is reflected in lengthening and tonal activity; metrical prominence is assigned postlexically in French, which means that it is not determined by the lexicon (see summary in Post, 2000: 8–9; Féry, 2014). In terms of the French ToBI (Delais-Roussarie et al., 2015), the last syllable of the accentual phrase is associated with a high tonal target (H−), while the last accentual phrase ends up with a low target (L−L%); see **Figure 4**. French accentual phrases may start with a rise within the initial syllable (German and D'Imperio, 2010; Delais-Roussarie et al., 2015). Since these events are associated with edge syllables, we code them as edge tones associated with the left edge of an accentual phrase (−L + H) (following Féry, 2014). Initial rises are reported to appear more often with contrastively focused constituents (see German and D'Imperio, 2010; Delais-Roussarie et al., 2015), especially in contexts of correction (Vander Klok et al., 2018); however, the function of these events is controversial, since they may be used to draw the attention of the hearer to not focused constituents and there are also empirical studies disputing its correlation with contrastive focus (Cole et al., 2019: 130). The data in **Figure 4** illustrate this contrast: focused subjects may be realized with an initial rise (**Figure 4A**), such that the high target is aligned with the right edge of the first syllable; non-focused subjects are realized with a (lower scaled) high edge tone aligned with the right edge of the accentual phrase (**Figure 4B**). The initial rise can also appear with focused objects (**Figure 4B**), while objects are not accented when following the focused subject (**Figure 4A**).

The averages per experimental condition (**Figure 5**) confirm that the introduced phenomena depend on information structure. The average f_o excursion of focused subjects (blue lines) targets an earlier local maximum than the corresponding excursion of non-focused subjects (red lines). Focused objects (red lines) also show an initial rise in contrast to non-focused objects (blue lines). Our data shows that tonal events following the nucleus are not necessarily erased in French (Di Cristo and Jankowski, 1999: 1567; Jun and Fougeron, 2000: 230; Féry, 2014):[9] prosodic words in the postfocal domain display the same type of f_o excursion with their focused counterparts – but with a compressed pitch range.

[9]Di Cristo and Jankowski (1999: 1567) report that postnuclear pitch variation is reduced, but not eliminated, Jun and Fougeron (2000: 230) conclude that postnuclear domains are deaccented but not dephrased, Féry (2014) argues that only prosodic phrases but not prosodic words are deaccented in French.

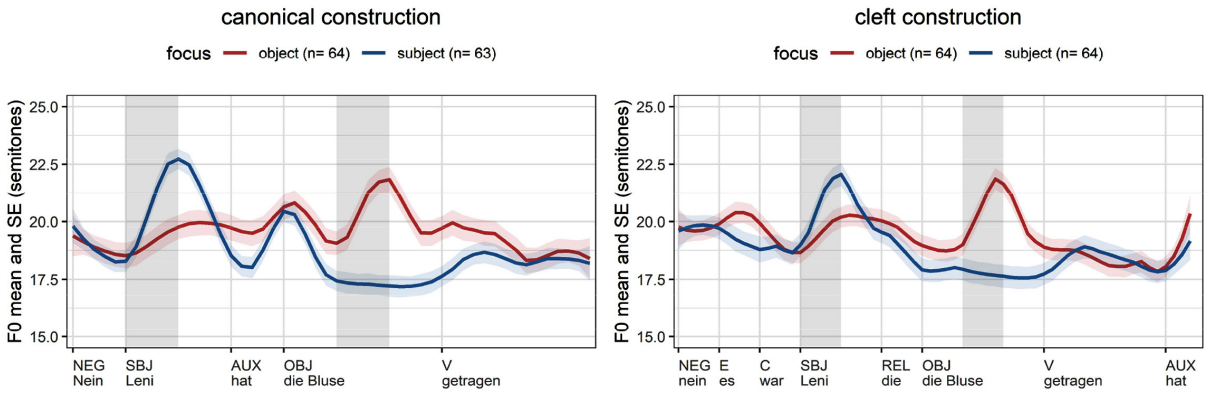

FIGURE 3 | Average f_0 measurements in German (time normalization based on five equal intervals per syllable; vertical lines: word edges; gray cells: areas of interest, stressed syllable of subject and object).

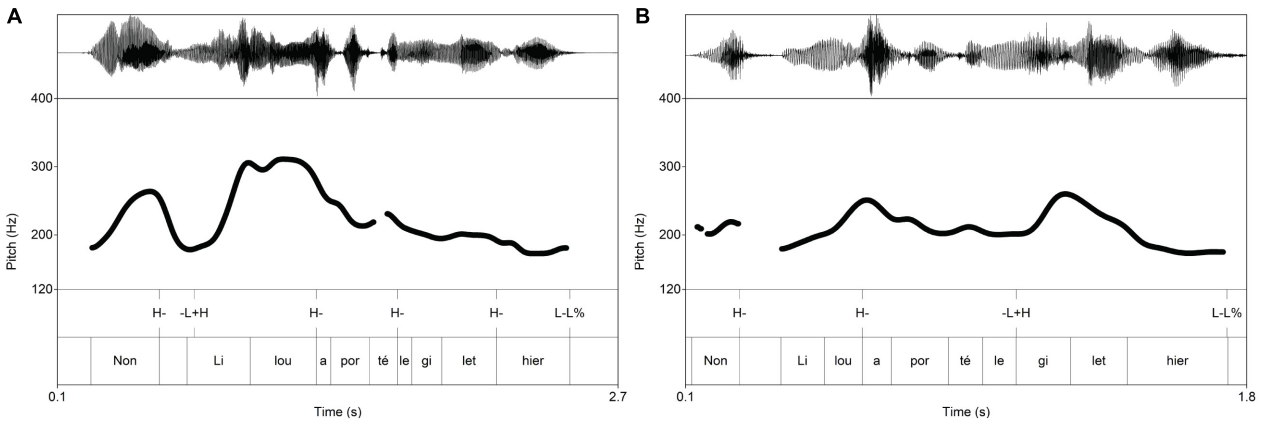

FIGURE 4 | Prosodic realization of **(A)** subject and **(B)** object focus in French.

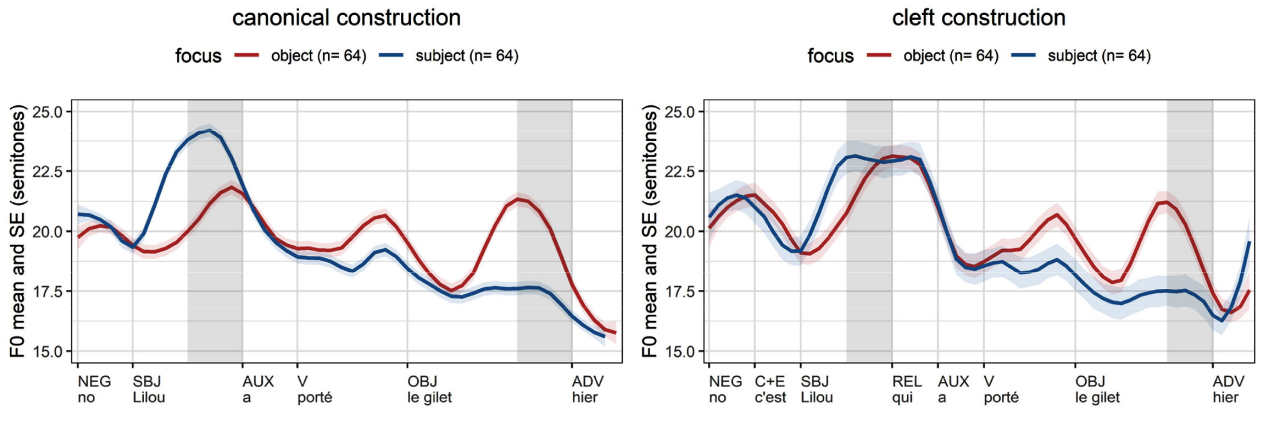

FIGURE 5 | Average f_0 measurements in French (time normalization based on five equal intervals per syllable; vertical lines: word edges; gray cells: areas of interest, stressed syllable of subject and object).

Mandarin Chinese displays a phonological contrast between four lexical tones (T1: high level; T2: rise; T3: fall-rise; T4: fall). The target words in our material contain the simple contour tones T2 and T4 that are comparable since they consist of two tonal targets (i.e., T2: LH, T4: HL). All items have the tonal sequence T2-T2 (rise-rise) for subjects and T2-T4

(rise-fall) for objects; see (8) and **Supplementary Material**, Section 2.1.4. The choice of T2/T4 was just determined by convenience for the selection of appropriate lexical material and maintained constant across items. Word stress is not applicable to Chinese. Even if some studies report a preference for initial prominence in compounds (Duanmu, 2007: 135, 142), both syllables are areas of interest for our study (see **Figure 6**), since there is no reason to expect reflexes of focus only in the initial syllable. Focus is reported to be reflected in an expansion of the pitch range of lexical tones, with a greater effect on f_o maxima than f_o minima (Xu, 1999: 69; Greif, 2012: 38) as well as by a general increase of the distinctness of tonal targets, which resembles hyperarticulation effects of focus on vowel quality (Chen and Gussenhoven, 2008: 744). This kind of hyperarticulation is also seen in our data: the T2–T2 sequence in the subject is realized with two distinct rising excursions when the subject is focused, but this contour is leveled out into a single rise when the subject is out of focus. A similar contrast applies to the object constituents. The T2–T4 sequence results in a hat contour (LHL), whose peak is reached beyond the offset of the first syllable (Xu and Wang, 2001: 331): this hat contour appears with a reduced pitch range when the object follows the focus, which is evidence for postfocal tonal compression. The asymmetry between prenuclear and postnuclear tonal compression is similar to the asymmetry between prenuclear and postnuclear deaccenting in Germanic languages (Chen, 2010: 520). While the pitch compression is radical in the postnuclear domain, prenuclear tones only slow slight differences in terms of pitch range (see lexical tones of subjects under object focus).

Linear mixed-effects models with the factors FOCUS, CONSTRUCTION, and TIME were fitted on the f_o measurements within the stressed syllables (for objects and subjects separately; see details in 2.2.5). In Chinese, we analyzed the first and the second syllable separately, in order to maintain the same degrees of freedom in all analyses and since we cannot reduce the analysis to a single syllable based on assumptions about word stress.

The maximal random-effects structure that converges in all analyses for subjects contains random intercepts for PARTICIPANTS and ITEMS and a by-PARTICIPANTS random slope of CONSTRUCTION. The models of maximal fit for the f_o measurements in the stressed syllable of the subject are listed in **Table 1**. German is the only language with a significant threefold interaction (CONSTRUCTION × FOCUS × TIME), indicating that the effect of FOCUS on the f_o slope is modulated by CONSTRUCTION, such that the difference between focused and non-focused subjects is greater in canonical clauses (therefore the interaction effect is negative); compare blue and red lines in the area of subjects in **Figure 3**. In all languages, we obtain a significant FOCUS × TIME interaction, whose direction is language specific: it is positive with rising accents (English, German, Chinese/syllable 1) and negative with falling accents (French). In either case, this effect means that the f_o change is more rapid when the subject is focused. The models of maximal fit in English and Chinese (syllable 1) contain a negative interaction CONSTRUCTION × TIME, indicating that the f_o change is slower in cleft than in canonical constructions.

The f_o measurements in the object constituent reveal similar results in all languages (**Table 2**). There is a clear interaction effect FOCUS × TIME, which is negative in English, German, and Chinese/syllable 1, since the baseline of object focus is a rise in these languages, while the same syllables in the postfocal domain (subject focus) are rather flat or slightly falling. The corresponding FOCUS × TIME interaction effects are positive in French and in Chinese/syllable 2, in which case the f_o excursion of the object focus is falling. There is no evidence that the difference between canonical vs. cleft constructions (CONSTRUCTION × TIME) plays a role.

Discussion

The results of the present study reveal that all examined languages show prosodic reflexes of focus, either through the prosodic prominence of the focused constituent or through leveling out the prosodic events of the postfocal domain.

All languages have a significant FOCUS × TIME interaction within the subject area (**Table 1**), whose properties vary depending on the language-specific tonal events. In German and English, this effect is positive, reflecting the use of rising accents for marking foci in these languages (Grice et al., 2005: 65, 71; Ladd, 2008: 96). A similar effect is found in the first syllable of the subject in Chinese, reflecting a more rapid rise of rising tones (T2) under focus. Our findings are in line with previous results on pitch range expansion of lexical tones under focus, especially applying to the rising tone (T2) (Xu, 1999; Wang and Xu, 2011; Greif, 2012: 75; Ouyang and Kaiser, 2015: 65). In particular, the average contours in **Figure 6** show an increase of distinctness between subsequent rises within focus, which is in line with the view that tonal realizations are hyperarticulated under focus (Chen and Gussenhoven, 2008: 744). In French, contrastive focus on the subject frequently induces initial rises in the focused constituent resulting in a falling contour within the last syllable (German and D'Imperio, 2010). Hence, focus has an effect on f_o excursions in all languages in our sample, as summarized in (10).

(10) Prosodic prominence of focus

Evidence for prosodic prominence of foci is found in all languages for both subject and object foci and both canonical and cleft constructions. The nature of the obtained effects depends on the specific properties of the languages at issue.

(a) In English and German the focused constituent bears the nuclear accent, which contains a high peak within the stressed syllable; the effects on the f_o slope come from the contrast of the nuclear accents with prenuclear accents (area of interest: subject) or with deaccented domains (area of interest: object).

(b) In French and Chinese, the obtained effects come from phenomena increasing the saliency of prosodic entities: initial rises in French are a general strategy for demarcating prosodic constituents that appear more often with foci; in Chinese, focus is reflected

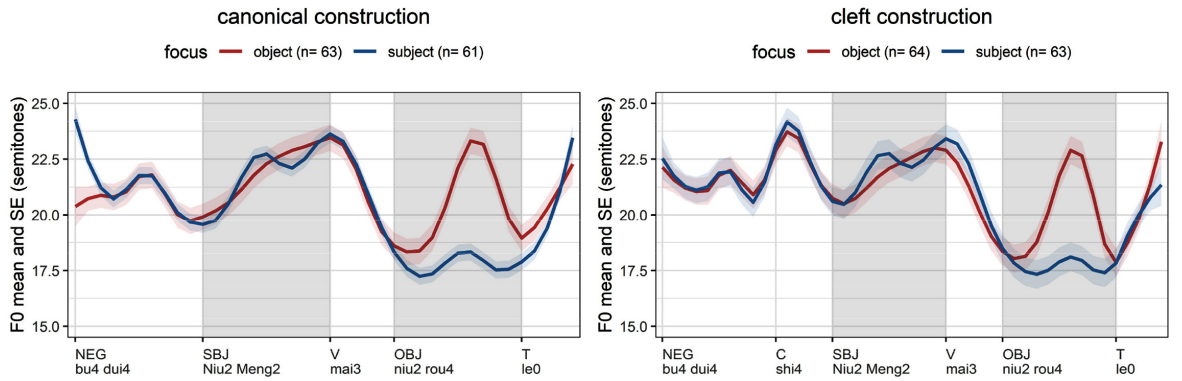

FIGURE 6 | Average f_0 measurements in Chinese (time normalization based on five equal intervals per syllable; vertical lines: word edges; gray cells: areas of interest, subject and object).

TABLE 1 | Linear fixed-effects models of best fit on the f_0 measurements (semitones): subject.

Language	Factor	β	SE	*t*	*p* (<)	Likelihood Ratio Test	
						χ^2	*p* (<)
English	Intercept	19.842	1.494	13.281	0.001		
	Construction (cleft)	1.776	0.197	9.015	0.001		
	Focus (subject)	−0.667	0.167	−3.989	0.001		
	Time	0.409	0.04	10.158	0.001		
	Construction × Focus	−0.329	0.131	−2.519	0.05	6.329	0.05
	Construction × Time	−0.243	0.046	−5.246	0.001	27.218	0.001
	Focus × Time	0.452	0.046	9.769	0.001	91.939	0.001
German	Intercept	18.022	1.226	14.702	0.001		
	Construction (cleft)	0.113	0.254	0.444	–		
	Focus (subject)	−1.327	0.21	−6.316	0.001		
	Time	0.312	0.045	6.994	0.001		
	Construction × Focus	0.619	0.296	2.093	0.05		
	Construction × Time	0.091	0.063	1.449	–		
	Focus × Time	0.912	0.063	14.42	0.001		
	Construction × Focus × Time	−0.302	0.089	−3.388	0.001	11.425	0.001
French	Intercept	19.149	1.182	16.206	0.001		
	Construction (cleft)	1.02	0.259	3.938	0.001		
	Focus (subject)	4.996	0.306	16.339	0.001		
	Time	0.624	0.06	10.406	0.001		
	Construction × Focus	−1.87	0.24	−7.787	0.001	59.2	0.001
	Focus × Time	−0.751	0.085	−8.844	0.001	75.843	0.001
Chinese (syllable 1)	Intercept	19.107	1.341	14.245	0.001		
	Construction (cleft)	1.03	0.173	5.968	0.001		
	Focus (subject)	−0.729	0.137	−5.304	0.001		
	Time	0.497	0.036	13.765	0.001		
	Construction × Time	−0.224	0.041	−5.436	0.001	29.193	0.001
	Focus × Time	0.352	0.041	8.528	0.001	70.603	0.001
Chinese (syllable 2)	Intercept	22.623	1.273	17.77	0.001		
	Focus (subject)	0.336	0.146	2.302	0.05		
	Time	0.245	0.031	7.9	0.001		
	Focus × Time	−0.133	0.044	−3.031	0.01	9.15	0.01

in the hyperarticulation of the tonal targets of phonological events that are independent of focus (lexical tones).

Postnuclear prosodic events are leveled out, which gives rise to a significant Focus × Time interaction in all languages (**Table 2**). Postnuclear leveling encompasses two types of

TABLE 2 | Linear fixed-effects models of best fit on the f_0 measurements (semitones): object.

Language	Factor	β	SE	*t*	*p* (<)	Likelihood-Ratio Test	
						χ^2	*p* (<)
English	Intercept	19.558	1.445	13.539	0.001		
	Focus (subject)	−1.558	0.412	−3.78	0.001		
	Time	0.413	0.034	12.183	0.001		
	Focus × Time	−0.603	0.05	−12.158	0.001	138.68	0.001
German	Intercept	17.996	1.107	16.259	0.001		
	Focus (subject)	0.334	0.286	1.17	–		
	Time	1.016	0.031	32.392	0.001		
	Focus × Time	−1.003	0.045	−22.074	0.001	406.02	0.001
French	Intercept	22.117	1.292	17.115	0.001		
	Construction (cleft)	−0.569	0.247	−2.31	0.05		
	Focus (subject)	−4.666	0.465	−10.039	0.001		
	Time	−0.536	0.051	−10.531	0.001		
	Construction × Focus	0.653	0.205	3.185	0.001	10.095	0.001
	Focus × Time	0.47	0.073	6.485	0.001	41.316	0.001
Chinese	Intercept	17.606	1.396	12.607	0.001		
(syllable 1)	Focus (subject)	0.478	0.211	2.266	0.05		
	Time	0.372	0.032	11.76	0.001		
	Focus × Time	−0.561	0.045	−12.359	0.001	143.51	0.001
Chinese	Intercept	24.411	1.355	18.019	0.001		
(syllable 2)	Construction (cleft)	−0.4	0.14	−2.854	0.05	6.649	0.05
	Focus (subject)	−5.343	0.359	−14.897	0.001		
	Time	−0.841	0.048	−17.407	0.001		
	Focus × Time	0.468	0.07	6.69	0.001	43.902	0.001

phenomena, namely deaccenting and tonal compression. In German and English, the postfocal domain is deaccented: the average excursions of postfocal objects reveal a falling contour without any significant prosodic events, sharply contrasting to the corresponding contour of accented constituents. This finding is in line with previous findings in English (Liberman and Pierrehumbert, 1984; Ladd, 2008: 231–236) and German (Féry and Kügler, 2008; Baumann and Riester, 2013: 20; Féry, 2017: 154). The postfocal excursions in French and Chinese have the same prosodic pattern as the corresponding conditions in focus, realized with a reduced pitch range, which is evidence for tonal compression. In French, tonal compression applies to edge tones: the rising contours encompassing prosodic words are visible in focus or out of focus, with a difference in pitch range, which confirms the view that the reflexes of prosodic phrasing on intonation are still visible in the postfocal domain (Di Cristo and Jankowski, 1999: 1567; Jun and Fougeron, 2000: 230; Féry, 2014). In Chinese, tonal compression applies to lexical tones: the hat contour (T2-T4) is realized with reduced pitch range when the object follows the focus, as already reported in instrumental phonetic studies (Xu, 1999: 69; Chen, 2010; Greif, 2012: 82–88, 110–116). This result is not generalizable for all tone languages but confirms the view that Mandarin Chinese belongs to the subclass of tonal languages that have postfocal tonal compression (Xu et al., 2012). Our conclusions are summarized in (11).

(11) Postfocal tonal leveling

The postfocal domain is prosodically leveled out in all languages:

(a) English and German: the postfocal material is deaccented;
(b) French and Chinese: the available tonal events (edge tones in French, lexical tones in Chinese) are visible after the focus but tonally compressed.

The effects of second-occurrence focus are only confirmed by a significant Construction × Focus × Time interaction in German. This result is in line with previous studies on second-occurrence focus in non-final contexts, in particular Féry and Ishihara (2006) on German. We refrain from any strong statement about a difference between languages with respect to second-occurrence foci: prenuclear accents are optional in general and a prosodic marking of second occurrence focus is not mandatory in these constructions, since it is already expressed through the cleft construction. Nevertheless, the fact that the only language for which we obtained evidence for prosodic reflexes of second-occurrence focus is German is in line with the view that signaling second-occurrence focus entails signaling focus. Languages with a contrast between accent types for the expression of focus are more likely

to employ this contrast for second-occurrence foci as well.

Finally, the prosodic devices that can be used for signaling focus are equally used in canonical and cleft constructions. The interaction effects of CONSTRUCTION × TIME in the subject region in English and Chinese are accounted by specific properties of the constructions at issue. In both languages, cleft constructions show a tonal event that is immediately left-adjacent to the first syllable: in English it is a pitch accent aligned with the pronoun *it* (see **Figure 2**), while in Chinese it is the falling tone (tone 4) on the copula *shi* (see **Figure 6**). The reflex of these accentual events on the immediately adjacent high target is that the f_0 rise starts later and from a higher pitch level in these constructions, which results into the significant interaction effect in these languages. Hence, this effect relates to language-specific properties of the material and is not informative for a difference between canonical and cleft constructions in terms of the prediction in (9b). An interaction effect of FOCUS × TIME (across constructions) is available in all languages, both in the analyses of subjects (**Table 1**) as well as in the analyses of objects (**Table 2**). We conclude from these facts that all languages have the potential to realize different prosodic structures depending on focus with canonical and cleft constructions.

CONTEXTUAL FELICITY OF SYNTACTIC CONSTRUCTIONS

Aims

The aim of the present experiment is to test whether the contextual felicity of cleft constructions with a contrastive focus in the cleft clause depends on the prosodic typology. For this purpose, we collected judgments of the appropriateness of target utterances in certain contexts by means of written questionnaires. The typological distinction between plastic and non-plastic languages (based on the flexibility of nuclear-accent placement) predicts an advantage for languages such as German and English. However, our study on speech production revealed that focus is associated with various reflexes of prosodic prominence in all examined languages (Section 2.4).

Method

Participants

The participants were explained that their participation was voluntary and that the data will be used in anonymized form for research purposes and will be made available through the internet. Participants signed a written consent form (translated into their native language). Participants were paid for their contribution to the experiment studies. This experiment was conducted independently of the experiment on the prosodic reflexes of focus in Section 2 (the participant samples are different). While sex was controlled in the prosodic study, there was no reason to control sex in the study on contextual felicity: English (n = 32, female = 14, age range = 18–38, average = 24.3; collected in London), German (n = 32, female = 26, age range = 19–32, average = 22.8; collected in Bielefeld), French (n = 32, female = 18, age range: 18–46 = average 30.1; collected in Lyon), and Chinese (n = 32, female = 28, age range = 18–45, average = 21.9; collected in Beijing).

Factorial Design

Participants were presented with a written dialogue containing a context of a speaker A and two alternative target utterances of speaker B (either B_1 or B_2) and were instructed to estimate the contextual felicity of the target utterances with respect to the context (see details in Section 3.2.4); see (12).

(12) A: *They auctioned off many things today. Peter sold the bicycle.*
B_1: *No, John sold the bicycle.*
B_2: *No, it's John that sold the bicycle.*

In order to assess the effect of correction in modulating the association of certain constructions with certain focus structures, we designed an experiment with the factors CONSTRUCTION, FOCUS, and CONTEXT; see (13). The context created by speaker A contained an initial sentence that was kept constant across experimental conditions and was used in order to create a richer situation in which alternative focus structures of the final utterance can be accommodated. The target utterances illustrate the two levels of the factor CONSTRUCTION: canonical sentence in B_1 or cleft construction in B_2. Their FOCUS (subject or object) depends on their relation to the antecedent statement (last utterance of A). The form of the antecedent statement determines the CONTEXT, being either a canonical or a cleft construction. Cleft constructions are expected to be accommodated in this context by assuming a richer Common Ground: "it's Peter that sold the car" implies that 'that somebody sold the car' is shared knowledge between the interlocutors and the contribution of this utterance to the discourse is that 'Peter (and not somebody else) did it.'

(13) Factorial design of the contextual felicity study

(a) FOCUS: subject, CONTEXT: canonical
A: *They auctioned off many things today. Peter sold the bicycle.*
B_1: *No,* $[John]_F$ *sold the bicycle.*
B_2: *No, it's* $[John]_F$ *that sold the bicycle.*

(b) FOCUS: subject, CONTEXT: cleft
A: *They auctioned off many things today. It's Peter that sold the car.*
B_1: *No,* $[John]_F$ *sold the bicycle.*
B_2: *No, it's* $[John]_F$ *that sold the bicycle.*

(c) FOCUS: object, CONTEXT: canonical
A: *They auctioned off many things today. John sold the car.*
B_1: *No, John sold* $[the\ bicycle]_{rmF}$.
B_2: *No, it's John that sold* $[the\ bicycle]_F$.

(d) FOCUS: object, CONTEXT: cleft
A: *They auctioned off many things today. It's John that sold the car.*
B_1: *No, John sold* $[the\ bicycle]_F$.
B_2: *No, it's John that sold* $[the\ bicycle]_F$.

Material

The conditions in (13) were implemented in 16 items with different lexicalizations of simple transitive clauses; see item list in **Supplementary Material**, Section 2.2. The native speakers that created the material were encouraged to create situations that are maximally natural in the languages at issue and contain the target structures – without being necessarily literal translations of the English version. The target utterances contain either canonical or cleft constructions; see discussion of the constructions in Section 2.2.2. The cleft constructions used in the present experiment are illustrated in (14).

(14) (a) German

Nein,	*es*	*ist*	*Johannes*,	*der*	*das*
NEG	it	be:3SG	Johannes	who	DEF
Auto	*verkauft*	*hat.*			
car	sell:PTCP	have:3SG			

'No, it was Johannes that sold the car.'

(b) French

Non,	*c'*	*est*	*Jean*	*qui*	*a*
NEG	this	be:3SG	Jean	who	have:3SG
vendu	*le*	*vélo.*			
sell:PTCP	DEF	bicycle			

'No, it's Jean that has sold the bicycle.'

(c) Chinese

Bù	*duì,*	*shì*	*wáng*	*nán*	*mài*	*le*
NEG	correct	be	Wang	Nan	sell	PFV
zìxíngchē.						
bicycle						

'No, it was Wang Nan that sold the bicycle.'

The material was presented in written questionnaires. The background assumption is that participants consider the range of prosodic structures that are active in memory in order to evaluate the felicity of the written utterance in a certain context. Hence, a target utterance may be judged as not felicitous if the participants cannot find an implicit prosodic structure that renders the realization of the utterance congruent with the given context, which may either mean that an appropriate prosodic structure is marginal in language use or that it is not considered sufficient to accommodate the utterance in the given context.

Procedure

The participants were presented with a context A and two target utterances B_1/B_2 as in (12), whereby each target utterance was accompanied by a scale from 1 to 7 (see Destruel et al., 2019 for a previous study on contextual felicity with a 1–7 scale). The participants were instructed to evaluate the extent that each contribution B_1/B_2 was felicitous regarding the context A. The level 1 of the scale stands for 'the contribution B does not fit to the context A' and the level 7 for 'the contribution B fits to the context A.' The order of presentation of canonical and cleft constructions was randomized in the trials. The reasoning for presenting both utterances in the same trial was motivated by the aim to understand native speakers' intuitions when considering the paradigmatic alternatives for expressing the same propositional content in certain contexts. We decided to not elicit a single judgment of the comparison between both options since it would not be informative for the felicity of the individual options (two options with the same score could be both felicitous or both non-felicitous).

The material was distributed into four different lists, with each list using each of the 16 items once in a Latin square design. The experimental items of each questionnaire were mixed with fillers at a 1 (targets): 3 (fillers) proportion. Each list was presented to eight participants, which renders (4 lists × 8 participants =) 32 participants (per language). In sum, the dataset of each language contains 16 items × 2 target utterances × 32 participants = 1024 judgments of the contextual felicity of target utterances in context.

Data Analysis

The response categories of a Likert scale form an ordinal variable, most importantly because it cannot be warranted that the differences between the numeric values of the 1-to-7 scale reflect equal distances of the estimations of contextual felicity (Bürkner and Vuorre, 2019: 77). We assessed the statistical significance of the examined effects by fitting cumulative link mixed-effects models for ordinal regression (function *clmm* of the package *ordinal* in R; Christensen, 2019). These models estimate the probability of each increase between the levels of the ordinal scale by adding a corresponding intercept to the regression model coefficients (Bürkner and Vuorre, 2019: 79).

The dependent variable of the ordinal regression was the CONTEXTUAL FELICITY, which contains an ordinal scale of ratings (1 to 7). The factors of interest were FOCUS (referring to the focus domain: level 0 = object; level 1 = subject) and CONSTRUCTION (referring to the structure of the target utterance; level 0 = canonical; level 1 = cleft) and CONTEXT (referring to the structure of the last utterance in the context; level 0 = canonical; level 1 = cleft). The random-effects structure contained intercepts for PARTICIPANTS and ITEMS as well as by-PARTICIPANTS and by-ITEMS random slopes of the fixed effects, which converges in all languages; see formula in **Supplementary Material**, Section 1.2 and see text footnote 6 on the notion of convergence. The maximal fixed-effects structure (FOCUS × CONSTRUCTION × CONTEXT) was reduced with model comparison based on a backward-elimination procedure by means of Likelihood Ratio Tests. The random-effects structure was kept maximal in all compared models (Barr et al., 2013).

Predictions

The aim of this study is to test whether the contextual felicity of cleft constructions with a focus in the cleft clause is equally felicitous across languages or whether it depends on the prosodic type of the language at issue. The crucial effect for this question is the threefold interaction FOCUS × CONSTRUCTION × CONTEXT, which indicates that the effect of the cleft-focus principle that is reflected in the FOCUS × CONSTRUCTION interaction (subject clefts are felicitous in subject focus) is modulated by CONTEXT, such that the felicity of subject clefts with object focus increases if the antecedent statement is a subject cleft.

The use of cleft constructions with a focus domain in the cleft clause is the configuration introduced in (2), which requires an expression of the focus by prosodic means that deviates from the preferred focus structure of a cleft construction. Hence, the cross-linguistic question is whether this threefold interaction will appear in all languages or in a phonologically determined subtype of languages (see discussion in Section 1). The null hypothesis is that this configuration will be possible in all languages of our sample, since they have been shown to have prosodic reflexes of focus (see Section 2.4). However, we have seen that the observed effects come from different types of phenomena: English and German use certain pitch accents that unambiguously determine the intonational nucleus of the utterance (nuclear accents) and correspondingly the focus placement, while the effects in French and Chinese are general indicators of prosodic prominence (not reserved for focus), which maximize the demarcation of prosodic constituents (edge tones in French) or tonal targets (lexical tones in Chinese). If this difference is relevant for expressing different focus domains with one and the same syntactic construction, then we should obtain a three-way interaction in German/English and not so in French and Chinese. The predictions of our study are summarized in (15).

(15) Predicted interaction effects on CONTEXTUAL FELICITY

(a) CONSTRUCTION × CONTEXT: an advantage for using the same construction in the target utterance (CONSTRUCTION) and the antecedent utterance (CONTEXT) is predicted by syntactic priming in general and additionally by the preference for structural parallelism between corrective statements and their antecedents (Van Leusen, 2004: 437; Clifton and Frazier, 2016; see Section 1).

(b) FOCUS × CONSTRUCTION: an advantage for subject clefts in subject focus is predicted by the cleft-focus principle (Rochemont, 1986: 133; see Section 1).

(c) FOCUS × CONSTRUCTION × CONTEXT: the FOCUS × CONSTRUCTION interaction is modulated by CONTEXT; in particular a second-occurrence focus is expected to result in an advantage for a focus within a cleft clause if the antecedent statement has the same syntactic construction. The three-way interaction is expected to appear only in English/German if reflexes of second-occurrence focus only apply to languages with unambiguous cues of the intonational nucleus, or in all languages otherwise.

Results

The averaged judgments reveal a major difference between plastic (English and German) and non-plastic (French and Chinese) languages (see **Figure 7**). The canonical sentences in English and German are almost equally felicitous in subject and object focus, while subject cleft constructions are more felicitous with subject focus. The interesting result is the contextual felicity of subject clefts in an object focus context. In this case, we observe a difference depending on the structure of the context utterance: if this utterance is a cleft (gray dots), then the contextual felicity of the cleft construction increases. Finally, cleft constructions in German generally obtain lower judgments than the same constructions in English. French and Chinese differ. Canonical target utterances (solid lines) show an effect of FOCUS, such that the contextual felicity decreases with subject focus. The judgments of cleft constructions (dashed lines) show the mirror image, rendering a disordinal interaction: subject clefts in French and Chinese are highly felicitous with subject foci and not so with object foci. Crucially, the context utterance has a marginal role in these languages. In the non-canonical constructions, we observe a slight advantage of contextual felicity with object focus, when the same construction is presented in the context: see difference between gray and black dots in object focus with non-canonical constructions (dashed lines).

Cumulative link mixed-effects models for ordinal regression were fitted on the ratings of contextual felicity of each language separately (see Section 3.2.5). The estimates assessing the effect of each increase between the levels of the Likert scale (1 to 7) are listed in **Supplementary Material**, Section 4. These intercept values are added to the model coefficients rendering the logit of the probability that the outcome exceeds a certain threshold. The relevant information for our data is the general tendency that is captured by the average estimates in (16): the estimates of edge values have greater differences (2|3 minus 1|2 renders 1.6 on average; 6|7 minus 5|6 renders 1.4) than estimates of middle values (all further average differences are below 1). Hence, the estimates reveal that the levels of the Likert scale are not equidistant.

(16) Average threshold intercepts of the increases in the ordinal-scale ratings (1 to 7)

1|2, average estimate: −6.177
2|3, average estimate: −4.767
3|4, average estimate: −3.834
4|5, average estimate: −3.136
5|6, average estimate: −2.161
6|7, average estimate: −0.510

The coefficients of the fixed factors (**Table 3**) show that all models contain a significant effect of CONSTRUCTION × FOCUS, indicating that cleft constructions with a subject pivot reach a better fit (compared to the canonical constructions) if the subject is focused. While the threefold interaction (CONSTRUCTION × FOCUS × CONTEXT) is significant in English and German, it is not so in French and Chinese. The negative interaction effect indicates that the effect of the cleft-focus principle (CONSTRUCTION × FOCUS) is modulated by CONSTRUCTION, such that contextual felicity increases in the baseline of the FOCUS factor (object focus) with subject clefts in the target utterance (CONSTRUCTION) and the context (CONTEXT). In English, German, Chinese, we obtained a positive interaction

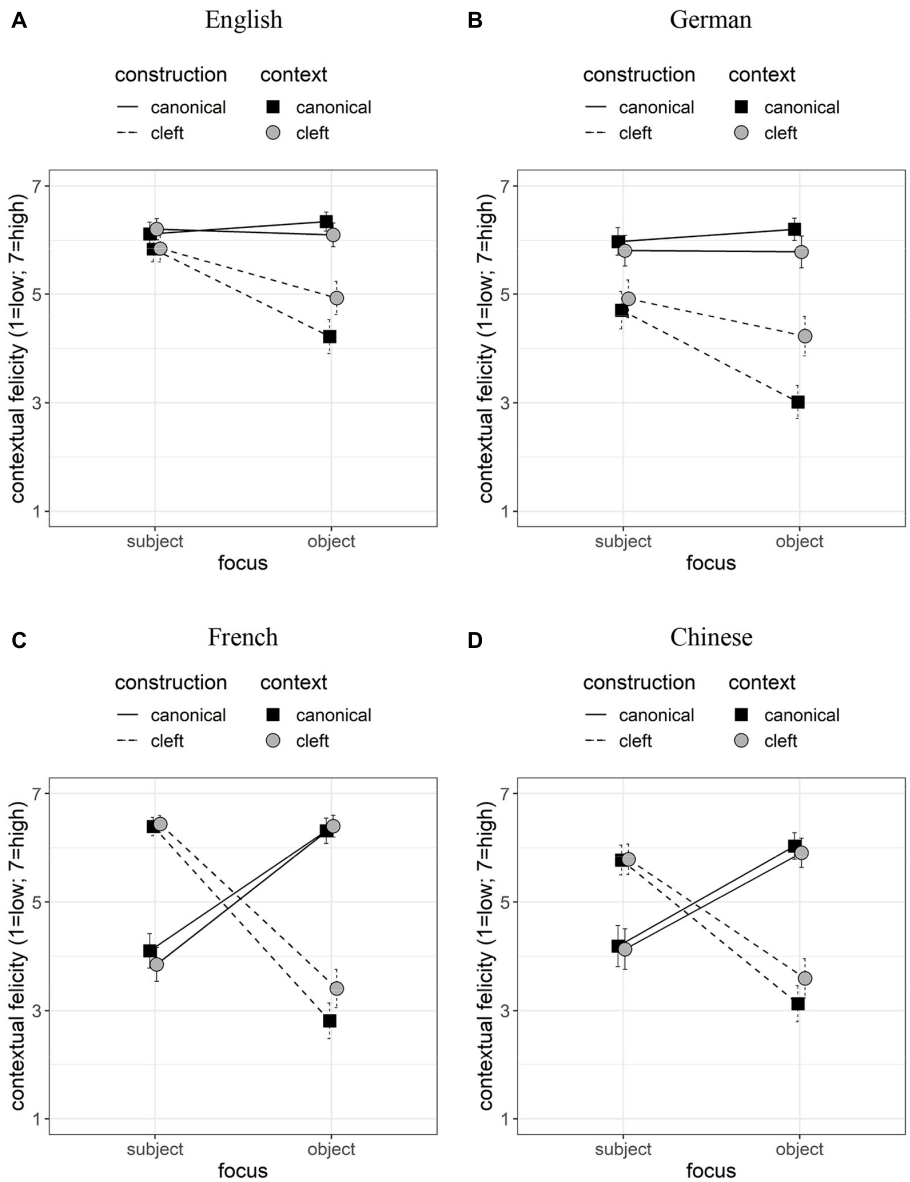

FIGURE 7 | (A–D) Contextual felicity (Y-bars: confidence intervals with probability 0.95).

CONSTRUCTION × CONTEXT, which means an advantage for using the same construction in the target as in the immediate context. The model of maximal fit in French contains a negative FOCUS × CONTEXT interaction, reflecting the fact that subject clefts obtain higher ratings than canonical constructions under subject focus but lower ratings under object focus (see **Figure 7C**). The main effect of CONSTRUCTION is negative in all languages, since canonical constructions achieve higher scores than cleft constructions across contexts. The main effect of FOCUS is also negative, reflecting a subject vs. non-subject asymmetry in focus. The factor CONTEXT has a significant negative effect in German (cleft constructions in the context utterance are judged to be less felicitous) and a significant positive effect in French (cleft constructions in the context utterance are judged to be more felicitous).

Discussion

Our findings show a major contrast between English and German on the one side and French and Chinese on the other. The major issue is the difference between languages: the felicity of the same constructions is judged differently under identical treatments depending on language. It is not the case that constructions with low scores are impossible. For instance, canonical sentences with a subject focus are possible albeit not the preferred option in French (Destruel, 2013: 162, Destruel, 2016: 310); cleft constructions with a focus in the embedded

TABLE 3 | Cumulative link models on the ordinal scale of contextual felicity.

Language	Factor	β	SE	*z*	*p* <	Log-Likelihood Test	
						χ^2	*p* <
English	Construction (cleft)	−3.817	0.402	−9.498	0.001		
	Focus (subject)	−0.269	0.342	−0.788	–		
	Context (cleft)	−0.478	0.326	−1.466	–		
	Construction × Focus	2.767	0.388	7.127	0.001		
	Construction × Context	1.487	0.37	4.017	0.001		
	Focus × Context	0.736	0.393	1.872	–		
	Construction × Focus × Context	−1.592	0.522	−3.048	0.01	9.351	0.01
German	Construction (cleft)	−4.288	0.377	−11.375	0.001		
	Focus (subject)	−0.351	0.339	−1.035	–		
	Context (cleft)	−0.593	0.271	−2.187	0.05		
	Construction × Focus	2.3	0.361	6.376	0.001		
	Construction × Context	1.871	0.356	5.26	0.001		
	Focus × Context	0.353	0.371	0.952	–		
	Construction × Focus × Context	−1.288	0.494	−2.605	0.01	6.813	0.01
French	Construction (cleft)	−5.059	0.308	−16.424	0.001		
	Focus (subject)	−3.579	0.266	−13.434	0.001		
	Context (cleft)	0.522	0.184	2.839	0.01		
	Construction × Focus	8.65	0.377	22.971	0.001	798.87	0.001
	Focus × Context	−0.629	0.247	−2.547	0.05	6.506	0.05
Chinese	Construction (cleft)	−3.571	0.367	−9.732	0.001		
	Focus (subject)	−2.378	0.221	−10.744	0.001		
	Context (cleft)	−0.197	0.181	−1.091	–		
	Construction × Focus	5.358	0.293	18.311	0.001	410.88	0.001
	Construction × Context	0.604	0.242	2.494	0.05	6.25	0.05

clause are possible in French (Dufter, 2009: 105, 114) and Chinese (Yan and Calhoun, 2019).

A first difference between languages relates to the felicity conditions of canonical sentences. The results for English and German indicate that canonical sentences are contextually unrestricted, as expected for languages with flexible placement of the nuclear accent. This result is in line with statements about the optionality of cleft constructions in English: cleft constructions are optionally used to express focus since the focus can be unambiguously identified through the phonological form (Kiss, 1998: 268). Crucially, the judgments differ for French and Chinese, in which case the contextual felicity of canonical sentences radically drops if the subject is focused. This result confirms intuitions about a constraint against focus on preverbal subjects in French (Lambrecht, 2001: 492; Hamlaoui, 2007), which is accounted for by the general preference of French for aligning the focus with the right edge of the intonation phrase (Féry, 2013: 698). Studies on speech production show that subjects are mostly focused through cleft constructions (Destruel, 2013: 162; Destruel, 2016: 310). Similar effects are reported for Chinese: canonical SVO sentences are typically mapped on a Topic-Comment articulation, which has specificity effects on the interpretation of preverbal subjects (Huang et al., 2009: 200). These effects are part of a cross-linguistic preference to map subjects on topics, which results to a subject vs. non-subject asymmetry in marking focus, such that subject focus often appears with additional marking (Lambrecht, 2001: 490; Hartmann and Zimmermann, 2007; Zerbian, 2007: 336; Destruel, 2016: 304). In our findings, the preference against a subject focus in canonical constructions depends on language; see (17).

(17) Contextual conditions for canonical constructions

In English and German, canonical constructions are judged as equally felicitous in subject and object focus contexts. In French and Chinese, canonical constructions are judged to be less felicitous in subject focus contexts.

The felicity of the subject clefts under subject focus confirms the association of the pivot of cleft constructions with contrastive focus. The only language in which subject clefts are judged differently from canonical sentences in subject focus is German; this finding is in line with the fact that cleft constructions in German are less frequent in corpora than the same constructions in Romance languages (Dufter, 2009: 90). The crucial finding of this study is the significant Focus × Context interaction in English and German – in contrast to French and Chinese. This result confirms the expectations concerning the flexibility in nuclear-accent placement. If the placement of the nuclear accent is flexible, as it is assumed for English and German, the use of cleft constructions with a later nuclear accent has an advantage in contextual felicity (in appropriate contexts); see (18). The question is how this finding combines with the results of the speech production study (see general discussion in Section 4).

(18) Contextual conditions for cleft constructions

> Across languages, the felicity of subject clefts increases if the pivot is focused. In English and German, but not in French and Chinese, the contextual felicity of cleft constructions with a focus in the cleft clause increases when the context motivates the use of the cleft construction (in our manipulation by the structural parallelism of a corrective statement to an antecedent statement).

A final note is due concerning the main effects of the fixed factors. The negative effect of CONSTRUCTION reflects the fact that cleft constructions are contextually restricted in comparison to canonical constructions. The negative effect of FOCUS reflects the subject vs. non-subject asymmetry in focus: the preferred option in discourse is a focus on objects or further verbal complements, while focus on subjects is the least preferred case (Lambrecht, 2001: 490; Hartmann and Zimmermann, 2007; Zerbian, 2007: 336; Destruel, 2016: 304). The effect of CONTEXT is negative in German and positive in French. This effect relates to the felicity of the context utterance (independent of the corresponding target). In a language such as German, in which cleft constructions are highly marked and rare in discourse, the presence of a cleft construction without an obvious contextual trigger in the context, is judged to be suboptimal. The opposite effect appears in French, a language in which subject clefts are very frequent in discourse and may appear without requiring a focus on the subject.

GENERAL DISCUSSION AND CONCLUSION

The presented studies examined hypotheses concerning the interaction of prosody and syntax in the expression of information structure. A reasonable hypothesis is the idea that the reflexes of focus on different layers of grammar are complementary: the syntactic expression of discourse notions is motivated if they are not expressed by prosody. Several versions of this complementarity have been used to explain cross-linguistic differences in the use of cleft constructions and other syntactic operations (Vallduví and Engdahl, 1996: 497; Zubizarreta, 1998: 89; Samek-Lodovici, 2005; discussion in Section 1).

We compared the use of cleft constructions in contexts licensing contrastive focus in four languages with different prosodic properties: two languages allowing for flexible nuclear-accent placement (English, German), a language that relies on prosodic phrasing (French) and a language with lexical tones (Mandarin Chinese). A speech production study has shown that all languages show prosodic reflexes of focus: the contrast between nuclear and prenuclear accents in English and German, initial rises demarcating prosodic constituents in French, increase of the distinctness of tonal targets in Chinese; see (10). The postfocal domain is also affected by effects of prosodic leveling in all languages, deaccenting in English and German and compression of edge and lexical tones in French and Chinese; see (11).

In a study on contextual felicity, we examined whether canonical and cleft constructions can be used in contexts licensing a focus on the subject and on the object. This study reveals a typological distinction between two classes of languages. Canonical constructions are contextually unrestricted in English and German, but less felicitous with subject focus in French and Chinese; see (17). Cleft constructions are more felicitous with a nucleus in the cleft clause, if the context motivates the use of the cleft as a corrective statement that relates to a cleft construction within the antecedent statement: crucially, this effect was statistically confirmed for English and German, but not for French and Chinese; see (18). This result has repercussions for the interpretation of the data collected in speech production. Our basic assumption regarding the contextual felicity judgments is that the participants evaluate a syntactic construction as felicitous in the presented context if they may recall a prosodic structure that fits to this context. Low scores of contextual felicity indicate that this structure is not active in memory, which may mean that this construction is marginal in language use. We conclude from the findings of this study that the French and Chinese data with focus on the cleft clause that we collected in speech production are marginal in language use.

Previous studies have shown that cleft constructions occur in a wider array of contexts in French and Chinese than in English and German. Cleft constructions require a contrastive context in English, while in French the same constructions also appear in non-contrastive contexts, such as answers to *wh*-questions (Skopeteas and Fanselow, 2010). In the same vein but based on acceptability and response time data, Destruel and De Veaugh-Geiss (2018) conclude that an exhaustive inference is part of the default interpretation of English clefts, which does not hold true in French. Beyond narrow focus on the subject, cleft constructions also appear in contexts in which the subject is part of a broader focus domain, such as sentence focus in French (Lambrecht, 2001; Karssenberg and Lahousse, 2018) and Chinese (Paul and Whitman, 2008: 426; Von Prince, 2012: 342), which does not apply to English and German (Dufter, 2009: 114). These comparisons may lead to the conclusion that cleft constructions are semantically bleached in French and Chinese and not so in English and German, such that they appear in a wider array of contexts in the latter type of languages than in the former.

However, our findings identified a type of context (focus within the cleft clause) in which cleft constructions have an advantage only in English and German. Hence, the array of contexts of English and German clefts is not a proper subset of the array of contexts of French and Chinese clefts, which is against the prediction of bleaching. A view from prosodic typology is relevant for understanding this difference, since the crucial context has exactly the property of requiring a prosodic marking of focus. A prosodic account also explains the occurrence of clefts in a wider array of focus types in French and Chinese: if cleft formation is the only means to focus a subject, as in these languages, then it follows that it will appear in any context involving subject focus, without contextual restrictions such as contrastivity or exhaustivity.

The question is how this typological distinction relates to the prosodic findings of the speech production study, which has shown that all languages have some prosodic reflexes of

information structure. Significant prosodic effects of focus were found in all four languages in line with earlier findings; see Vander Klok et al. (2018) on reflexes of different types of focus in French, Greif (2012) and Ouyang and Kaiser (2015) for the impact of contrastive focus in Chinese as well as Yan and Calhoun (2019) for effects of prosodic prominence in Chinese on interpretation (invoking alternatives). Moreover, our study shows that these prosodic effects equally appear in the constructions at issue (canonical and cleft constructions).

The crucial observation is that the prosodic reflexes found in these languages come from different classes of phenomena, as outlined in (10). In English and German, the focus determines the placement of the nuclear stress, which is reflected on the contrast between nuclear and prenuclear accents (area of interest: subject) or the contrast between nuclear accents and deaccenting (area of interest: object). On the other hand, focus is reflected on events that reinforce the articulation of prosodic constituents in French (initial rise) or the articulation of lexical tones in Chinese (distinctness of tonal targets). These classes of phenomena have distinct semantic-pragmatic import, such that only the first class of phenomena is an unambiguous indicator of the focus structure of the utterance. While nuclear-accent placement is directly determined by the focus structure, effects on the articulation of prosodic constituents or lexical tones may be employed in order to draw the attention of the hearer to a certain partition of the utterance without being unambiguously associated with a focus structure. The cross-linguistic differences in the flexibility of using canonical and cleft constructions in various contexts is straightforwardly accounted for by this distinction: our findings show that determining a layer of focus structure that is independent from syntax has an advantage in languages of the former type.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

MG and SS designed the experiments, analyzed the data, and prepared a first draft together. MG supervised the data collection and data processing in the individual languages. SS prepared the final manuscript. Both authors reviewed the manuscript and approved the final version to be published.

FUNDING

This study was supported by the Collaborative Research Center 673, *Alignment in Communication*, Bielefeld University, funded by the German Research Foundation (DFG).

ACKNOWLEDGMENTS

We are grateful to the cooperation partners who helped us to create the language-specific materials and conducted the experimental studies: Georg Höhn and Joseph DeVeaugh-Geiss (English), Sabrina Meier (German), Adrien Déchaine (French), Bei Wang (Chinese).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.648478/full#supplementary-material

REFERENCES

Alter, K., Mleinek, I., Rohe, T., Steube, A., and Umbach, C. (2001). Kontrastprosodie in sprachproduktion und -perzeption. *Linguistische Arbeitsberichte* 77, 59–79.

Arnhold, A. (2021). Prosodic focus marking in clefts and syntactically unmarked equivalents: prosody – syntax trade-off or additive effects? *J. Acoust. Soc. Am.* 149, 1390–1399. doi: 10.1121/10.0003594

Barr, D. J. (2008). Analyzing 'visual world' eyetracking data using multilevel logistic regression. *J. Memory Lang.* 59, 457–474. doi: 10.1016/j.jml.2007.09.002

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

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Statist. Software* 67, 1–48. doi: 10.18637/jss.v067.i01

Baumann, S., and Ishihara, S. Sh (2016). "Second occurrence focus," in *The Oxford Handbook of Information Structure*, ed. C. Féry (Oxford: Oxford University Press), 483–502.

Baumann, S., and Riester, A. (2013). Coreference, lexical givenness, and prosody in German. *Lingua* 136, 16–37. doi: 10.1016/j.lingua.2013.07.012

Baumann, S., and Winter, B. (2018). What makes a word prominent? predicting untrained German listeners' perceptual judgments. *J. Phonet.* 70, 20–38. doi: 10.1016/j.wocn.2018.05.004

Baumann, S., Mertens, J., and Kalbertodt, J. (2021). The influence of informativeness on the prosody of sentence topics. *Glossa* 35, 1–28. doi: 10.16995/glossa.5871

Beaver, D., Clark, B. Z., Flemming, E., Jaeger, T. F., and Wolters, M. (2007). When semantics meets phonetics: acoustical studies of second-occurrence focus. *Language* 83, 245–276. doi: 10.1353/lan.2007.0053

Boersma, P., and Weenink, D. (2020). *Praat: doing Phonetics by Computer [Computer program]. Version 6.1.37*. Available online at: http://www.praat.org/ (accessed 28 December, 2020).

Büring, D. (2015). A theory of second occurrence focus. *Lang. Cogn. Neurosci.* 30, 73–87. doi: 10.1080/01690965.2013.835433

Bürkner, P.-C., and Vuorre, M. (2019). Ordinal regression models in psychology: a tutorial. *Adv. Methods Pract. Psychol. Sci.* 2, 77–101. doi: 10.1177/2515245918823199

Calhoun, S. (2010). The centrality of metrical structure in signaling information structure: a probabilistic perspective. *Language* 86, 1–42. doi: 10.1353/lan.0.0197

Cheng, L. (2008). Deconstructing the shì . . . de construction. *Linguistic Rev.* 25, 235–266.

Chen, Y. (2010). Post-focus F0 compression: now you see it, now you don't. *J. Phonet.* 38, 517–525. doi: 10.1016/j.wocn.2010.06.004

Chen, Y., and Gussenhoven, C. (2008). Emphasis and tonal implementation in standard Chinese. *J. Phonet* 36, 724–746. doi: 10.1016/j.wocn.2008.06.003

Christensen, R. H. B. (2019). *Ordinal - Regression Models for Ordinal Data*. R package version 2019.12-10.

Clifton, C. Jr., and Frazier, L. (2016). Focus in corrective exchanges: effects of pitch accent and syntactic form. *Lang. Speech* 59, 544–561. doi: 10.1177/0023830915623578

Cole, J., Hualde, J. I., Smith, C. L., Eager, C., Mahrt, T., Napoleão, et al. (2019). Sound, structure and meaning: the bases of prominence ratings in English, French and Spanish. *J. Phonet.* 75, 113–147. doi: 10.1016/j.wocn.2019.05.002

Delais-Roussarie, E., Post, B., Avanzi, M., Buthke, C., Di Cristo, A., Feldhausen, I., et al. (2015). "Intonational phonology of French: developing a ToBI system for French," in *Intonation in Romance*, eds S. Frota and P. Prieto (Oxford: Oxford University Press), 63–100. doi: 10.1093/acprof:oso/9780199685332.003.0003

Delin, J. (1992). Properties of it-cleft presupposition. *J. Semant.* 9, 289–306. doi: 10.1093/jos/9.4.289

Delin, J. (1995). Presupposition and shared knowledge in it-clefts. *Lang. Cogn. Proc.* 10, 97–120. doi: 10.1080/01690969508407089

Destruel, E. (2013). *The French c'est-cleft: Empirical Studies of its Meaning and use* thesis, Ph.D. dissertation. Austin, TX: University of Texas.

Destruel, E. (2016). Focus marking asymmetries in Colloquial and Standard French: a stochastic optimality-theoretic account. *J. French Lang. Stud.* 26, 299–326. doi: 10.1017/S0959269515000265

Destruel, E., and De Veaugh-Geiss, J. (2018). On the interpretation and processing of exhaustivity: evidence of variation in English and French clefts. *J. Pragmat.* 138, 1–16. doi: 10.1016/j.pragma.2018.09.009

Destruel, E., Beaver, D. I., and Coppock, E. (2019). It's not what you expected! the surprising nature of cleft alternatives in french and english. *Front. Psychol.* 10:1400. doi: 10.3389/fpsyg.2019.01400

Di Cristo, A., and Jankowski, L. (1999). Prosodic organisation and phrasing after focus in French. *ICPhS* 99, 1565–1568.

Drenhaus, H., Zimmermann, M., Vasishth, and Sh. (2011). Exhaustiveness effects in clefts are not truth-functional. *J. Neurolinguist.* 24, 320–337.

Duanmu, S. (2007). *The Phonology of Standard Chinese*, 2nd Edn. Oxford: Oxford University Press.

Dufter, A. (2009). "Clefting and discourse organization - comparing Germanic and Romance," in *Focus and Background in Romance Languages*, eds A. Dufter and D. Jacob (Amsterdam, PA: John Benjamins), 83–121. doi: 10.1075/slcs.112.05duf

Féry, C. (2013). Focus as prosodic alignment. *Nat. Lang. Linguistic Theory* 31, 683–734. doi: 10.1007/s11049-013-9195-7

Féry, C. (2014). "Final compression in French as a phrasal phenomenon," in *Perspectives on Linguistic Structure and Context: Studies in honor of Knud Lambrecht*, eds S. K. Bourns and L. L. Myers (Amsterdam, PA: John Benjamins), 133–156. doi: 10.1075/pbns.244.07fer

Féry, C. (2017). *Intonation and Prosodic Structure*. Cambridge: Cambridge University Press.

Féry, C., and Ishihara, Sh (2006). "Interpreting second occurrence focus," in *Proceedings of the NELS 36*, eds Chr. Davis, Y. Zabbal, and A. R. Deal (Amherst, MA: GLSA), 371–384.

Féry, C., and Kügler, F. (2008). Pitch accent scaling on given, new and focused constituents in German. *J. Phonet.* 36, 680–703. doi: 10.1177/0023830916647204

Fischer, K. (2009). Cleft sentences: form, function, and translation. *J. Germanic Linguistics* 21, 167–191.

German, J., and D'Imperio, M. (2010). *Focus, Phrase Length, and the Distribution of Phrase-initial Rises in French. Speech Prosody 2010*. Chicago, IL: ISCA.

Gotzner, N. (2015). *Establishing Alternative Sets*. Thesis, Ph.D. dissertation. Berlin: Humboldt University.

Grabe, E., Kochanski, G., and Coleman, J. (2007). Connecting Intonation Labels to mathematical descriptions of fundamental frequency. *Lang. Speech* 50, 281–310. doi: 10.1177/00238309070500030101

Greif, M. (2012). *Corrective Focus in Mandarin Chinese: A Question of Belief?*. Munich: Lincom.

Greif, M., and Skopeteas, S. (2021). *Correction by Focus: Sound Files and Transcriptions (1.0)*. London: Zenodo. doi: 10.5281/zenodo.5092381

Grice, M., Baumann, S., and Benzmuller, R. (2005). "German intonation in autosegmental-metrical phonology," in *Prosodic Typology: The Phonology of Intonation and Phrasing*, ed. S.-A. Jun (Oxford: Oxford University Press), 55–83.

Grice, M., Baumann, S., and Jagdfeld, N. (2007). "Evidence for tonal identity from peak scaling under pitch span variations," in *Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS)* (Saarbrücken).

Gussenhoven, C. (2004). *The Phonology of Tone and Intonation*. Cambridge: Cambridge University Press.

Hamlaoui, F. (2007). "French cleft sentences and the syntax-phonology interface," in *Actes du congrès annuel de l'Association Canadienne de linguistique 2007*, ed. M. Radišić (Ottawa: Canadian Linguistic Association).

Hartmann, J. M. (2015). *The Syntax and Focus Structure of Specificational Copular Clauses and Clefts*. thesis, Tübingen: University of Tübingen.

Hartmann, K., and Zimmermann, M. (2007). "In place – out of place? focus strategies in Hausa," in *On Information Structure, Meaning and Form: Generalizations Across Languages*, eds K. Schwabe and S. Winkler (Amsterdam, PA: John Benjamins), 365–403. doi: 10.1075/la.100.20har

Hedberg, N. (1990). *Discourse Pragmatics and Cleft Sentences in English*. Thesis, Ph.D dissertation. Minneapolis, Min: University of Minnesota.

Hedberg, N. (2013). "Multiple focus and cleft sentences," in *Cleft Structures*, eds K. Hartmann and T. Veenstra (Amsterdam, PA: John Benjamins), 227–250. doi: 10.3389/fpsyg.2021.684639

Howell, J. (2011). "Second occurrence focus and the acoustics of prominence," in *Interfaces in Linguistics: New Research Perspectives*, eds R. Folli and Chr. Ulbrich (Oxford: Oxford University Press), 278–298. doi: 10.3109/14992027.2015.1030509

Huang, C.-T. J., Li, Y.-H. A., and Li, Y. (2009). *The Syntax of Chinese*. Cambridge: Cambridge University Press.

Isaacs, A., and Watson, D. G. (2010). Accent detection is a slippery slope: direction and rate of F0 change drives listeners' comprehension. *Lang. Cogn. Process* 25, 1178–1200. doi: 10.1080/01690961003783699

Jun, S. A., and Fougeron, C. (2000). "A phonological model of French intonation," in *Intonation. Text, Speech and Language Technology*, ed. A. Botinis (Dordrecht: Springer), 209–242. doi: 10.1177/0023830915587337

Karssenberg, L., and Lahousse, K. (2018). The information structure of French il y a clefts and c'est clefts: a corpus-based analysis. *Linguistics* 56, 513–548.

Kiss, K. É (1998). Identificational focus versus information focus. *Language* 74, 245–273. doi: 10.2307/417867

Krifka, M. (1997). "A second look at second occurrence expressions," in *Context-dependence in the Analysis of Linguistic Meaning*, eds H. Kamp and B. Partee (Stutgart: University of Stuttgart), 253–276.

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (2017). lmerTest package: tests in linear mixed effects models. *J. Statistical Software* 82, 1–26. doi: 10.18637/jss.v082.i13

Ladd, D. R. (2008). *Intonational Phonology*. Cambridge: Cambridge University Press.

Lambrecht, K. (2001). A framework for the analysis of cleft constructions. *Linguistics* 39, 463–516.

Liberman, M., and Pierrehumbert, J. (1984). "Intonational invariance under changes in pitch range and length," in *Language, sound, structure: Studies in Phonology Presented to Morris Halle by his Teacher and Students*, eds M. Aronoff and R. T. Oehrle (Cambridge, MA: MIT Press), 157–223.

Nolan, F. (2003). "Intonational equivalence: an experimental evaluation of pitch scales," in *Proceedings of the 15th International Congress of Phonetic Sciences* (Spain: Barcelona), 771–774.

Ouyang, I. Ch., and Kaiser, E. (2015). Prosody and information structure in a tone language: an investigation of Mandarin Chinese. *Lang. Cogn. Neurosci.* 30, 57–72. doi: 10.1080/01690965.2013.805795

Paul, W. (2015). *New Perspectives on Chinese Syntax*. Berlin: Mouton De Gruyter.

Paul, W., and Whitman, J. (2008). Shide focus clefts in Mandarin Chinese. *Linguistic Rev.* 25, 413–451.

Post, B. (2000). *Tonal and Phrasal Structures in French Intonation*. The Hague: Holland Academic Graphics.

Prince, E. F. (1978). A comparison of wh-clefts and it-clefts in discourse. *Language* 54, 883–906.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Repp, S. (2010). Defining 'contrast' as an information-structural notion in grammar. *Lingua* 120, 1333–1345. doi: 10.1016/j.lingua.2009.04.006

Rochemont, M. S. (1986). *Focus in Generative Grammar*. Amsterdam, PA: John Benjamins.

Rooth, M. (1996). "On the interface principles for intonation focus," in *Proceedings of Semantics and Linguistic Theory (SALT)*, eds T. Galloway and J. Spence (Ithaca, NY: Cornell University), 202–226.

Rooth, M. (2009). "Second occurrence focus and relativized stress F," in *Information Structure* eds M. Zimmermann and C. Féry (Oxford: Oxford University Press), 15–35. doi: 10.1093/acprof:oso/9780199570959.003.0002

Samek-Lodovici, V. (2005). Prosody-syntax interaction in the expression of focus. *Nat. Lang. Linguistic Theory* 23, 687–755. doi: 10.1007/s11049-004-2874-7

Skopeteas, S., and Fanselow, G. (2010). "Focus types and argument asymmetries: a cross-linguistic study in language production," in *Contrastive Information Structure*, eds C. Breul and E. Göbbel (Amsterdam, PA: John Benjamins), 169–197. doi: 10.1075/la.165.07sko

Steube, A. (2001). Correction by contrastive focus. *Theoretical Linguistics* 27, 215–249.

Titze, I. R., Baken, R. J., Bozeman, K. W., Granqvist, S., Henrich, N., Herbst, C. T., et al. (2015). Toward a consensus on symbolic notation of harmonics, resonances, and formants in vocalization. *J. Acoust. Soc. Am.* 137, 3005–3007. doi: 10.1121/1.4919349

Vallduví, E., and Engdahl, E. (1996). The linguistic realization of information packaging. *Linguistics* 34, 459–519.

Van Leusen, N. (2004). Incompatibility in context: a diagnosis of correction. *J. Semant.* 21, 415–441.

Vander Klok, J., Goad, H., and Wagner, M. (2018). Prosodic focus in English vs. French: a scope account. *Glossa* 3:71.

Veilleux, N., Shattuck-Hufnagel, S., and Brugos, A. (2006). *Transcribing Prosodic Structure of Spoken Utterances with ToBI. MIT Course Number 6.911*. Available online at: https://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-911-transcribing-prosodic-structure-of-spoken-utterances-with-tobi-january-iap-2006/index.htm (accessed November 4, 2021).

Von Prince, K. (2012). Predication and information structure in Mandarin Chinese. *J. East Asian Linguist.* 21, 329–366. doi: 10.1007/s10831-012-9091-x

Wang, B., and Xu, Y. (2011). Differential prosodic encoding of topic and focus in sentence-initial position in Mandarin Chinese. *J. Phonet.* 39, 595–611. doi: 10.1016/j.wocn.2011.03.006

Watson, D. G., Tanenhaus, M. K., and Gunlogson, Chr. A. (2008). Interpreting pitch accents in online comprehension: H* vs. L+H. *Cogn. Sci.* 32, 1232–1244. doi: 10.1080/03640210802138755

Xu, Y. (1999). Effects of tone and focus on the formation and alignment of f0 contours. *J. Phonet.* 27, 55–105.

Xu, Y., and Wang, Q. E. (2001). Pitch targets and their realization: evidence from Mandarin Chinese. *Speech Commun.* 33, 319–337. doi: 10.1121/1.2151799

Xu, Y., Chen, S.-W., and Wang, B. (2012). Prosodic focus with and without post-focus compression: a typological divide within the same language family? *Linguistic Rev.* 29, 131–147.

Yan, M., and Calhoun, S. (2019). Priming effects of focus in Mandarin Chinese. *Front. Psychol.* 10:1985. doi: 10.3389/fpsyg.2019.01985

Zerbian, S. (2007). "The subject/object asymmetry in Northern Sotho," in *On Information Structure, Meaning and Form: Generalizations across Languages*, eds K. Schwabe and S. Winkler (Amsterdam, PA: John Benjamins), 323–345.

Zubizarreta, M. L. (1998). *Prosody, Focus, and Word Order*. Cambridge, MA: MIT Press.

Conflict of Interest: 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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Greif and Skopeteas. 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) and the copyright owner(s) 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 November 2021
doi: 10.3389/fpsyg.2021.528862

Persistence Conditions of Institutional Entities: Investigating Copredication Through a Forced-Choice Experiment

Elliot Murphy[1,2,3]*

[1] Vivian L. Smith Department of Neurosurgery, McGovern Medical School, University of Texas Health Science Center, Houston, TX, United States, [2] Texas Institute for Restorative Neurotechnologies, University of Texas Health Science Center, Houston, TX, United States, [3] Department of Linguistics, University College London, London, United Kingdom

The conditions under which certain complex polysemous nominals can sustain coherent sense relations (informally, can "survive") is investigated through a two-alternative forced choice experiment. Written scenarios were constructed which permitted copredication, through which multiple, semantically different sense types are associated with a single nominal. Participants were presented with two scenarios involving a polysemous nominal (e.g., *bank*, *city*) and had to select which scenario (and, hence, which combination of predicates) appeared to be the most prototypical, faithful realization of the nominal. In order to achieve this, an additional manipulation was added, such that the number of senses hosted by each forced choice was either equal (2 senses choice vs. 2 senses choice) or unequal (1 sense choice vs. 2/3 senses choice). In order to address certain concerns in the literature about prototypicality, a core question addressed was whether the institutional sense of the nominals strongly determined the option chosen by participants, or whether the number of senses more strongly predicted this. It was found that the best predictor of sense "survival" was not sense frequency, but rather sense complexity or approximation to the institutional sense.

Keywords: copredication, polysemy, pragmatics, ship of theseus, persistence conditions, forced choice experiment

OPEN ACCESS

Edited by:
Pietro Perconti, University of Messina, Italy

Reviewed by:
Domenica Bruni, University of Messina, Italy
Guido Löhr, Eindhoven University of Technology, Netherlands

***Correspondence:**
Elliot Murphy
elliotmurphy91@gmail.com

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

Received: *22 January 2020*
Accepted: *08 November 2021*
Published: *30 November 2021*

Citation:
Murphy E (2021) Persistence Conditions of Institutional Entities: Investigating Copredication Through a Forced-Choice Experiment. Front. Psychol. 12:528862. doi: 10.3389/fpsyg.2021.528862

INTRODUCTION

In a recent discussion on the nature of consciousness, Chomsky (2018, p. 38) considers "one of the most ancient problems of philosophy: How can we cross the same river twice?" These are problems of "identity" and "individuation." Relatedly, Collins (2017, p. 680) presents an example demonstrating how river can license copredication through it being a geographical feature, an abstract relation or a body of water:

> "The Nile runs the length of Egypt and it serves as the most important trade route in the region as well as the source of irrigation for nigh-on all of Egypt's crop production."

This issue has been discussed mostly within the philosophical literature; witness Collins (2017, p. 686): "Thus, a group of people and a geographical area wildly dissociate in every conceivable sense save for them being referred to by *London*, say. We can kill the population of London, but not the area in south-east England. Equally, we can burn the city down while sparing the people, but rebuild the same city elsewhere, with a new population."

It has been argued that these issues pertaining to the semantic diversity of apparently simple entities may relate to the theme of polysemy. Cognitive models of polysemy have suggested that vagueness, polysemy and homonymy represent "a cline of diminishing schematicity and increasing instances of salience" (Lewandowska-Tomaszczyk, 2007, p. 139), such that polysemous senses bear a lower degree of salience than homonymous meanings do, and vague words have an even lower degree of salience for a particular meaning. For instance, *student* is not polysemic with respect to the distinction between *young student* and *old student* ("I gave the book to a student but not to a student" cannot refer to an old and young student, respectively); it is vague and unmarked. In a similar way that one would typically say "Milton Keynes is close to London" and not "London is close to Milton Keynes," due to proximation to a larger city being seen as a prototypical frame of geographical reference, it may be that there is an empirically detectable range of sense prototypicality in polysemy such that one sense may be seen as more essential to the polysemous nominal than others.

The main set of prototype effects are plotted in **Table 1**, loosely adapted from Lewandowska-Tomaszczyk (2007, p. 151) and modified to focus on the form of polysemy explored here. This shows how degrees of salience, clustering of senses, a lack of clear necessary and sufficient features, and variability in category membership are hallmarks of prototype effects. **Table 1** plots effects relating to *extensions* and *intensions*, which are, respectively, entities referred to by the concepts and senses which the concepts are composed of. These effects can in turn be categorized based on whether they effect the salience of the sense (how clear and prominent it is in a given interpretation) or its discreteness (related to issues of demarcation).

As the framework depicted in the above table would predict, it has been found that people tend not to categorize objects using necessary and sufficient features, but rather do so by comparing their similarity to a prototype of the candidate category (Rosch, 1975, 1978). As such, when judging whether a cluster of senses in a given context constitutes a *school* or not, comprehenders will presumably compare this structure to their stored prototypical representation.

Yet, the main theoretical difficulty in discussing this topic surrounds the issue of what precisely constitutes the prototype. As Gries (2015, p. 473–474) notes, the prototypical sense of a word could be the most frequent, most salient, most concrete, or the earliest acquired sense, and it is furthermore likely that the criteria of prototypicality differs across nominal classes. As such, there seems little fundamental connection between prototype structure and models of polysemy; rather, the prototype is a representation likely independent of semantic or pragmatic processes (Murphy, 2016) and, the criteria for being a prototype likely differs across word classes and conceptual domains.

TABLE 1 | Main prototype effects and their mutual relationships.

	Extentionally	**Intensionally**
Salience	Differences of salience among senses	Clustering of senses into nominal representations
Discreteness	Senses can demarcate different entities	Absence of definitions in terms of necessary and sufficient features

As Geeraerts (1989) also proposed, there are certain concepts (for example, INSTITUTION-related polysemous nominals) which may not be able to be defined through a set of necessary and sufficient features, and which exhibit a semantic structure which assumes the form of a set of clustered and overlapping interpretations. For example, "The school with large windows starts at 9 a.m. and has a strict headmaster and unruly students" contains a number of clustered senses being attributed to a single nominal. However, it may be the case that one of these senses is more salient and prototypical than the others, but intuition alone does not seem a powerful enough measure to expose this. As such, behavioral data is needed.

All of these effects are clearly present in a particular instantiation of polysemy known as *copredication*, whereby multiple, semantically different sense types are associated with a single nominal ("Lunch was delicious but was delayed"; "The newspaper that I held this morning has been sued"). Progressing on from recent research into the acceptability dynamics of copredication (Murphy, 2019, 2021a,b) and possible lower-level accounts (Murphy, 2018, 2020), I will investigate what I will term the *persistence conditions* of copredication, making close contact with the nature of prototypical copredications. The main research question addressed will be: What are the conditions under which the identity of a given entity can survive? This will act as a refined, controlled version of the classical Ship of Theseus paradox and the river paradox of Heraclitus. Crucially, the very notion of persistence conditions more readily lends itself to forms of polysemy involving copredication, since the notion of copredication is rooted in a sense of semantic conflict and incompatibility, rendering the construction of scenarios involving some aspect of competition between senses feasible.

In terms of focal predictions, the persistence conditions of these entities could be primarily determined by the number of senses being referred to in the discourse. This would suggest that sense number renders the ongoing representation of the entity salient, supporting its ultimate representational perseverance. Alternatively, one particular sense may more strongly predict how the object persists, such that, for example, the institutional sense of *school* determines its persistence, and not any other sense (e.g., a *school* might not be conceived as a building with an institution, but an institution with a building). One might relate these predictions to certain existing models of polysemy. For example, the Sense Enumeration Lexicon Hypothesis maintains that distinct senses of polysemous words like *school* are in fact represented as separate lexical entries, such that the persistence of any individual sense would be predicted not to be directly reliant upon any other sense, since these are lexically independent (supporting the sense number prediction) (Lehrer, 1990; Foraker and Murphy, 2012). On the other hand, the One Representation Hypothesis maintains, broadly speaking, that a word like *school* has multiple, underspecified representations connected to a single lexical entry, and would more directly be related to the sense type prediction (Frisson, 2009, 2015). More specifically, Löhr's (2021) distinction between rich and thin semantic representations of polysemy (both of which are in accord with the One

Representation Hypothesis) are relevant here. The thin view maintains that one specific sense forms the core meaning, around which other senses are clustered, while the rich view sees multiple senses forming distinct contributions to lexical meaning whilst still maintaining the existence of a single lexical entry. For more extensive discussion of polysemy storage and processing models (see Carston, 2016, 2019; Vicente, 2018; Ortega-Andrés and Vicente, 2019).

MATERIALS AND METHODS

Materials

An online acceptability judgment experiment was carried out using Qualtrics[1] and sourcing participants from Prolific Academic (prolific.ac). The purpose of the experiment was to present participants with scenarios involving a nominal that licenses copredication (e.g., *bank*, *city*) and ask them to make a forced choice to determine which outcome of the scenario (and, hence, which combination of predicates) appears to be the most prototypical, faithful realization of the nominal. In order to achieve this, an additional manipulation was added, such that the number of senses hosted by each choice was either equal (2 senses choice vs. 2 senses choice) or unequal (1 sense choice vs. 2/3 senses choice). All participants saw all trials across both types.

A core question of interest was whether the INSTITUTION sense strongly determined the option chosen by participants, or whether the number of senses predicted this. Addressing this question required balancing the number of scenarios in which the institution sense appeared either in isolation or with other senses. As such, in three scenarios the institution sense was isolated, in three scenarios it appeared with other senses while a different sense (e.g., PHYSICAL) was isolated, and the remaining six scenarios contained an equal number of senses.

The scenarios involved a two-alternative forced choice between qualified (licensed/acceptable) copredications of differing types, with participants being asked to choose *Which is the bank?* or *Which is the city?* 12 narratives were constructed, 6 of which presented two options exhibiting an equal number of senses, and the other 6 presented two options with an unequal number of senses. The narratives themselves exhibited all of the senses in the choices for participants. The central question which arises here is: What determines the more popular choice amongst participants? One possibility is *sense frequency*, i.e., participants will choose whichever option hosts the most frequent sense. Another possibility is *sense complexity*, with the most semantically complex sense (the one able to host the greatest number of senses and which is related to the greatest number of "core knowledge systems" or cognitive modules; Carey, 2009) determining true or prototypical objecthood for a given nominal.

Narratives of the following type were constructed (see **Supplementary Appendix** for full list). In the experiment, each narrative was followed by a choice: "Which is the X?":

> *Library*: A library catches on fire and is shut down. A new building across the street with self-service machines is built to help the public take out books. However, the employees of the original building protest and insist that the old library can simply be repaired, and refuse to hand over most of the books to the new building.
>
> Choice: Old building—New building
> Senses: Physical, People—Physical, Process
>
> *Village*: The King of a medieval village becomes corrupt and so the folk stage a rebellion, burning down the whole place in the process. Taking the village's original architects and main political leaders with him, the King relocates to a new site to accurately rebuild it. The village's entire population, however, move to a different site and also bring with them one of the original architects who helps them rebuild.
>
> Choice: King's site—People's site
> Senses: Physical, Polity—Physical, Populace

Six fillers were also used to ensure participants were paying attention, with these narratives having clearer and more obvious answers. These were of the following kind:

> *Sandwich*: John decides to make a sandwich. He slices it in half, begins to eat the first half, but then finds the bread very hard and difficult to swallow. He decides to take the second half and blend it into a smoothie.
>
> Choice: First half—Second half

Procedure

Scenarios were presented as single paragraph blocks, over a white background. Scenarios were presented in a random order. Below each paragraph, two options were presented corresponding to either choice, and participants were tasked with selecting their choice. After the presentation of the two-alternative forced choice, the screen was refreshed and participants were tasked with selecting one option from a 1–5 confidence metric, introduced by the question "How confident are you about your choice?" This allows differentiation between cases in which participants strongly believed in their choice, and cases when they were more ambivalent. Finally, to ensure that participants paid attention, they were forced to explain their reasons for selecting either option.

Participants

79 native English speakers (average age: 36, range: 20–60, 60 female) took part in the experiment, sourced from Prolific Academic and having an approval rating on the site of at least 90%. Participants were paid £6 per hour, with the average finishing time being 10 min. This study was approved by the UCL Research Ethics Committee and participants consented immediately prior to the experimental procedure to their recording responses being used for academic research purposes.

RESULTS

Confidence Scores

The average confidence score for the fillers was 4.2, and was lower for the experimental items (3.6). Confidence scores for the experimental items ranged from 3.01 (*city*) to 4.06 (*bank*),

[1] qualtrics.com/uk

suggesting that the responses can be taken as an accurate and genuine reflection of the participant's semantic intuitions. Confidence scores for experimental items can be found in **Table 2**.

Sense Type Score

Participants scored correctly on most fillers, and the minority of deviations (25/474 total responses) provided well-reasoned explanations. The results for the experimental narratives are plotted in **Figure 1**.

A multiple regression analysis was conducted using SPSS 24 with Sense Equality (Equal vs. Unequal) and Nominal (1–12 of the nominal list) as independent variables and Sense Type (Institution vs. Non-Institution) as the score. Sense Equality and Nominal significantly predicted Sense Type [$F(2, 945) = 19.659$, $p < 0.001$]. While Sense Equality added statistical significance to the prediction ($p < 0.001$), Nominal did not ($p = 0.438$). The average Sense Type score for the Unequal nominals was 0.59 (with 1 being 100% Institution sense chosen) and was 0.39 for the Equal nominals.

DISCUSSION

The results suggest that adding a greater degree of sense variability to each decision (e.g., Village #1 vs. Village #2) results in a more stable role for the institutional sense in determining objecthood. Participants were more willing to select a Non-Institution sense when the number of senses was matched across choices, but when the number of senses was unequal this seems to have exposed the possibility that the institutional sense was somehow more primary and essential, since when the number of senses is equal the (putative) primacy of the institutional sense would be harder for participants to detect, only becoming clear when they were forced to make a more stark choice between the institutional or non-institutional senses.

Comparing the present results to those from a previous frequency experiment (Murphy, 2019) allows us to determine whether the results were modulated by sense dominance. For *village* and *town*, PHYSICAL is the most frequent sense (based on results reported in Murphy, 2019, 2021a), which initially appears to predict the present results, yet participants made their decision based on the PEOPLE sense, not PHYSICAL, since both options hosted a PHYSICAL sense, and indeed both these decisions may have been influenced more by the POINT OF ORIGIN than by PEOPLE. Either way, it is difficult to determine whether frequency was the determining factor. PHYSICAL is also the dominant sense of *factory*, and two new locations were part of this narrative so POINT OF ORIGIN played no role, and both choices also hosted a PHYSICAL sense, but as with *village* and *town* the PEOPLE sense appears to have determined the outcome; that is, the sense semantically closest to INSTITUTION (and also the most abstract sense) determined the outcome. While the participant's choices for *company*, *school* and *university* on average matched the dominant sense of these nominals (INSTITUTION), twice as many other nominals (*province, city, library, church, shop, bank*) deviated from this.

In conclusion, the best predictor of the results was not sense frequency, but rather sense complexity or approximation to INSTITUTION. It was reasonable to hypothesize that frequency would have had at least some influence, not least because the present experiment was explicitly oriented toward pitting two entities against each other, with prototypicality likely being one of the best guides for participants to make a judgment. Yet, even under these circumstances, sense frequency was not the determining factor.

It might also initially appear that persistence conditions could more easily be explored by giving participants stages of decomposition, such that parts of a city (its people, institutions, etc.) are removed one-by-one in a given scenario, with participants being asked "Is this a city?" every step of the way. However, since objects can exist with only a single polysemous sense persisting this proposal would likely result in all participants (rightfully) selecting "Yes" up to the point when the last sense remains. As such, not only would this design not speak to the question of *core, essential* senses of complex polysemous nominals, it would also likely result in all participants claiming that each nominal survived the full destructive process simply because such a process would necessarily be a "sense-by-sense" level of destruction, since this is the only level of granularity one can operate at (with the only other possibility being manipulating pragmatic factors such as the context that the object denoted by the nominal found itself in, in which case we are back to the original design of the present experiment).

One potential objection to the present study is that it did not control for pragmatic factors, such as the tendency for participants to "side" with the underdog against a putative antagonist, encouraging them to choose "People's location" over "King's location" for the *village* narrative. However, this seems unlikely since both *village* and *town* contain a clear "antagonist" (the corrupt King representing the INSTITUTION sense and the violent gangs representing the POPULACE sense) but this did not predict responses, since participants on average chose the ("good") POPULACE-PHYSICAL senses for *village* but the ("bad") POPULACE-PHYSICAL senses for *town*, even when the PHYSICAL sense of *town* was described in the narrative as being doomed for demolition. Likewise, the majority of participants

TABLE 2 | Average confidence scores across all nominals.

Nominal	Score (1–5)
Library	*3.75*
Village	*3.94*
Factory	*3.51*
City	*3.01*
University	*3.64*
Town	*3.71*
Church	*3.84*
Shop	*3.88*
Bank	*4.06*
Company	*3.29*
Province	*3.16*
School	*3.46*

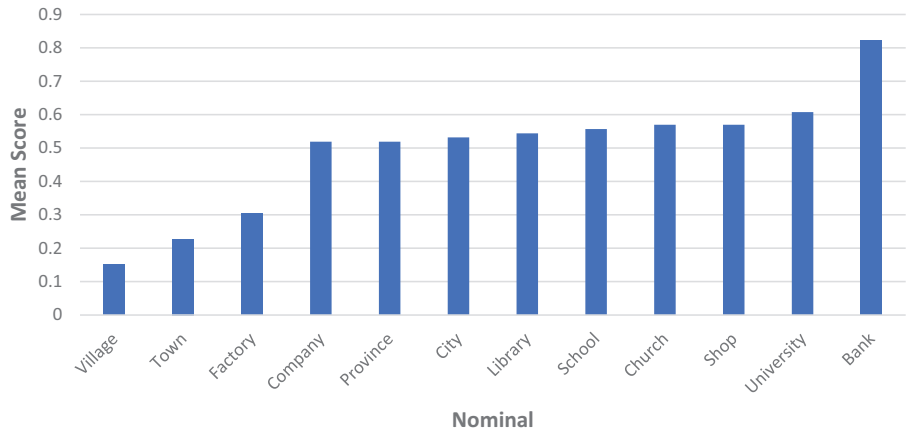

FIGURE 1 | Average scores for all nominals, with the *Y*-axis representing percentage of Institutional sense selection (0.1 = 10%), where POLITY is subsumed as an institutional sense and narratives involving no explicit institutional sense were sorted such that the sense most approximate to INSTITUTION acted as an institutional proxy (e.g., POPULACE/EMPLOYEES is more semantically related to INSTITUTION than PHYSICAL).

opted for the "antagonist" PHYSICAL-INSTITUTION senses for *university* rather than the PHYSICAL-POPULACE senses. The notion of POINT OF ORIGIN therefore seems to play a role in individuating institutional entities like *town*, as it also does for *school*, *company* and *shop*—but not for *bank*, seemingly due to FUNCTION trumping POINT OF ORIGIN. For Pustejovsky (1995), these two notions of FUNCTION and POINT OF ORIGIN are, respectively, referred to as the TELIC and AGENTIVE Qualia roles. It appears that, much as how Pustejovsky originally claimed, different Qualia roles are foregrounded across different nominals based on context—although, going somewhat beyond this, the present experiment also suggests that the relations of these Qualia roles might exhibit a more robust, generalizable structure or hierarchy, such that the telicity of *bank* is foregrounded not simply due to context/narrative, but because of the internal structure of the senses which compose the (highly) polysemous nominal. This hypothesis ties in with Lang and Maienborn's (2011, p. 719) interesting proposal that institutional nominals appear to have a common PURPOSE semantic feature (essentially Pustejovsky's TELIC role), with the senses enveloping schools and banks and shops ultimately being centered on a core lexical meaning: "[A] legal entity that organizes purposeful events to be performed and/or received by authorized groups of persons in specific locations." Nevertheless, it is remains possible that certain pragmatic factors were a confounding factor (e.g., narrative perspective; coherence relations), and future research should aim to more carefully examine this possibility.

CONCLUSION

This experiment has shown that specific semantic and pragmatic factors enter into judgments about the persistence of polysemous entities in complex scenarios. These results provide support for a version of the One Representation Hypothesis of polysemous lexical representations, through which different senses of polysemous words can be navigated around a mereologically focal and essential representation. As such, Löhr's (2021) thin view of polysemy representation appears to be supported, although the specific sense which forms the thin "core" seems to vary across nominals. Further research is required to more systematically relate Löhr's (2021) model to psycholinguistic concerns. The present findings do not directly support Vicente's (2017) claim that the INSTITUTION sense is necessary for the persistence of these nominal types, nor do they support (Arapini's, 2013, 2015) belief that the multiple senses of institutional entities are "clustered in a symmetric structure" (2013, p. 35), since there is variability in spite of the very strong trend in INSTITUTION-dominance. While there may in fact be no such thing as a core, essential sense for any of the nominals discussed (with each nominal being a cluster of senses with pragmatic factors determining which one is brought to the fore), the results suggest that there is considerable variation in the level of INSTITUTION-dominance the sense-cluster of each nominal exhibits.

Future experiments involving a larger range of scenarios could introduce additional factors to test, such as sentence type/syntax, or the presence of coherence relations between components of the scenario. Addressing the issue of pragmatics, narrative frames could also be kept consistent across nominals; statistical power could be boosted by increasing the number of nominals; and all nominals could be presented under both equal/non-equal sense number conditions, to more directly test the extent to which variations across nominals impacts persistence.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/**Supplementary Material**.

ETHICS STATEMENT

The study was approved by the Department of Linguistics Ethics Committee (LING-2013-3). Written informed consent to participate in the study was provided by all subjects.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

FUNDING

This research was funded by an Economic and Social Research Council scholarship (1474910).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.528862/full#supplementary-material

REFERENCES

Arapini's, A. (2013). Referring to institutional entities: Semantic and ontological perspectives. *Appl. Ontol.* 8, 31–57. doi: 10.3233/AO-130122

Arapinis, A. (2015). Whole-for-part metonymy, classification, and grounding. *Linguist. Philos.* 38, 1–29. doi: 10.1007/s10988-014-9164-6

Carey, S. (2009). *The Origin of Concepts*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780195367638.001.0001

Carston, R. (2016). The heterogeneity of procedural meaning. *Lingua* 175-176, 154–166. doi: 10.1016/j.lingua.2015.12.010

Carston, R. (2019). "Ad hoc concepts, polysemy and the lexicon," in *Relevance, Pragmatics and Interpretation*, eds K. Scott, B. Clark, and R. Carston (Cambridge: Cambridge University Press), 150–162. doi: 10.1017/9781108290593.014

Chomsky, N. (2018). "Mentality beyond consciousness," in *Ted Honderich on Consciousness, Determinism, and Humanity*, ed. G. Caruso (London: Palgrave Macmillan), 33–46. doi: 10.1007/978-3-319-66754-6

Collins, J. (2017). The copredication argument. *Inquiry* 60, 675–702. doi: 10.1080/0020174X.2017.1321500

Foraker, S., and Murphy, G. L. (2012). Polysemy in sentence comprehension: effects of meaning dominance. *J. Memory Lang.* 67, 407–425. doi: 10.1016/j.jml.2012.07.010

Frisson, S. (2009). Semantic underspecification in language processing. *Lang. Linguist. Compass* 3, 111–127. doi: 10.1111/j.1749-818X.2008.00104.x

Frisson, S. (2015). About bound and scary books: the processing of *book* polysemies. *Lingua* 157, 17–35. doi: 10.1016/j.lingua.2014.07.017

Geeraerts, D. (1989). Prospects and problems of prototype theory. *Linguistics* 27, 587–612. doi: 10.1515/ling.1989.27.4.587

Gries, S. (2015). "Polysemy," in *Handbook of Cognitive Linguistics*, eds E. Dąbrowska and D. S. Divjak (Berlin: de Gruyter Mouton), 472–490. doi: 10.1515/9783110292022-023

Lang, E., and Maienborn, C. (2011). "Two-level semantics: semantic form and conceptual structure," in *Semantics*, eds C. Maienborn, K. von Heusinger, and P. Portner (Berlin: De Gruyter), 709–740. doi: 10.1515/9783110226614.709

Lehrer, A. (1990). Polysemy, conventionality, and the structure of the lexicon. *Cognit. Linguist.* 1, 207–246. doi: 10.1515/cogl.1990.1.2.207

Lewandowska-Tomaszczyk, B. (2007). "Polysemy, prototypes, and radial categories," in *The Oxford Handbook of Cognitive Linguistics*, eds D. Geeraerts and H. Cuyckens (Oxford: Oxford University Press), 139–169.

Löhr, G. (2021). Does polysemy support radical contextualism? On the relation between minimalism, contextualism and polysemy. *Inquiry* 2021:1868329. doi: 10.1080/0020174X.2020.1868329

Murphy, E. (2016). Phasal eliminativism, anti-lexicalism, and the status of the unarticulated. *Biolinguistics* 10, 21–50.

Murphy, E. (2018). "Interfaces (travelling oscillations) + recursion (delta-theta code) = language," in *The Talking Species: Perspectives on the Evolutionary, Neuronal and Cultural Foundations of Language*, eds E. Luef and M. Manuela (Graz: Unipress Graz Verlag), 251–269.

Murphy, E. (2019). "Acceptability properties of abstract senses in copredication," in *Perspectives on Abstract Concepts: From Cognitive Processing to Semantic Representation and Linguistic Expression*, eds M. Bolognesi and G. Steen (Amsterdam: John Benjamins), 145–165. doi: 10.1075/hcp.65.08mur

Murphy, E. (2020). *The Oscillatory Nature of Language*. Cambridge: Cambridge University Press. doi: 10.1017/9781108864466

Murphy, E. (2021a). *Linguistic Representation and Processing of Copredication*. Ph. D. thesis. London: University College London. doi: 10.31234/osf.io/yubkz

Murphy, E. (2021b). Predicate order and coherence in copredication. *Inquiry* 2021:1958054. doi: 10.1080/0020174X.2021.1958054

Ortega-Andrés, M., and Vicente, A. (2019). Polysemy and co-predication. *Glossa J. General Linguist.* 4:564. doi: 10.5334/gjgl.564

Pustejovsky, J. (1995). *The Generative Lexicon*. Cambridge, MA: MIT Press.

Rosch, E. (1975). Cognitive reference points. *Cognit. Psychol.* 7, 532–547. doi: 10.1016/0010-0285(75)90021-3

Rosch, E. (1978). "Principles of categorization," in *Cognition and Categorization*, eds E. Rosch and B. B. Lloyd (Hillsdale: Lawrence Erlbaum), 27–48.

Vicente, A. (2017). What words mean and express: semantics and pragmatics of kind terms and verbs. *J. Pragmat.* 117, 231–244. doi: 10.1016/j.pragma.2017.07.007

Vicente, A. (2018). Polysemy and word meaning: an account of lexical meaning for different kinds of content words. *Philosop. Stud.* 175, 947–968. doi: 10.1007/s11098-017-0900-y

Conflict of Interest: 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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Murphy. 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) and the copyright owner(s) 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.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership