frontiers | Frontiers in Communication

TYPE Original Research
PUBLISHED 06 November 2025
DOI 10.3389/fcomm.2025.1696818

Check for updates

OPEN ACCESS

EDITED BY
Valentina Cuccio,
University of Messina, Italy

REVIEWED BY
Dusan Stamenkovic,
Södertörn University, Sweden
Francesca Strik Lievers,
University of Genoa, Italy

*CORRESPONDENCE
Mojca Brglez
✉ mojca.brglez@ff.uni-lj.si

RECEIVED 01 September 2025
ACCEPTED 10 October 2025
PUBLISHED 06 November 2025

CITATION
Brglez M and Pahor de Maiti Tekavčič K (2025) Enhancing transparency in source domain disambiguation for metaphor analysis: a cross-lingual approach integrating lexical resources, word embeddings, and human annotation. *Front. Commun.* 10:1696818. doi: 10.3389/fcomm.2025.1696818

Enhancing transparency in source domain disambiguation for metaphor analysis: a cross-lingual approach integrating lexical resources, word embeddings, and human annotation

Mojca Brglez[1,2]* and Kristina Pahor de Maiti Tekavčič[1,3]

[1]Faculty of Arts, University of Ljubljana, Ljubljana, Slovenia, [2]Jožef Stefan Institute, Ljubljana, Slovenia, [3]Institute for Contemporary History, Ljubljana, Slovenia

Contemporary cognitive-linguistic research often seeks to consolidate metaphorical expressions into systematic mappings between source and target domains. However, the formulation of such mappings in natural language remains insufficiently systematized, frequently relying on intuition or on lexical resources that are not available for all languages. In this study, we propose a systematic, semi-automatic approach to source domain identification that enhances transparency, objectivity, and replicability in metaphor analysis while reducing annotator reliance on intuition. We build on an established semantic ontology, bilingual lexical resources, and distributional semantic representations to assign semantic domains to words, which serve as proxies for conceptual source domains. We manually validate the data and quantitatively evaluate the method via automatic metrics. Furthermore, we perform a qualitative evaluation of annotation disagreements and a detailed error analysis. Results indicate that the approach provides a promising foundation for semantic tagging and metaphor analysis in Slovene. The qualitative analysis of disagreements demonstrates how individual linguistic variation and cognitive biases influence domain attribution, and often prevent reaching a complete consensus between annotators. The error analysis further identifies specific limitations of the proposed approach, which arise from gaps in lexical resources and from the inherent properties of distributional semantic modeling. Overall, the findings underscore both the methodological challenges of automatic domain attribution and the cognitive complexity of source domain mapping in metaphor analysis.

KEYWORDS

metaphor analysis, semantic tagging, source domains, word embeddings, USAS ontology, Slovene

1 Introduction

Metaphor, as a phenomenon of language, thought, and communication, serves as a powerful tool for exploring the conceptual frameworks through which we express ourselves and reason about the world. Since the cognitive turn in linguistics, and particularly with the rise of Conceptual Metaphor Theory (CMT, Lakoff and Johnson, 1980) in the 1980s, metaphors have been studied as systematic mappings between source and target domains that reveal how abstract concepts are framed through more concrete, physical, and familiar concepts which we experience in our everyday interaction with the physical world. In most recent approaches to metaphor, the Neural Theory of Metaphor (Lakoff, 2008, 2014) further advances the view that metaphor is found/experienced in the brain due to the activity between the neural nodes when a metaphor is cognitively processed (Feldman, 2008). This embodied perspective has shown that metaphor is not merely a rhetorical device, but a fundamental cognitive mechanism shaping how we reason about the world (Gibbs, 1994; Gibbs et al., 2004). Still, even these recent neurolinguistic approaches incorporate insights from the old CMT theory, which already advanced some ideas about metaphor that still hold true: that metaphors are parts of our brain circuitry, that they consist of conceptual mappings, that they operate on source domain structures which are used for reasoning about the target, and that many of those mappings are very conventionalized (Lakoff, 2008).

However, the CMT theory, albeit groundbreaking and influential, has also provoked a range of responses, with criticism focusing on the lack of empirical evidence or a clear method to find such empirical evidence in language for the lack of corpus evidence (Deignan, 2005), lack of a systematic procedure of finding metaphors (Group, 2007; Kövecses, 2008; Dobrovol'skij and Piirainen, 2022), and on the fact that many conceptual metaphors put forth by (Lakoff and Johnson, 1980) could also be phrased differently or correspond to other conceptual metaphors (Ritchie, 2003; Clausner and Croft, 1997; Deignan, 2008). In parallel to CMT, psychologists and cognitive linguists developed alternative models of figurative thought and approaches to the processing of metaphor and similar phenomena. For example, Gentner (1983) introduced structure-mapping theory, where she defined analogy as a process of aligning relational structures between two domains, i.e., as cross-domain mappings from "base to target." Such research laid an important foundation for metaphor studies, as metaphor can be seen as a special case of analogy. Furthermore, Glucksberg (2001) advanced the class-inclusion view, where metaphors instantiate *ad-hoc* categories (e.g., "my lawyer is a shark" creates a "predatory person" category). Building on both Gentner's and Glucksberg's views, Bowdle and Gentner (2005) proposed the "career of metaphor" hypothesis. According to the latter, novel metaphorical meanings are processed through comparison, much like analogies, then, as they become conventionalized, the meanings are directly accessed in terms of categories. This emphasizes that the mapping-view proposed by CMT is not as simple, and differently applies to novel and more conventional metaphors (see also Hanks, 2006). Metaphor has also been approached from the perspective of integration and blending theory. Fauconnier and Turner (2002) argue that any language comprehension, including metaphor, arises from blending different conceptual spaces, which can be seen similarly to domains. In the blended space, only certain attributes and structures from the input spaces are selected, where they are further combined and built upon to create other emergent properties.

All theories are faced with the problem of delimiting what is a metaphor, which is usually based on the opposition of the metaphorical to the "literal" or "basic" meaning. However, even the latter is not easy to define. As mentioned by Steen (2007), (at least) five definitions of the literal meaning were previously given by cognitive scientists, including those that define it as the "conventional," "usual," "context-free," "objectively true," and "directly meaningful." He argues that for metaphor, the latter perspective should be taken, i.e., that the literal or basic meaning is the "direct" meaning which is not understood in terms of or with the help of another meaning. This was also adopted in the first formalized approach to identifying metaphors in text (Metaphor Identification Procedure or MIP, Group, 2007), where the authors emphasize the difference between the contextual and basic meanings. Here, basic meanings are defined as more concrete (what they evoke is easier to imagine, see, hear, feel, smell, and taste), they relate to bodily action, are more precise (as opposed to vague), and are historically older. Secondly, as the authors note, basic meanings are not necessarily the most frequent meanings of lexical items. However, a group of researchers around Gerard Steen have problematized certain aspects of basic meaning, for example, the one that the basic meaning is historically older, which is why they proposed a further development of the procedure named MIP—Vrije Universiteit (MIPVU, Steen et al., 2010). Here, the authors define "basic meaning" as "a more concrete, specific, and human-oriented sense in contemporary language use. /.../ always to be found in a general users' dictionary." (Steen et al., 2010, p. 32). MIPVU thus changes the previous view of basic meaning as the historically older one, as such etymological meanings may already be obsolete, absent from speakers' mental lexicons, and thus irrelevant for cognitive-linguistic analysis of contemporary language. However, MIPVU may not be directly applicable to other languages or linguistic communities for reasons of both linguistic diversity and lack of contemporary, well-structured corpus-based resources, which the method precludes (cf. Nacey et al., 2019). Nonetheless, MIPVU has proved to be a reliable tool for metaphor identification that can lead to a high agreement among annotators. Identifying metaphors is typically followed by the more consequential task of analyzing or interpreting them, which provides the basis for testing hypotheses and inferring the broader metaphorical frameworks at play. This step can range from simply formulating the conceptual mapping (i.e., defining the source and target domain), paraphrasing the linguistic metaphor in (more) literal terms, explaining the grounds of the metaphor (i.e., the common aspects between domains that enable the comparison), or predicting likely entailments or implicatures of a metaphor in a given context. In a broader sense, this task falls within the realm of semantic annotation, where meaning is assigned to linguistic units based on either structured resources or human intuition. Numerous studies have noted the importance and challenges of relying on human judgments, especially when dictionary definitions are incomplete, ambiguous, or language-specific (e.g., Joshi et al., 2013; Beck et al., 2020). In such cases, annotators often default to their intuitive understanding, whereas human language processing is subject to high individual variation (Lewellen et al., 1993; Kidd et al., 2018; Boland et al., 2016; Marti et al., 2023; Ramsey, 2021; Stacy et al., 1997). These differences are particularly important in metaphor interpretation, where–unlike the relatively well-defined procedure for metaphor identification–the task remains less structured and more dependent on the intuition of the individual researcher. Most previous metaphor studies have based their domain labels on the Master Metaphor List

(MML, Lakoff et al., 1991) stemming from Lakoff and Johnson's Conceptual metaphor theory (Lakoff and Johnson, 1980) despite several critiques. Among others, researchers criticized both the lack of a clear methodology and the ambiguity caused by the use of variably phrased domain labels (Krennmayr, 2013; Cameron, 2003; Vervaeke and Kennedy, 1996; Ritchie, 2003). Only recently, Reijnierse and Burgers (2023) proposed "MSDIP," a method for coding source domains. However, the scope or specificity of possible domain labels remains undetermined.

The lack of a clearly defined set of domain labels, along with the issues of the level of specificity and synonymous wordings, poses at least two hurdles. On the one hand, without a clearly defined set of labels, it is difficult for an annotator not trained in metaphor analysis and unfamiliar with the established conceptual metaphor phrasings to formulate denominations for these domains. On the other hand, the presuppositions of metaphor researchers may drive them to choose one over many possible alternative domains. Some efforts have been made to use domain labels from existing ontologies such as FrameNet (Baker et al., 1998; Boas et al., 2024), USAS (Rayson et al., 2004; Piao et al., 2015), and WordNet (Fellbaum, 1998, 2005). However, these resources are specific to English, and, even when adapted to other languages, do not necessarily reflect the particularities of those systems. In light of conducting more comprehensive analyses of metaphorical frameworks in and across languages, the different labeling approaches make it impossible to compare and contrast findings from different studies.

In this work, we present a computer-assisted, semi-automatic approach to conceptual domain attribution to support metaphor identification and interpretation in Slovene. We base our approach on CMT and source-to-target domain mapping operationalized in MIPVU, as although only one of the possible approaches, it has proven to be a reliable tool for metaphor analysis. In this view, we note that the analysis in terms of broad conceptual domains, presented in this paper, is proposed only as one of the many possible schematic structures through which metaphors *may* be understood. The algorithm we propose annotates words with their respective conceptual domains taken from a well-established general semantic ontology (USAS, Rayson et al., 2004) and is based on lexical (Open Slovene WordNet—OSWN 1.0, Čibej et al., 2023 and Open English WordNet—OEWN, McCrae et al., 2019) and distributional-semantic resources (Slovene word embeddings CLARIN.SI, Terčon et al., 2023). Information about the conceptual domains of words in a given sentence streamlines both metaphor identification and interpretation. On the one hand, in a text not yet annotated for metaphors, it is possible to discern potentially metaphorical units by the incongruity of the domains within a sentence or text (i.e., a word originating in the conceptual domain of ANIMAL[1] in a text about human migrations alerts the analyst of a possible metaphor). On the other hand, in a text already annotated for metaphorical expressions, the domain labels help the analyst to interpret metaphorical units in terms of cross-domain mappings. This reduces reliance on intuition, eliminates the issue of determining the exact label wording and the level of specificity, and facilitates comparative metaphor analysis.

The remainder of this paper is organized as follows. Section 2 provides an overview of the existing approaches to metaphor identification and interpretation. In Section 3, we introduce our methodology, including the underlying knowledge resources and the algorithm to consolidate resources and annotate words with semantic domains. In Section 4, we assess the performance using manual annotation, detailed error analysis, and metric-based evaluation. Section 5 concludes the study with a discussion of the main findings and suggestions for future research.

2 Related work

Many studies since the advent of CMT have used metaphor as a tool to find entrenched conceptual mappings and their proposed implications/inferences. However, the theory has triggered critical discussion, including doubts toward he fact that many conceptual metaphors put forth by Lakoff and Johnson (1980) could also be phrased differently or correspond to other conceptual metaphors (Clausner and Croft, 1997; Deignan, 2008; Ritchie, 2003). Despite these critiques, a significant number of metaphor annotation projects (e.g., Eilts and Lönneker, 2002; Shutova and Teufel, 2010) adopted the phrasings proposed by the Master Metaphor List (MML, Lakoff et al., 1991). Conversely, some studies have opted for alternative domain labeling practices, stepping away from "conceptual" to "systematic" metaphors (e.g., Maslen, 2010), or focusing on other levels of schematic structures, such as scenarios, image schemas, or conceptual spaces (see Kövecses, 2020). To address the issue of variability and to provide more solid grounds for metaphor analysis, attempts have been made to systematize the process, primarily based on dictionary or corpus evidence (e.g., Deignan, 2016; Krennmayr, 2013; Semino, 2018; Steen, 1999). Most recently, Reijnierse and Burgers (2023) proposed Metaphor Source Domain Identification Procedure ("MSDIP"). In MSDIP, the annotators consult a dictionary to discern and rank word senses, which facilitates the assignment of semantic domains. However, the scope or specificity of possible domain labels remains undetermined, and due to the reliance on a dictionary, the procedure is not directly transferable to all languages. For example, the current reference dictionary for Slovene[2] is not a contemporary corpus-based resource. It frequently lacks explicit sense definitions altogether[3] and confounds distinct senses.[4]

Metaphor interpretation has also been addressed computationally, by extracting domain labels from existing general lexico-semantic resources (e.g., Demmen et al., 2015; Dodge et al., 2015; Ge et al., 2022; Koller et al., 2008; Mason, 2004; Sengupta et al., 2023; Shaikh et al., 2014). The task can range

[1] We use small caps to indicate conceptual or semantic domain labels following the tradition of other works on metaphor.

[2] Slovar slovenskega knjižnega jezika "The Dictionary of standard Slovene language" [digital edition], second, supplemented and partially revised edition, available at www.fran.si.

[3] Instead, it points to the base word from which it derives. For example, the definition of the noun *petje* "singing" is "*glagolnik od peti*," meaning "gerundive" form of *peti* "to sing."

[4] Metaphorical uses are commonly presented as examples under a more basic sense, without separate definitions.

from paraphrasing the metaphorical expression to (more) literal language (Shutova, 2010), source domain identification (Sengupta et al., 2022), conceptual/systematic metaphor identification (Ge et al., 2022), modeling the highlighting in metaphor, i.e., the characteristics or aspects transferred from source to target (Sengupta et al., 2023; Wang et al., 2023), and even intentions behind using a metaphor (Michelli et al., 2024). Most recently, generative LLMs are also being employed to perform both identification and interpretation (Hicke and Kristensen-McLachlan, 2024; Ichien et al., 2024; Tian et al., 2024; Wachowiak and Gromann, 2023). However, studies such as those by Pedersen et al. (2025) on Danish or Puraivan et al. (2024) on Spanish data show a heavy bias of English: the models perform better on English data or on metaphors that have conceptual equivalents in the English language.

Both manual and computational approaches can present limitations. Manual approaches often lack a clear or comprehensive set of domain labels, making comparisons between studies, especially contrastive cross-linguistic studies, particularly challenging. Secondly, both manual and computational approaches often base their findings on a handful of very specific conceptual or discourse domains. Furthermore, an overwhelming number of studies focus on English, leaving a large gap in metaphor analysis for smaller or less-studied languages, which often lack the resources and tools needed for robust research. In our study, we address this gap by introducing a novel semi-automatic approach to conceptual domain attribution to support metaphor interpretation in Slovene, using bilingual lexical resources and domains taken from a well-established general semantic ontology.

3 Materials and methods

In this chapter, we first present the resources and tools we use in our approach to basic domain disambiguation. This includes the USAS semantic ontology and lexicon (Rayson et al., 2004), Slovene (Čibej et al., 2023) and English (McCrae et al., 2019) WordNets, and Slovene word embeddings (Terčon et al., 2023). In Section 3.5, we outline our approach–the algorithm that integrates these knowledge resources to tag words in Slovene texts. Finally, in Section 3.7, we present the dataset, manual annotation procedure, and quantitative metrics used to evaluate the source domain disambiguation procedure.

3.1 The USAS ontology and lexicon

Our approach rests primarily on the semantic tagset used by the UCREL Semantic Analysis System (USAS, Rayson et al., 2004). On the top level, the USAS tagset contains a set of 21 major discourse fields. These are further separated into a total of 232 fine-grained semantic categories, hierarchically organized into maximally four sublevels or specifications. Both the major discourse fields and semantic categories have natural language labels and corresponding codes or tags, for example, the label SUBSTANCES, MATERIALS, OBJECTS AND EQUIPMENT is also referred to with the tag "O," and a more fine-grained category Texture is referred to with the tag "O4.6." A visualization of the hierarchical structure of the "O" discourse domain with two sublevels is shown in Figure 1. To facilitate the semantic analysis of running text, the UCREL system uses a lexicon of 55,662 English words and multi-word expressions (MWEs), disambiguated by part-of-speech. These words and MWEs are tagged with semantic categories from the tagset that correspond to their multiple senses, ordered by their frequency of use in discourse. For instance, the noun *object* is tagged with O2 and X7+, corresponding to the categories "Objects generally" and "Wanting; planning; choosing" with a positive connotation.

The UCREL system has been primarily applied to ascribe semantic domains to words and phrases in English texts in general, as well as specifically in the analysis of metaphoric expressions (Demmen et al., 2015; Koller et al., 2008). The underlying tagset and lexicon have also been previously manually (e.g., Löfberg et al., 2005) and automatically (e.g., Piao et al., 2015; Brglez and Pahor de Maiti, 2024) transported to other languages. In our approach, we manually translate the tagset, i.e., the discourse domain and semantic category labels, to create a Slovene tagset.

3.1.1 Slovene translation of the tagset

The English tagset was translated into Slovene in several stages. First, we produced two independent draft translations that aimed to mirror the original wording as faithfully as possible. One of the authors then resolved the disputed translations, creating a single provisional list. The list was used in the algorithm to assess how well the translated terms behave in the matching process. We identified several problems with the translations that negatively impacted the algorithm performance, for example, the use of conjunctions and punctuation in the labels, or too many synonyms for one tag. We also conducted preliminary experiments where we compared embedding labels in lower vs uppercase, embedding multiword labels as sentences vs. using the average of word vectors, and averaging with or without punctuation. The best results were obtained by embedding the labels in lowercase, and, in the case of multiword labels, removing punctuation and averaging over the embedding of each word. Furthermore, we experimented with translating the tags only as single-word equivalents, but this proved less optimal compared to rich multiword labels, especially in cosine-based similarity ordering (see Sections 3.4 and 3.5). These findings resulted in the adoption of the following translation guidelines:

- top-level tags are translated as nouns, while lower-level tags are translated as adjectives using the neuter gender, or, if the adjectival form is not viable, as adverbs;
- whenever possible, a single-word Slovene equivalent is preferred;
- the singular form is preferred, although plural can be used when sensible, e.g., "People" (S2) was translated to *ljudje* "people," not *človek* "a human";
- within a single label, only commas are used as separators (conjunctions and other punctuation functioning as separators in the English tagset were discarded);

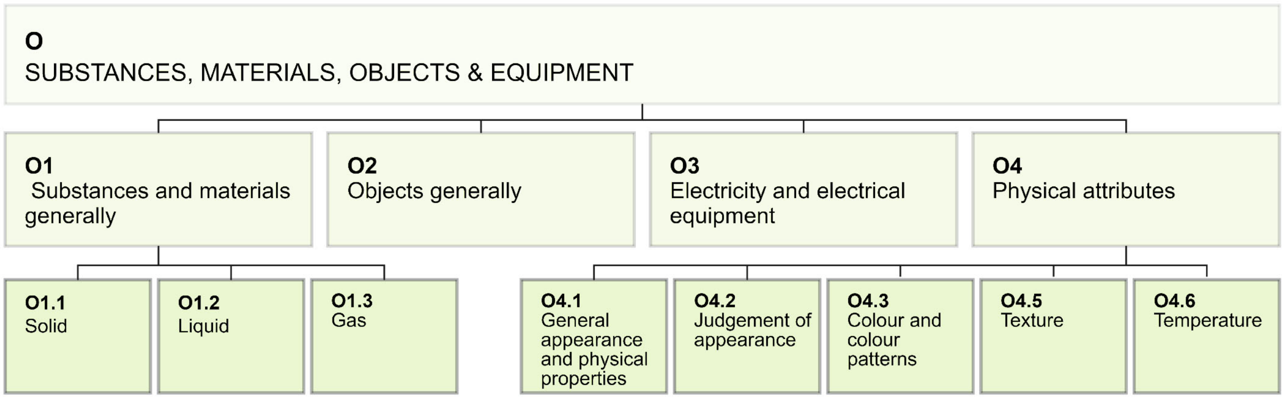
O
SUBSTANCES, MATERIALS, OBJECTS & EQUIPMENT
O1
Substances and materials generally
O2
Objects generally
O3
Electricity and electrical equipment
O4
Physical attributes
O1.1
Solid
O1.2
Liquid
O1.3
Gas
O4.1
General appearance and physical properties
O4.2
Judgement of appearance
O4.3
Colour and colour patterns
O4.5
Texture
O4.6
Temperature

FIGURE 1
Visualization of the O discourse domain (*Substances, materials, objects & equipment*) and its subcategories.

- when Slovene regularly uses several distinct equivalents for a single English term, a maximum of two synonyms[5] with different lexical roots are listed, e.g., "Participation" (S1.1.3) was translated to *udeleženost, sodelovanje* "participation, involvement";
- unlike the English tagset, the Slovene tagset excludes the word *general* (*splošno* in Slovene) or its equivalents from the tag, e.g., "Time: General" (T1.1) was translated to only *čas* "time," and "Objects generally" (O2) to *predmeti* "objects";
- when the English tagset combines two hierarchical levels in one label, the Slovene version retains only the lower-level label, e.g., "Evaluation: Bad" (A5.1-) was translated to *slabo* "bad."

This translation process enabled us to create a tagset that remains faithful to the original, ensuring cross-linguistic comparability, while also aligning with natural Slovene language use. The translated tagset also improves the internal coherence with regard to orthographic and lexical conventions. Finally, we empirically tested different label phrasings to inform the final guidelines and to enhance the performance of the algorithm.

3.2 Open English WordNet

Open English Wordnet (OEWN; McCrae et al., 2019) is a fork of the original Princeton WordNet (Fellbaum, 1998, 2005) and provides a lexical network of the English language. The network consists of lexical items or "literals" (words and multiword expressions), which are grouped into synonymous groups or "synsets" according to their sense. Synsets are linked through lexical relationships such as hypernymy and antonymy, and lexical items are linked via morphological relations such as participle and derivation. In this study, we use the 2023 edition of OEWN, which contains 161,338 literals, 120,135 synsets, and 415,905 relations.

3.3 Open Slovene WordNet

Open Slovene WordNet 1.0 (OSWN, Čibej et al., 2023) is a counterpart of OEWN for Slovene. OSWN 1.0 was constructed by using several existing mono- and bilingual resources and automatic methodologies, resulting in 95,262 synsets and a total of 164,904 literals. All synsets were manually checked by at least one expert annotator. OSWN currently does not include any lexical or morphological relations. Although it naturally contains literals (words) different from those in OEWN, both resources share an overwhelming majority of synsets representing possible concepts. For example, Figure 2 illustrates a synset with two Slovene synonymous words *bucika* and *risalni žebljiček*. The OSWN has a corresponding (or identical, by ID) synset in OEWN, which groups the English synonyms *pushpin*, *thumbtack*, and *drawing pin*. Figure 2 also depicts that the synset is linked to two other synsets via the hypernym relation, which contain the lexemes *tack* and *paper fastener*, respectively. Although not depicted in the image, the lexical entry *thumbtack* also has an additional morphological relation "derivation" which links it to the verbal lexeme *thumbtack*.

The alignment on the level of synsets provides a foundation for meaning-based mappings between the two languages. By connecting words through their shared synsets, we can link not only their surface forms but, more importantly, the underlying concepts they represent. Furthermore, we can leverage the lexical relationships unique to the English WordNet to extrapolate to Slovene, for example, finding potential hypernyms in the Slovene WordNet. We believe that regardless of the level of generality/specificity differentiating between hypernyms and hyponyms, lexical items from both ends should share the overarching (or at least one) semantic domain. In the case of *pushpin, drawing pin, thumbtack*, none of the entries are included in the USAS lexicon. However, the lexicon includes the hypernym *tack* and tags it with semantic domains OBJECTS GENERALLY; LIVING CREATURES.

Our approach rests on the assumption that, at a minimum, basic meaning(s) are represented and included within the lexical resources for English and Slovene. While OSWN and OEWN both contain around 160,000 lexical entries, there are

[5] If more than two suitable equivalents exist, the reference corpus of Slovene Gigafida 2.0 (Krek et al., 2020) is consulted to choose the most frequent ones.

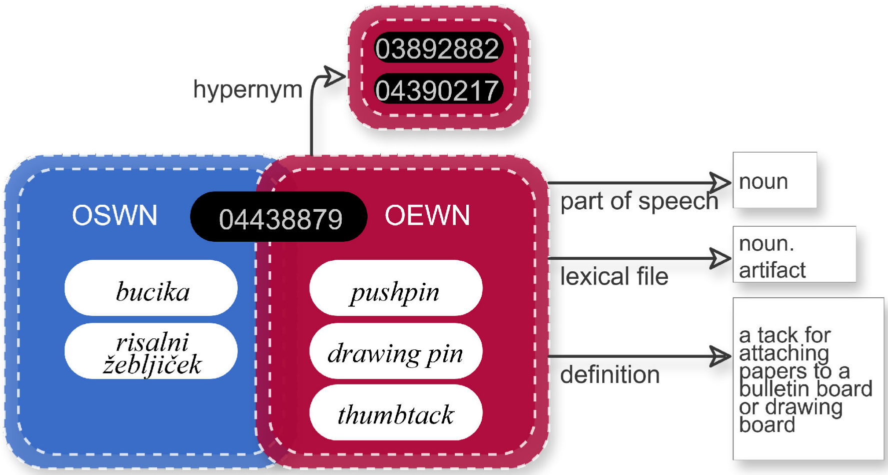

FIGURE 2
Members and relations of a WordNet synset containing the word *bucika* "pushpin."

as many as 321,994 unique mappings (defined here as unique combinations of a synset ID, a Slovene literal, and an English literal). However, the synsets and their relations are not always symmetric. Polysemy varies across languages, and lexical senses rarely map one-to-one between languages, leading to asymmetries in meaning (Cruse, 1986). The OEWN also enumerates senses of lexical entries, ordering synsets by sense rank. A possible approach would be to restrict our analysis to mappings involving only the first (i.e., most basic or most frequent) sense of English lexical entries. However, Slovene lexical entries that share the same synset do not necessarily reflect this basic sense. Additionally, while OSWN includes some sense numbering, we have found these mostly unreliable, often failing to correspond with dictionary definitions or intuitive sense hierarchies.[6] We acknowledge and try to mitigate such meaning asymmetries by relying on the semantic information captured in the space of Slovene word embeddings, which we describe in the next chapter.

3.4 Word embeddings

Modern computational methods for meaning representation rely on word embeddings–dense vectors situated in a continuous space and grounded in the principle of distributional semantics (Harris, 1954; Firth, 1957). Word embeddings encapsulate semantic and syntactic relationships by leveraging statistical patterns of co-occurrence within language data. In our approach, we use embeddings to embed words and domain labels in vector space. We then use these embeddings to compute their semantic similarity. As a measure for the latter, we use cosine similarity, which is calculated as the cosine of the angle between two vectors. It ranges from −1 to 1, where −1 indicates maximum dissimilarity (the vectors point in opposite directions), and 1 indicates maximum similarity (the two vectors point in the same direction). In our experiments, we use the CLARIN.SI-embed.sl 2.0 static word embeddings for Slovene (Terčon et al., 2023), which are skip-gram fastText embeddings trained on a large collection of Slovene corpora. The embeddings are used primarily to select the semantically most similar tags from the candidate tags mapped via the English synsets.

3.5 Mapping the USAS ontology to OSWN

Our approach uses a semantic ontology to label words and is primarily aimed at disambiguating the basic, literal meaning of words from metaphorical or other meanings. It is based on the consolidation of three main knowledge resources: the semantic ontology and lexicon (USAS), the English WordNet (OEWN), and the Slovene WordNet (OSWN). The approach rests on the assumption that, at a minimum, the lexical resources include the basic meaning of words, as this is either the most frequent or the most conventional meaning from which metaphoric and other meanings are drawn.

The overall algorithm for consolidating these resources is illustrated in Figure 3. The first step of our method involves consolidating OEWN and the USAS lexicon, that is, finding exactly matching entries from the two resources. For single words, we match entries based on lemma and part-of-speech. For multiword expressions (MWEs) in the USAS lexicon, which include information about each component's part of speech and wildcard characters (e.g., *old_ADJ timer*_NOUN*),[7] we search for the abbreviated, simplified expression (*old timer*) first. There are

[6] For example, the first synset of the lexical entry *mačka* "cat" represents the sense "very attractive woman" while the arguably most basic, animal sense of the word comes only third.

[7] The lexicon features wildcard characters such as ∗ in order to match various forms in running text.

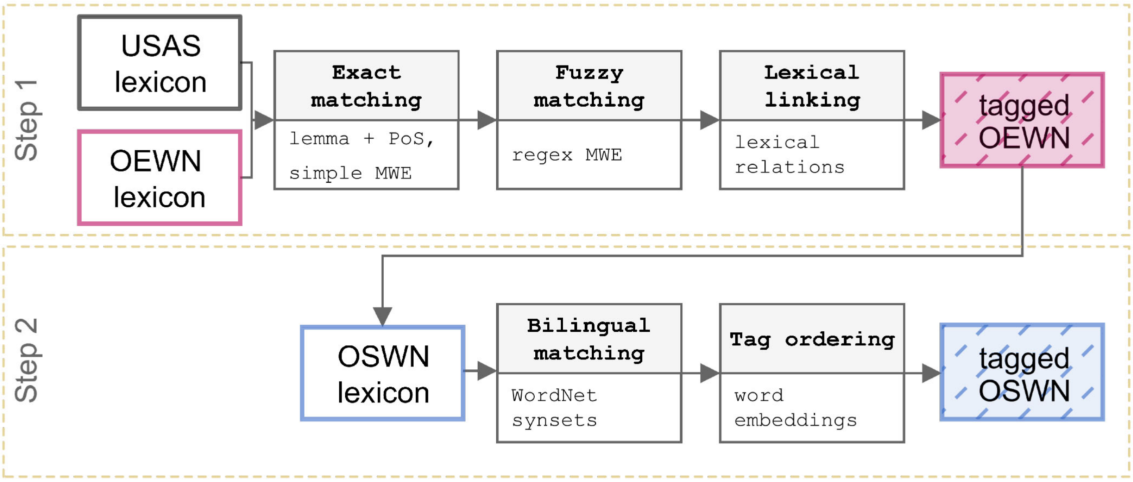

FIGURE 3
The process of lexical resource consolidation.

29,572 single-word lexical entries and 4,089 MWEs in OEWN with such exact matches to the USAS lexicon.

Then, to expand the overlap between USAS and OEWN and thus maximize the coverage, we also apply fuzzy and lexico-semantic matching for the remaining OEWN lexical entries without a match in USAS. Fuzzy matching only concerns MWEs since it refers to transforming the original USAS patterns into corresponding regular expressions, e.g., *old timer.**. This results in 2,138 additional matching literals in OEWN. Then, for the remaining single and multi-word literals, we search for lexico-semantically connected lexical entries, i.e., those linked via lexical relations. As described in Section 3.3, OEWN connects lexical items via lexical relations such as synonymy (items in the same synset) or hyponymy and hypernymy (items in synsets connected via the corresponding relation) as well as other morphological relations (derivation, participle). Items in these relations should, in principle, share the overarching semantic domain.[8] Thus, for lexical entries in OEWN not covered by the USAS lexicon, we successively search for their synonyms, hypernyms, and, finally, any other relations, and collect all the USAS domains associated with their lexically related items. For example, the noun *disquiet* does not appear in USAS; however, it appears in the same synset as and is thus a synonym of *unease* and *uneasiness*, which are both tagged with WORRY in USAS. An example of hyponym-hypernym matching is between the noun *zestfulness* and its hypernyms *zest, gusto, relish*, which contribute the tags INTERESTED/EXCITED/ENERGETIC, LIKE, GETTING AND POSSESSION, HAPPY. An example of matching based on other relations is the word *ellipse*, which is linked to its adjectival derivative *elliptical*. The adjective features in USAS and is tagged with SHAPE, CLOSED, and HIDING/HIDDEN. Lexico-semantic matching allows us to link an additional 32,976 single-word and 25,850 multi-word literals. Altogether, the algorithm mapped a total of 62,548 single-word and 32,077 multi-word lexical units in OEWN to at least 1 candidate domain from USAS, by which we attained 68% coverage for single-word literals and 47% coverage for MWEs.[9]

The second step concerns consolidating information from the USAS-tagged OEWN and OSWN and thus creating a Slovene semantically-tagged lexicon. In this step, the lexical approach using a bilingual resource is augmented by a distributional semantics approach using word embeddings.

As already mentioned in Section 3.3, the two lexical resources (OSWN, OEWN) contain the same synsets, i.e., groups of synonymous lexical units or “literals” denoting the same sense or concept. Literals can also be polysemous and thus belong to different synsets; however, their multiple senses do not always align across languages. Thus, we do not just map the semantic tags of English equivalents directly, nor only those including the first senses. This could, on the one hand, include very specific meanings of the English lexemes and thus conceptual domains not relevant to their Slovene equivalents, and, on the other hand, discard relevant basic meanings and domains. To mitigate this, we opt for collecting all possible mappings between English and Slovene entries (in all their senses) and all semantic domains of the English entries (which correspond to their different senses). Then, we use Slovene word embeddings to select the distributionally most similar domain(s). For example, for a candidate Slovene entry *biser*, we first collect its English equivalents *bead, pearl, gem, jewel*. Next, we collect all semantic categories attributed to these English equivalents: OBJECTS GENERALLY, SUBSTANCES AND MATERIALS, JUDGEMENT OF APPEARANCE: BEAUTIFUL, MENTAL OBJECT: CONCEPTUAL OBJECT, CLOTHES AND PERSONAL BELONGINGS, and PEOPLE. To find the most suitable basic conceptual domain(s),

[8] We conducted a preliminary analysis where we checked the overlap in USAS domains between lexically related items. For synonyms, they shared at least one USAS tag in 71% of cases in their first sense. The first sense of hyponyms shared a tag with their hypernyms in 58% of cases, and lexical items in other relations shared at least one tag in 81% of cases.

[9] The remaining literals for which we could not find a tag mostly concern proper nouns, such as *Grand Canyon*, or rare terms, such as *hypocycloid*.

we embed both the word *biser* and the Slovene domain labels (PREDMETI; TRDNO; LEPO; OBLEKE, LASTNINA; POJMOVNO; LJUDJE), and rank them by cosine similarity to the word. In this particular case, the closest domains are LEPO "beautiful," OBLEKE, LASTNINA "clothes and belongings," and PREDMETI "objects."

3.6 Tagging text

The USAS-tagged OSWN can be used to tag (lemmatized) text in Slovene by finding the corresponding entry and associated semantic tags. Additionally, we devise a way to annotate out-of-vocabulary (OOV) words, that is, those that are not present in the initial set of English-Slovene mappings. To this end, we again employ word embeddings and cosine similarity. Specifically, we search for N similar words among the Slovene lexical entries and gather the nearest k domains of those. We have experimented with various N and k and find optimal results with $N = 3$ and $k = 3$. For example, the word *religija* "religion" is not included among the OSWN lexical entries. However, we can to find three similar entries: *religioznost* "religiousness," *ideologija* "ideology," *religiozna oseba* "religious person," and the three closest conceptual domains of each of them. After embedding both the word and the collected domains and ranking the latter by cosine similarity, the three most suitable domains for the OOV in question are RELIGION AND THE SUPERNATURAL, CONCEPTUAL OBJECTS, EDUCATION.

3.7 Evaluation and annotation

No gold-standard datasets for semantic domain disambiguation exist for Slovene, which is why automatic evaluation cannot be straightforwardly performed. To assess the quality of the algorithm and the semantic domain tags, we apply the algorithm to unlabeled data and manually validate the tags. Then, we use the manual annotations as silver-standard data to perform automatic, metric-based evaluation.

3.7.1 Dataset

We use a subset of KOMET 1.0 (Antloga, 2020), a corpus of metaphors in Slovene covering various domains and genres. The subset was compiled by Brglez and Vintar (2025) and consists of noun-phrase constructions, i.e., syntactic constructions with a noun head modified by either an adjective (amod relation) or a noun (nmod relation). Examples of these include amod phrases such as *oster ritem* "sharp rhythm," *strokovna širina* "professional breadth," *poceni filmska finta* "cheap movie trick," and nmod phrases such as *podiranje sveta* "collapse [of the] world," *pot do sreče* "path to happiness," *slika naše realnosti* "image [of] our reality." These phrases may correspond to many different metaphorical constructions (Croft, 2003; Sullivan, 2009) featuring a conceptually dependent element evoking the source domain of the metaphor and a conceptually autonomous element evoking the target domain of a metaphor. In addition to the metaphoric examples, the subset includes the same amount of literal phrases (those not containing metaphor-related words) and amounts to a total of 2,658 examples or phrase-sentence pairs. To assess the algorithm, we sample 50 unique examples per semantic-syntactic type, i.e., 50 literal amod, 50 literal nmod, 50 metaphoric amod, and 50 metaphoric nmod constructions, and tag all the constituent words. The output of the algorithm consists of all the collected semantic tags, ordered by cosine similarity to the target word (Table 1).

TABLE 1 Example output for the two lemmas in the phrase *višjo tarifo* "higher fee."

Lemma	Tags, ordered by similarity
visok [high]	**DOLGO, VISOKO, ŠIROKO [long, high, wide]**
	DOBRO [good]
	VELIKO [big, large]
	ZVOK [sound]
tarifa [fee]	**CENA, STROŠKI [price, expenses]**
	DENAR, FINANCE [money, finance]

Selected tag in bold.

3.7.2 Manual annotation

The 200 examples are annotated by two expert annotators trained in linguistics and metaphor theory. The annotation procedure consists of first reading the example phrase and sentence. Then, for each of the words in the phrase, the task of the annotator is to select the semantic domain of the word's most basic meaning. To do that, the annotator either *(a)* selects 1–3 tags among the automatically provided ones that correspond(s) to the basic domain of the word, or *(b)* annotates with a special label "X" indicating none of the provided tags apply.

Many semantic annotation guidelines prescribe the use of dictionaries or other linguistic resources. In MSDIP (Reijnierse and Burgers, 2023), the annotators consult a dictionary to discern and rank word senses, which facilitates the assignment of semantic domains. However, this procedure cannot be directly applied to Slovene. As previously mentioned, the current reference dictionary for Slovene (see text footnote[2]) is not a contemporary corpus-based resource, and confounds distinct senses. It was thus not possible to strictly follow MIPVU or MSDIP to define senses and associated domains. This is why we decided to adopt a more flexible annotation schema in which the annotators rely on their own linguistic intuition instead of basing their decisions on (unreliable) normative resources. This is also in line with the approach to basic meaning proposed by other researchers (e.g., Nacey et al., 2019; Cruse, 2006). Ideally, the annotation of basic meaning would involve only isolated words, relying on the assumption that annotators can identify the default meaning without contextual cues. However, our evaluation setup provides annotators with words embedded in phrases and sentences for two reasons. First, this aids in disambiguating homonymous words (e.g., *klop* "bench" and *klop* "tick"). Secondly, the dataset is intended to be used for further annotation of target domains, i.e., the domains of the contextual sense of the metaphorically used words.

As demonstrated by the example in Table 1, the extracted phrase *višjo tarifo* "higher fee" consists of the lemmas *visok* "high" and *tarifa* "fee." There are four semantic tags provided for the first lemma and two semantic tags provided for the second. In this example, both annotators agreed that the first provided semantic tag is most appropriate, meaning they would classify the basic meaning of *visok* "tall" as belonging to the domain of DOLGO, VISOKO, ŠIROKO "long, high, wide," and the basic meaning of *tarifa* "fee" into the domain of CENA, STROŠKI "price, expenses."

To assess the overall agreement between the annotators, we calculate perfect and partial percentage agreement. Due to the non-exclusive selection of one or multiple tags, we also calculate the Jaccard index and MASI similarity (Passonneau, 2006) to account for partially matching tags. To further assess the reliability and validity of the annotation, we calculate the inter-annotator agreement (IAA) with Krippendorff's α using either MASI or Jaccard as the distance function to account for actual and expected (dis)agreements as done by, e.g., Passonneau et al. (2006) and Lim and Lauw (2024).

3.7.3 Metric-based evaluation

Using the annotations from the manual evaluation, we evaluate the performance of the proposed algorithm in terms of common accuracy metrics. To create a "silver standard," we consider annotations with partial overlap or in perfect agreement, and aggregate them by choosing intersecting tags. Then, for each of the words in our silver standard, we calculate accuracy (A), precision (P), recall (R), and F1 score at k, where $k \in \{1, 3, 5\}$ represents the number of candidate tags output by the algorithm considered for metric calculation. In other words, we want to assess whether the silver-standard semantic domain(s) appear(s) among the top k candidates provided by the algorithm, and where in the output they appear (i.e., how large does k have to be?). On the one hand, this gives us an indication of whether the bilingual lexical resource pivoting component of the algorithm captures and provides appropriate tags, and, on the other hand, the value of k indicates the quality of the tag ordering using word embeddings and cosine distance.

4 Results and discussion

4.1 Reliability and agreement

First, we calculated the percentage agreement based on all the unique annotated lemmas. As reported in Table 2, the annotators were in perfect agreement for 60% of all lemmas, in partial agreement for 24% of cases, and completely disagreed in 15% of cases. The average MASI similarity between the selected tag sets was 0.675, while the average Jaccard similarity was 0.715. For only partially matching tags, the average MASI amounted to 0.294 and Jaccard to 0.455.

TABLE 2 Percentage agreement (%) and label similarity (MASI/Jaccard) in different agreement types.

Agreement type	*N*	Agreement (%)	MASI	Jaccard
Perfect	121	60.2	1	1
Partial	50	23.4	0.294	0.455
No agreement	30	15.4	0	0
Average			0.675	0.715

TABLE 3 Inter-annotator agreement (Krippendorff's α) calculated on multiple tags vs. first tag-only.

Tags considered	Distance function	α
All	MASI	0.563
All	Jaccard	0.603
First only	Standard nominal	0.732

To further assess the reliability and validity of the annotation, we calculated the inter-annotator agreement (IAA) with Krippendorff's α using either MASI or Jaccard as the distance function. Table 3 reports our results. We obtained $\alpha = 0.563$ (95% CI: 0.493–0.628) using MASI and $\alpha = 0.603$ (95% CI: 0.522, 0.687) using Jaccard. According to previous recommendations (e.g., Artstein and Poesio, 2008; Krippendorff, 2018), the magnitude of α falls below the recommended benchmark ($\alpha \geq 0.67$). This suggests our coding scheme and/or rater calibration require improvement before substantive analysis. However, due to the nature of the annotation and goals of our study, we believe this level of agreement is not necessarily fatal. Compared to most annotation studies, the pool of labels in this study is much larger, and the severity of disagreement is not identical throughout (some labels are more similar than others).[10] For example, by only calculating agreement on the first selected tag by each annotator, it rises to $\alpha = 0.732$ (95% CI: 0.670, 0.792).

To optimize the IAA, annotation campaigns may reconvene the annotators to discuss and resolve disagreements, refine the annotation guidelines, and re-annotate a new data sample. In the next subsection, we analyze and discuss disagreements that emerged in the first annotation round.

4.2 Qualitative evaluation of disagreements

After the first annotation round, the two annotators reconvened to discuss the largest disagreements, i.e., those with completely divergent or minimally overlapping annotations. A few disagreements were simply due to errors in the annotation (wrong label chosen) and thus promptly corrected; a few were re-annotated after deliberation.

Shown in Table 4 is how the inter-annotator agreement slightly improved after the deliberation round and reaching a consensus on some examples. The IAA measured with MASI-based α reached 0.614 (95% CI: 0.536-0.674), and measured with Jaccard-based α it reached 0.655 (95% CI: 0.581–0.720). What is also evident from these results is that even after deliberation, there were instances where the annotators could not reach an agreement.

[10] Some previous semantic/sense annotation studies have opted to group similar, fine-grained senses into supersenses to increase IAA, e.g., Duffield et al. (2007). We could also assign varying weights to disagreements, for example, by putting a lesser weight on disagreements between semantically similar or related domains.

TABLE 4 Inter-annotator agreement (Krippendorff's α) calculated on multiple tags vs. first tag-only (post-deliberation).

Tags considered	Distance function	α
All	MASI	0.614
All	Jaccard	0.655
First only	Standard nominal	0.759

We found four main factors that contributed to annotation disagreements and affected the task in general:

1. **Cognitive-linguistic bias**, by which we refer to idiosyncratic differences in the conceptualization and organization of word meaning, especially for highly polysemous words. The annotators perceived different meanings as more basic or as the "first meaning." An example of such a polyseme was the adjective *blag* "mild," which can be used in various contexts, such as *blaga duša* "gentle soul," *blag okus* "mild taste," *blag vonj* "faint smell," *blag nasmeh* "gentle smile," *blagi ovinek* "soft curve." Another example is the word *svet* "world." One of the annotators chose the domain of VESOLJE "universe" while the other chose ZEMLJEPIS, GEOGRAFIJA "geography." Annotator 1 saw the most basic meaning of the word as an abstract reference to any "world" in the universe or the universe itself, while annotator 2 argued that they deem the basic meaning synonymous with *planet Earth*.
2. **Dictionary override**, where the dictionary diverged from the annotators' intuition of basic meaning. In one example, one of the annotators who was not completely certain of their decision consulted with the dictionary, which overrode their initial decision. As discussed in Section 3.7.2, a word's basic meaning is defined as the meaning assigned in the absence of any contextual cues. For this reason, the annotation guidelines did not anticipate or encourage the use of external resources; however, they also did not explicitly prohibit them.
3. **Knowledge bias**, which stems from specific knowledge of either a discipline (conceptual metaphor research) or a resource (the USAS ontology). For example, both annotators were well versed in conceptual metaphors, which is why in annotating the word *nosilec* "carrier, pylon, girder" the annotators immediately opted to tag it with DELI ZGRADB "parts of buildings" although the basic sense in the dictionary refers to a human carrying something. In another example, one of the annotators, who was more knowledgeable in the USAS categorization, opted for a more specific label outside the provided ones. An example is the adjective *osebnosten* "relating to personality," for which annotator 1 proposed the domain of OSEBNOST "personality," whereas the other annotator was initially content with one of the provided domains, specifically ČUSTVOVANJE, ČUSTVENA STANJA, ČUSTVENI PROCESI "feeling, emotional states, emotional processes." On the other hand, this example also illustrates potential **algorithmic priming**: the algorithm's suggested labels may influence the annotator's choice, whereas without seeing them first, the annotator might select a more general or specific label.
4. **Contextual meaning bias**, which stemmed from the annotation task setting in which the annotators received the word in a phrase and sentence context as opposed to seeing it in isolation. Such a setting was intended to help annotators differentiate between potential homographs and enable later annotations of the target domain. In this regard, we also noticed that annotators had issues differentiating between the general conceptual domain and specific properties or aspects. For example, *senca* "shadow" comes from the domain of SVETLOBA "light," specifically the absence of it. However, the list of potential domains also featured the domain *temno* "dark." The annotators agreed that this is not the basic domain of the concept but represents one of the properties, especially a property that can be transferred when using a metaphor.

These factors partially overlap with insights put forward by Passonneau et al. (2009), who conducted sense annotation for highly polysemous words. They observed that higher specificity of the contexts in which a word appeared and more concrete uses of words led to higher agreement, whereas uses with closely related senses gave rise to lower agreement. Nevertheless, in our study, we are first and foremost interested in identifying the basic meaning of the word, i.e., the out-of-context meaning of a word. Providing the context, if only as a way to disambiguate potential homonyms, occasionally hindered rather than helped the task. Conversely, providing the word in the context where it is used in its basic meaning would presume already knowing the basic meaning and selecting the appropriate example, which would provide an additional source of bias.

Ultimately, both annotators agreed that many examples exist where the subjectivity of the task due to individual linguistic and conceptual differences precludes reaching a consensus. This is in line with many psycholinguistic studies showing individual differences in language acquisition and processing (Lewellen et al., 1993; Kidd et al., 2018; Boland et al., 2016) as well as the differences in the conceptualizations of meaning specifically (Marti et al., 2023; Ramsey, 2021; Stacy et al., 1997).

4.3 Error analysis: words with no suitable tag

In the annotated examples, 12 unique lemmas received the tag "X" by both annotators,[11] indicating none of the provided labels were suitable to represent the basic conceptual domain of the word. This included six adjectives *avtorski* "authorial," *lasten* "own," *pevski* "singing, vocal," *večen* "eternal," *naraven* "natural," *ponorel* "crazed," and six nouns *ime* "name," *ogled* "viewing," *življenje* "life," *luža* "puddle," *dogajanje* "happening," *senca* "shadow."[12] To better understand the gaps in the algorithm or the annotation procedure, we inspected these twelve cases further and report our findings below.

[11] Six in the first round, and an additional six after deliberation.

[12] None of these can be considered a rare word. We checked the frequency of each word in the deduplicated version of Gigafida 2.0 (Krek et al., 2020), the reference corpus of Slovene. The rarest word from this list was *ponorel* "crazed" with a frequency of 1,348.

The first issue we observed is primarily due to **missing or overly specific domains in the USAS lexicon**. For example, the first tag provided by the algorithm for the word *večen* "eternal" was ZAČETNI ČAS "start time." The most suitable tag would belong to the same overarching domain of TIME (T). Although the first provided domain ZAČETNI ČAS "start time" (T2+) stems from a suitable overarching domain, it was considered too specific. In turn, the annotators suggested two alternative domains: a more general OBDOBJE "period" or the more specific DOLGO OBDOBJE "long period."

In the case of the adjective *avtorski* "authorial," several potential issues contribute to the error. First, the word does not feature in OSWN, which is why the algorithm first searched for the nearest neighbors of the word. These were *odrski* "scenery, stage" (ADJ), *avtorsko zaščiten* "author copyrighted," and *avtorski honorar* "author fee/royalty." None of these provided a tag to match the most basic meaning of the word. The annotators agreed that the word should be tagged similarly as the noun *avtor* "author," which it derives from, namely a combination of GENERAL ACTIONS, MAKING, PEOPLE, and INVESTIGATE, EXAMINE, TEST, SEARCH. However, the noun *author* only has two possible domains in USAS, which are BOOKS and PEOPLE. The annotators agree that the domains such as BOOKS and PAPER, DOCUMENTS, WRITING could be suitable but were deemed too specific, as the act of authorship is not necessarily only literary.

Secondly, we observe two cases where the suitable label was not chosen because of the **label description, i.e., the wording of the semantic tag**. In the example of the noun *ime* "name," the seemingly most suitable OEWN equivalents *name* and *given name* provided only the domain GOVORNA DEJANJA "speech acts." The more specific equivalent *first name* provided OBDOBJE "period," and the even more specific *gens* contributed the potential semantic labels SORODSTVO, RODBINA "kin," VRSTE, SKUPINE, PRIMERKI "kinds, groups, examples," and GLASBA "music."[13] In the USAS lexicon, similar words such as *title*, *nickname*, *apellation*, *denominator*, *label*, which share the basic sense of *name*,[14] are indeed tagged with *govorna dejanja* "speech acts." However, a possible label could also be found under the general discourse domain RAZVRŠČANJE "classification" or the subtag VRSTE, SKUPINE, PRIMERKI "kinds, groups, examples." Here, we believe the issue is with the USAS label phrasing. Both annotators believe that the domain labeled GOVORNA DEJANJA "speech acts" specifically denotes actions performed by utterances (i.e., by speaking, enunciating), potentially only suitable for the verbal form *to name*.

We also observed that many errors stemmed from the **lack of mappings in OSWN**. In the next example, a similar contextual meaning is drawn from different source domains in the two languages. The noun *ogled* "viewing, tour" was annotated with semantic tags through several steps. First, OSWN only lists the lexeme *grand tour* as the English equivalent, which does not have a direct semantic tag attributed, therefore taking the semantic tags of its hypernym *tour*. The first tag provided by the algorithm matches the basic meaning of the English *tour*: PREMIKANJE, PRIHAJANJE, ODHAJANJE "moving, coming, going." However, this is not the basic meaning of the Slovene lexeme, which stems not from the domain of MOVING but from the domain of SIGHT. The error could be prevented if OSWN also included the English equivalent *viewing*, which has a suitable tag in the USAS lexicon.

Another example of missing English equivalents can be demonstrated with the word *ponorel* "crazed, deranged." The only equivalent in OEWN is *rampageous*, which provided the tags NASILNO, JEZNO "aggressive, angry" and POŠKODOVANJE, UNIČEVANJE "destroying, damaging." While these can be considered semantic features related to the basic meaning of the word (i.e., a crazy person can be aggressive and damage things) and thus potentially express metonymic or metaphorical usage, the annotators believed these were not the optimal labels to represent the basic domain. Alternatively, the annotators proposed the domain of PSIHOLOŠKA DEJANJA, PSIHOLOŠKA STANJA, PSIHOLOŠKI PROCESI "psychological actions, psychological states, psychological processes." The latter appears in the USAS lexicon under the lexemes *crazy*, *crazed*, and *mad*, so the domain could be captured by the algorithm if these were also provided as equivalents in OSWN.

Lastly, we also note some limitations stemming from the use of **cosine similarity** as the metric to assess semantic similarity between words or phrases. This is the case for the word *Življenje* "life." Although appropriate tags appeared among the tags fetched from the English translations *living* and *life*, such as ŽIVO "alive," the cosine similarity module ranked it much lower than the other candidates and therefore fell outside of the tags provided by the algorithm. Specifically, the cosine similarity between the candidate lexeme *življenje* "life" and the domain word ŽIVO "alive" was calculated at only 0.399, whereas the labels most promoted by cosine similarity VERA, NADNARAVNO "religion, supernatural," DELO, ZAPOSLITEV "work, employment," and PRISTNO "authentic" achieved a similarity value of 0.614, 0.579, and 0.555, respectively. We presume the drift toward other domains may be due to the diverse use of the word in the training data. The word *življenje* "life" strongly collocates with the nouns *način, človek, kakovost, smrt, delo* "manner," "human," "quality," "death," "work," and adjectives *vsakdanje, zasebno, družinsko, skupno, človeško* "everyday," "private," "family," "joint," "human,"[15] which possibly explains the drift in vector space toward the diverse domains promoted by the algorithm. This indicates that the vector space captures a prevalence of the metaphorical meaning in language use and that the basic, literal meaning is less represented.

4.4 Metric-based evaluation

Using the silver-standard data produced in the manual annotation (post-deliberation), we evaluate the algorithm with

[13] Since there were no semantic tags associated directly with the word *gens* in USAS, the indicated domains were gathered from its hypernyms *family* and *folk*.

[14] Defined here as "the word or words that a person, thing, or place is known by," Cambridge English Dictionary, https://dictionary.cambridge.org/, accessed May 1, 2025.

[15] The examples are taken from the list of the strongest collocations with nouns or adjectives from the reference corpus Gigafida 2.0 https://viri.cjvt.si/gigafida, which was used to train the word embeddings.

TABLE 5 Algorithm performance, calculated on first *k* items for *N* = 178 words.

k	1	3	5
A	0.551	**0.775**	0.837
P	**0.551**	0.273	0.176
R	0.533	0.770	**0.831**
F1	**0.538**	0.399	0.289

Best result per metric in bold.

common accuracy metrics. Here, we only consider annotations with partial overlap or in perfect agreement, and aggregate them by choosing intersecting tags.

We calculate accuracy (A), precision (P), recall (R), and F1 score at k, where $k \in \{1, 3, 5\}$ represents the number of candidate tags output by the algorithm considered for metric calculation. We define Accuracy@k (A@k) as the proportion of instances where at least one of the top k predicted tags matches a correct tag. The metrics tell us (a) whether the silver-standard semantic domain(s) appear(s) among the top k candidates, and (b) where in the output they appear (i.e., how large does k have to be?). This gives us an indication of whether the bilingual lexical resource pivoting component of the algorithm captures and provides appropriate tags, and, on the other hand, the value of k indicates the quality of the tag ordering using word embeddings and cosine distance.

We report the results in Table 5. The accuracy (A) is relatively low when only considering the first tag (A@1 = 0.551). However, considering the first 3 or 5 tags provided by the algorithm significantly increases the accuracy, respectively reaching A@3 = 0.775 and A@5 = 0.837. In other words, this means that in more than 83% of cases, at least one tag from the first provided five corresponds to the correct semantic domain. This indicates the first component, pivoting using a bilingual lexicon, provides good candidates for the domains of basic word meaning.

On the other hand, the algorithm achieves somewhat lower Precision (P) and Recall (R). For example, Precision@5 = 0.176 indicates that only 17.6% of the first five tags are correct. However, it is important to note that even with the maximum number of chosen labels (3) according to the annotation procedure, the maximum Precision@5 amounts to 60.0%. The algorithm attained a satisfactory R value at both $k = 3$ and $k = 5$, with the highest R@5 = 0.831. This means that on average, 83.1% of all the silver-standard tags appear among the top 5 candidates. All the metrics calculated at $k = 1$ were quite low (approx. 0.55), indicating that ordering by cosine similarity does not always prioritize the correct tag or the one representing the basic meaning of the word. Nevertheless, by using a larger set of candidates with a k of 3 or 5, the algorithm does provide the analyst with one or more relevant tags.

5 Conclusion

We have presented an approach to basic semantic domain disambiguation that combines the knowledge from multiple lexical and semantic resources in English and applies modern distributional semantic methods to transfer knowledge from English to Slovene. The method provides a reliable and interpretable starting point for semantic annotation, reducing the initial reliance on intuition to formulate and select domain labels. An additional feature of our approach is its robustness, as it proposes domains even for out-of-vocabulary words not covered in the existing resources, thus ensuring better coverage.

Our findings reveal the complexity of attributing conceptual domains to metaphorically used words, both from a methodological and cognitive perspective. We tested the validity of our approach by conducting manual annotation, which included 201 unique words. The inter-annotator agreement (Krippendorff's α between 0.56 and 0.614) falls within the expected range for nuanced, multi-label semantic tasks. Agreement improved after a round of deliberation, but complete consensus could not be achieved. We observe that some disagreements stem from individual linguistic knowledge, intuition, and processing, which influence the conceptualization of a word's basic meaning. This is in line with psycholinguistics studies showing great individual differences in language processing (Kidd et al., 2018; Lewellen et al., 1993; Marti et al., 2023).

Qualitative analysis further highlighted several key sources of disagreement apart from individual linguistic knowledge or intuition. Notably, the inclusion of example sentences in the annotation sometimes confused the annotator, who chose the domain of the word's contextual meaning instead of the basic one. We have also observed some cases where the annotators chose a particular label after consulting a dictionary, which thus "overrode" their initial judgments. This illustrates how lexical resources may diverge from speaker judgment (Joshi et al., 2013) and should thus not always act as a primary reference. This, in particular, holds for languages like Slovene that lack a well-structured, corpus-based dictionary. This experience gained in this study also suggests that future work should more carefully define the annotation procedure to address (or analyze) ambiguity, uncertainty, and biases, which frequently emerge in annotation projects (Beck et al., 2020; Frenda et al., 2024). In the future, we also plan to construct a gold-standard dataset by employing a larger number of annotators, which would, on the one hand, enhance the validity and reliability of the annotations, and, on the other, allow us to gather a wider scope of individual cognitive-linguistic variability.

From a computational perspective, the algorithm performed moderately well in finding candidate tags as it achieved over 83% recall and hit-rate within the top five suggestions. However, the cosine similarity component proved less effective at ranking, as many of the correct tags appeared lower in the list and were not prioritized among the top candidates. In the error analysis, we have uncovered several instances where the embedding-based similarity ranking prioritizes metaphorical senses rather than basic meanings, which most likely reflects the prevalence of metaphorical senses in general language use. This is in line with (Li et al., 2023), who note that word embeddings do not necessarily reflect a single basic meaning, but rather what they describe as "aggregated meaning"—a blend of all word senses, weighted by their frequency in actual language use. Future developments of the method could opt for re-ranking strategies which would draw knowledge from other lexically-related items or from other conceptual resources, such as word concreteness and specificity ratings (e.g., Brysbaert et al., 2014; Ravelli et al., 2025). A major source of error arose from limitations in lexical resources, such as missing or overly specific semantic tags in the USAS semantic lexicon, and missing mappings between Slovene and English lexical items in Open Slovene WordNet. Although our approach was aimed at achieving the greatest possible coverage, further improvements would only be possible with more comprehensive and curated lexical resources and computationally heavier approaches, which were out of the scope of this study.

There are several other avenues we wish to pursue in future research. First, our approach could benefit from using newer, contextual embeddings, which would also allow us to model the contextual meaning, i.e., the target domain activated by the metaphor. Moreover, for a deeper conceptual understanding and reasoning about specific aspects of the domains highlighted in a metaphor, the approach could be expanded using more fine-grained semantic resources such as FrameNet (Boas et al., 2024). Another promising avenue of research that would allow us to overcome the limitations of lexical resources would be to include generative language models. However, while these offer greater flexibility and perhaps better performance, we nevertheless believe our method has advantages in greater interpretability and consistency.

Finally, our results indicate that basic domain attribution is not a straightforward task. The insights from manual annotation spell out the need to carefully consider various factors, such as the annotation setting and the context available to the annotators. Secondly, we believe that some disagreements between the annotators do not necessarily mean failure; on the contrary, they reflect the inherent diversity of speakers, their intuitions, and mental lexicons. Diversity is present even in metaphor theory, where various views exist on how humans process metaphors, which consequently entail different approaches to metaphor analysis. For this reason, we believe future metaphor research, as well as similar cognitive-linguistic endeavors, should be based on diverse sources, combining and cross-validating data from automatically produced annotations, psycholinguistic studies, dictionaries, and corpora.

Data availability statement

Data is available from the corresponding author upon request. Requests to access these datasets should be directed to mojca.brglez@ff.uni-lj.si.

Author contributions

MB: Investigation, Conceptualization, Writing – review & editing, Data curation, Writing – original draft, Visualization, Methodology. KP: Writing – review & editing, Data curation, Conceptualization.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This work received financial support from the Slovenian Research Agency (ARRS/ARIS) under the research core funding P6-0215, the research core funding P6-0436, and under the project Large language models for Digital Humanities (LLM4DH), project code GC-0002.

Acknowledgments

The authors thank Professor Špela Vintar for providing supervision and reviewing the manuscript draft.

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.

Generative AI statement

The author(s) declare that Gen AI was used in the creation of this manuscript. To improve fluency and style of the manuscript text.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

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.

References

Antloga, P. (2020). *Metaphor Corpus KOMET 1.0. Slovenian Language Resource Repository CLARIN.SI*. Available online at: http://hdl.handle.net/11356/1293

Artstein, R., and Poesio, M. (2008). Survey article: Inter-coder agreement for computational linguistics. *Comput. Ling*. 34, 555–596. doi: 10.1162/coli.07-034-R2

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). "The Berkeley FrameNet project," in *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1* (Montreal, QC: Association for Computational Linguistics), 86–90. doi: 10.3115/980845.980860

Beck, C., Booth, H., El-Assady, M., and Butt, M. (2020). "Representation problems in linguistic annotations: ambiguity, variation, uncertainty, error and bias," in *Proceedings of the 14th Linguistic Annotation Workshop*, eds. S. Dipper, and A. Zeldes (Barcelona: Association for Computational Linguistics), 60–73.

Boas, H. C., Ruppenhofer, J., and Baker, C. (2024). Framenet at 25. *Int. J. Lexicography* 37, 263–284. doi: 10.1093/ijl/ecae009

Boland, J. E., Kaan, E., Kroff, J. V., and Wulff, S. (2016). Psycholinguistics and variation in language processing. *Linguist. Vanguard* 2:20160064. doi: 10.1515/lingvan-2016-0064

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

Brglez, M., and Pahor de Maiti, K. (2024). "Conceptual domain disambiguation for metaphor identification and interpretation," in *Sprache & Einstellung (DGfS 2024 Workshop)* (Bochum), 155.

Brglez, M., and Vintar, P. (2025). In search of semantic distance: metaphorical and non-metaphorical constructions in static and contextual embeddings. *J. Lang. Model*. 13, 207–260. doi: 10.15398/jlm.v13i2.437

Brysbaert, M., Warriner, A. B., and Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behav. Res. Methods* 46, 904–911. doi: 10.3758/s13428-013-0403-5

Cameron, L. (2003). *Metaphor in Educational Discourse*. London: Continuum.

Čibej, J., Terčon, L., Krek, S., Repar, A., Novak, E., Gantar, P., et al. (2023). *Open Slovene WordNet OSWN 1.0. Slovene language resource repository CLARIN.SI*. Available online at: http://hdl.handle.net/11356/1888

Clausner, T. C., and Croft, W. (1997). Productivity and schematicity in metaphors. *Cogn. Sci*. 21, 247–282. doi: 10.1207/s15516709cog2103_1

Croft, W. (2003). "The role of domains in the interpretation of metaphors and metonymies," in *Metaphor and Metonymy in Comparison and Contrast*, eds. R. Dirven, and R. Pörings (Berlin: Mouton de Gruyter), 161–206. doi: 10.1515/9783110219197.2.161

Cruse, A. (1986). *Lexical Semantics. Cambridge Textbooks in Linguistics*. Cambridge: Cambridge University Press.

Cruse, A. (2006). *A Glossary of Semantics and Pragmatics*. Edinburgh: Edinburgh University Press. doi: 10.1515/9780748626892

Deignan, A. (2005). *Metaphor and Corpus Linguistics, Volume 6 of Converging Evidence in Language and Communication Research*. Amsterdam: John Benjamins. doi: 10.1075/celcr.6

Deignan, A. (2008). "Corpus linguistics and metaphor," in *The Cambridge Handbook of Metaphor and Thought*, ed. R. W. Gibbs Jr. (Cambridge, MA: Cambridge University Press), 280–294. doi: 10.1017/CBO9780511816802.018

Deignan, A. (2016). *Metaphor and Corpus Linguistics*. Amsterdam: John Benjamins.

Demmen, J., Semino, E., Demjén, Z., Koller, V., Hardie, A., Rayson, P., et al. (2015). A computer-assisted study of the use of violence metaphors for cancer and end of life by patients, family carers and health professionals. *Int. J. Corpus Linguist*. 20, 205–231. doi: 10.1075/ijcl.20.2.03dem

Dobrovol'skij, D., Piirainen, E. (2022). *Figurative Language: Cross-Cultural and Cross-Linguistic Perspectives*. Berlin: De Gruyter Mouton. doi: 10.1515/9783110702538

Dodge, E., Hong, J., and Stickles, E. (2015). "MetaNet: deep semantic automatic metaphor analysis," in *Proceedings of the Third Workshop on Metaphor in NLP* (Denver, CO: Association for Computational Linguistics), 40–49. doi: 10.3115/v1/W15-1405

Duffield, C. J., Hwang, J. D., Brown, S. W., Dligach, D., Vieweg, S. E., Davis, J., et al. (2007). "Criteria for the manual grouping of verb senses," in *Proceedings of the Linguistic Annotation Workshop*, eds. B. Boguraev, N. Ide, A. Meyers, S. Nariyama, M. Stede, J. Wiebe, et al. (Prague: Association for Computational Linguistics), 49–52. doi: 10.3115/1642059.1642067

Eilts, C., and Lönneker, B. (2002). The Hamburg metaphor database. *Metaphorik.de* 3, 100–110.

Fauconnier, G., and Turner, M. (2002). *The Way We Think: Conceptual Blending and the Mind's Hidden Complexities*. New York, NY: Basic Books.

Feldman, J. (2008). *From Molecule to Metaphor: A Neural Theory of Language*. Cambridge, MA: MIT press.

Fellbaum, C. (2005). "WordNet and wordnets," in *Encyclopedia of Language and Linguistics*, ed. K. Brown (Oxford: Elsevier), 665–670. doi: 10.1016/B0-08-044854-2/00946-9

Fellbaum, C. (Ed.) (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/7287.001.0001

Firth, J. (1957). "A synopsis of linguistic theory 1930–1955," in *Studies in Linguistic Analysis*. (Oxford: Philological Society), 1–32.

Frenda, S., Abercrombie, G., Basile, V., Pedrani, A., Panizzon, R., Cignarella, A., et al. (2024). Perspectivist approaches to natural language processing: a survey. *Lang. Resour. Eval*. 59, 1719–1746. doi: 10.1007/s10579-024-09766-4

Ge, M., Mao, R., and Cambria, E. (2022). "Explainable metaphor identification inspired by conceptual metaphor theory," in *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence, Vol. 36* (Vancouver, BC), 10681–10689. doi: 10.1609/aaai.v36i10.21313

Gentner, D. (1983). Structure-mapping: a theoretical framework for analogy. *Cogn. Sci*. 7, 155–170. doi: 10.1016/S0364-0213(83)80009-3

Gibbs, R., Lima, P., and Françozo, E. (2004). Metaphor is grounded in embodied experience. *J. Pragmat*. 36, 1189–1210. doi: 10.1016/j.pragma.2003.10.009

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

Glucksberg, S. (2001). *Understanding Figurative Language: From Metaphors to Idioms*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780195111095.001.0001

Group, P. (2007). MIP: a method for identifying metaphorically used words in discourse. *Metaphor Symb*. 22, 1–39. doi: 10.1080/10926480709336752

Hanks, P. (2006). "Metaphoricity is gradable," in *Corpora in Cognitive Linguistics. Vol. 1: Metaphor and Metonymy* (Boston, MA: Mouton de Gruyter), 17–35. doi: 10.1515/9783110199895.17

Harris, Z. S. (1954). Distributional structure. *Word* 10, 146–162. doi: 10.1080/00437956.1954.11659520

Hicke, R. M. M., and Kristensen-McLachlan, R. D. (2024). "Science is exploration: computational frontiers for conceptual metaphor theory," in *Proceedings of CHR 2024: Computational Humanities Research Conference* (Aarhus), 1105–1116.

Ichien, N., Stamenkovic, D., and Holyoak, K. (2024). Large language model displays emergent ability to interpret novel literary metaphors. *Metaphor Symb*. 39, 296–309. doi: 10.1080/10926488.2024.2380348

Joshi, S., Kanojia, D., and Bhattacharyya, P. (2013). "More than meets the eye: study of human cognition in sense annotation," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, eds. L. Vanderwende, H. Daumé III, and K. Kirchhoff (Atlanta, GA: Association for Computational Linguistics), 733–738.

Kidd, E., Donnelly, S., and Christiansen, M. H. (2018). Individual differences in language acquisition and processing. *Trends Cogn. Sci*. 22, 154–169. doi: 10.1016/j.tics.2017.11.006

Koller, V., Hardie, A., Semino, E., and Rayson, P. (2008). Using a semantic annotation tool for the analysis of metaphor in discourse. *Metaphorik.de* 15, 141–160.

Kövecses, Z. (2008). Conceptual metaphor theory: some criticisms and alternative proposals. *Ann. Rev. Cogn. Linguist*. 6, 168–184. doi: 10.1075/arcl.6.08kov

Kövecses, Z. (2020). *Extended Conceptual Metaphor Theory*. Cambridge University Press. doi: 10.1017/9781108859127

Krek, S., Arhar Holdt, P., Erjavec, T., Čibej, J., Repar, A., Gantar, P., et al. (2020). "Gigafida 2.0: the reference corpus of written standard Slovene," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, eds. N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, et al. (Marseille: European Language Resources Association), 3340–3345.

Krennmayr, T. (2013). Adding transparency to the identification of cross-domain mappings in real language data. *Rev. Cogn. Linguist*. 11, 163–184. doi: 10.1075/rcl.11.1.05kre

Krippendorff, K. (2018). *Content Analysis: An Introduction to Its Methodology*, 4th Edn. Thousand Oaks, CA: SAGE Publications.

Lakoff, G. (2008). "The neural theory of metaphor," in *The Cambridge Handbook of Metaphor and Thought, Cambridge Handbooks in Psychology*, ed. R. W. Gibbs (Cambridge, MA: Cambridge University Press), 17–38. doi: 10.1017/CBO9780511816802.003

Lakoff, G. (2014). Mapping the brain's metaphor circuitry: Metaphorical thought in everyday reason. *Front. Hum. Neurosci*. 8:958. doi: 10.3389/fnhum.2014.00958

Lakoff, G., Espenson, J., and Schwartz, A. (1991). *Master Metaphor List*, 2nd Edn. Technical report. Berkeley, CA: Cognitive Linguistics Group, University of California, Berkeley.

Lakoff, G., and Johnson, M. (1980). *Metaphors We Live By*. Chicago, IL: University of Chicago Press.

Lewellen, M. J., Goldinger, S. D., Pisoni, D. B., and Greene, B. G. (1993). Lexical familiarity and processing efficiency: individual differences in naming, lexical decision, and semantic categorization. *J. Exp. Psychol. Gen.* 122, 316–330. doi: 10.1037/0096-3445.122.3.316

Li, Y., Wang, S., Lin, C., and Guerin, F. (2023). "Metaphor detection via explicit basic meanings modelling," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, eds. A. Rogers, J. Boyd-Graber, and N. Okazaki (Toronto, ON: Association for Computational Linguistics), 91–100. doi: 10.18653/v1/2023.acl-short.9

Lim, J. P., and Lauw, H. W. (2024). Aligning human and computational coherence evaluations. *Comput. Linguist.* 50, 893–952. doi: 10.1162/coli_a_00518

Löfberg, L., Piao, S., Rayson, P., Juntunen, J.-P., Nykänen, A., Varantola, K., et al. (2005). "A semantic tagger for the Finnish language," in *Proceedings from the Corpus Linguistics Conference Series Online e-Journal* (Birmingham), 1.

Marti, L., Wu, S., Piantadosi, S., and Kidd, C. (2023). Latent diversity in human concepts. *Open Mind* 7, 1–14. doi: 10.1162/opmi_a_00072

Maslen, L. C. R., editor (2010). *Metaphor Analysis: Research Practice in Applied Linguistics, Social Sciences and the Humanities*. Toronto, ON: University of Toronto Press.

Mason, Z. J. (2004). CorMet: A computational, corpus-based conventional metaphor extraction system. *Comput. Linguist.* 30, 23–44. doi: 10.1162/089120104773633376

McCrae, J. P., Rademaker, A., Bond, F., Rudnicka, E., and Fellbaum, C. (2019). "English WordNet 2019-an open-source wordnet for English," in *Proceedings of the 10th Global Wordnet Conference* (Wrocław: Global Wordnet Association), 245–252.

Michelli, G., Tong, X., and Shutova, E. (2024). A framework for Annotating and Modelling Intentions Behind Metaphor Use. *Comput. Res. Reposit.* arXiv:2407.03952. doi: 10.48550/arXiv.2407.03952

Nacey, S., Dorst, A. G., Krennmayr, T., and Reijnierse, W. G. editors (2019). *Metaphor Identification in Multiple Languages: MIPVU around the World*. Amsterdam: John Benjamins Publishing Company. doi: 10.1075/celcr.22

Passonneau, R., Salleb-Aouissi, A., and Ide, N. (2009). "Making sense of word sense variation," in *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009)*, eds. E. Agirre, L. Màrquez, and R. Wicentowski (Boulder, CO: Association for Computational Linguistics), 2–9. doi: 10.3115/1621969.1621972

Passonneau, R. J. (2006). "Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation," in *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)* (Genoa: European Language Resources Association), 831–836.

Passonneau, R. J., Habash, N., and Rambow, O. (2006). "Inter-annotator agreement on a multilingual semantic annotation task," in *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)* (Genoa: European Language Resources Association), 1951–1956.

Pedersen, B. S., Sørensen, N., Nimb, S., Hansen, D. H., Olsen, S., and Al-Laith, A. (2025). "Evaluating LLM-generated explanations of metaphors - a culture-sensitive study of Danish," in *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, eds. R. Johansson, and S. Stymne (Tallinn: University of Tartu Library), 470–479.

Piao, S., Bianchi, F., Dayrell, C., D'Egidio, A., and Rayson, P. (2015). "Development of the multilingual semantic annotation system," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Denver, CO: Association for Computational Linguistics), 1268–1274. doi: 10.3115/v1/N15-1137

Puraivan, E., Renau, I., and Riquelme, N. (2024). Metaphor identification and interpretation in corpora with ChatGPT. *SN Comput. Sci.* 5:976. doi: 10.1007/s42979-024-03331-0

Ramsey, R. (2021). Individual differences in word senses. *Cogn. Linguist.* 33, 65–93. doi: 10.1515/cog-2021-0020

Ravelli, A. A., Bolognesi, M. M., and Caselli, T. (2025). Specificity ratings for English data. *Cogn. Process.* 26, 283–302. doi: 10.1007/s10339-024-01239-4

Rayson, P., Archer, D., Piao, S., and McEnery, T. (2004). "The UCREL semantic analysis system," in *Proceedings of the Workshop on Beyond Named Entity Recognition Semantic Labelling for NLP Tasks* (Lisbon), 7–12.

Reijnierse, W. G., and Burgers, C. (2023). MSDIP: a method for coding source domains in metaphor analysis. *Metaphor Symb.* 38, 295–310. doi: 10.1080/10926488.2023.2170753

Ritchie, D. (2003). "ARGUMENT IS WAR"—or is it a game of chess? Multiple meanings in the analysis of implicit metaphors. *Metaphor Symb.* 18, 125–146. doi: 10.1207/S15327868MS1802_4

Semino, E. (2018). *Metaphor, Cancer and the End of Life: A Corpus-Based Study*. London: Routledge. doi: 10.4324/9781315629834

Sengupta, M., Alshomary, M., Scharlau, I., and Wachsmuth, H. (2023). "Modeling highlighting of metaphors in multitask contrastive learning paradigms," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, ed.s H. Bouamor, J. Pino, and K. Bali (Singapore: Association for Computational Linguistics), 4636–4659. doi: 10.18653/v1/2023.findings-emnlp.308

Sengupta, M., Alshomary, M., and Wachsmuth, H. (2022). "Back to the roots: predicting the source domain of metaphors using contrastive learning," in *Proceedings of the Third Workshop on Figurative Language Processing (FLP)* (Abu Dhabi), 137–142. doi: 10.18653/v1/2022.flp-1.19

Shaikh, S., Strzalkowski, T., Cho, K., Liu, T., Broadwell, G. A., Feldman, L., et al. (2014). "Discovering conceptual metaphors using source domain spaces," in *Proceedings of the 4th Workshop on Cognitive Aspects of the Lexicon (CogALex)*, eds. M. Zock, R. Rapp, and C. Huang (Dublin: Association for Computational Linguistics and Dublin City University), 210–220. doi: 10.3115/v1/W14-4725

Shutova, E. (2010). "Automatic metaphor interpretation as a paraphrasing task," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, eds. R. Kaplan, J. Burstein, M. Harper, and G. Penn (Los Angeles, CA: Association for Computational Linguistics), 1029–1037.

Shutova, E., and Teufel, S. (2010). "Metaphor corpus annotated for source - target domain mappings," in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)* (Valletta: European Language Resources Association), 3255–3261.

Stacy, A. W., Leigh, B. C., and Weingardt, K. (1997). An individual-difference perspective applied to word association. *Pers. Soc. Psychol. Bull.* 23, 229–237. doi: 10.1177/0146167297233002

Steen, G. (2007). "Finding metaphor in grammar and usage: a methodological analysis of theory and research," in *Converging Evidence in Language and Communication Research* (Amsterdam: John Benjamins Publishing Company). doi: 10.1075/celcr.10

Steen, G., Dorst, A. G., Herrmann, J. B., Kaal, A. A., Krennmayr, T., Pasma, T., et al. (2010). *A Method for Linguistic Metaphor Identification: From MIP to MIPVU, Volume 14 of Converging Evidence in Language and Communication Research*. Amsterdam: John Benjamins Publishing Company. doi: 10.1075/celcr.14

Steen, G. J. (1999). "From linguistic to conceptual metaphor in five steps," in *Metaphor in Cognitive Linguistics: Selected Papers from the International Conference on Cognitive Linguistics*, eds. D. Geeraerts, and H. Cuyckens (Amsterdam: John Benjamins), 51–77. doi: 10.1075/cilt.175.05ste

Sullivan, K. (2009). "Grammatical constructions in metaphoric language," in *Studies in Cognitive Corpus Linguistics*, eds. B. Lewandowska-Tomaszczyk, and K. Dziwirek (Frankfurt am Main: Peter Lang), 57–80.

Terčon, L., Ljubešić, N., and Erjavec, T. (2023). *Word embeddings CLARIN.SI-embed.sl 2.0. Slovene language resource repository CLARIN.SI*. Available online at: http://hdl.handle.net/11356/1791

Tian, Y., Xu, N., and Mao, W. (2024). "A theory guided scaffolding instruction framework for LLM-enabled metaphor reasoning," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, eds. K. Duh, H. Gomez, and S. Bethard (Mexico City: Association for Computational Linguistics), 7738–7755. doi: 10.18653/v1/2024.naacl-long.428

Vervaeke, J., and Kennedy, J. M. (1996). Metaphors in language and thought: Falsification and multiple meanings. *Metaphor Symb. Act.* 11, 273–284. doi: 10.1207/s15327868ms1104_3

Wachowiak, L., and Gromann, D. (2023). "Does GPT-3 grasp metaphors? Identifying metaphor mappings with generative language models," in Proceedings *of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, eds. A. Rogers, J. Boyd-Graber, and N. Okazaki (Toronto, ON: Association for Computational Linguistics), 1018–1032. doi: 10.18653/v1/2023.acl-long.58

Wang, Z., Peng, S., Chen, J., Zhang, X., and Chen, H. (2023). ICAD-MI: interdisciplinary concept association discovery from the perspective of metaphor interpretation. *Knowl. Based Syst.* 275:110695. doi: 10.1016/j.knosys.2023.110695