Song, Chenchen (2024) Two linguistic levels of lexical ambiguity and a unified categorical representation. Journal of Computer-Assisted Linguistic Research 8: 77–107. ISSN: 2530-9455.

Two linguistic levels of lexical ambiguity and a unified categorical representation

Chenchen Song*
Zhejiang University, P.R. China
*Corresponding author: cjs021@zju.edu.cn

Received: 2 September 2024 / Accepted: 31 October 2024 / Published: 15 November 2024

Abstract
Lexical disambiguation is one of the oldest problems in natural language processing. There are three main types of lexical ambiguity: part-of-speech ambiguity, homonymy, and polysemy, typically divided into two tasks in practice. While this division suffices for engineering purposes, it does not align well with human intuition. In this article, I use lexical ambiguity as a representative case to demonstrate how insights from theoretical linguistics can be helpful for developing more human-like meaning and knowledge representations in natural language understanding. I revisit the three types of lexical ambiguity and propose a structured reclassification of them into two levels using the theoretical linguistic tool of root syntax. Recognizing the uneven expressive power of root syntax across these levels, I further translate the theoretical linguistic insights into the language of category theory, mainly using the tool of topos. The resulting unified categorical representation of lexical ambiguity preserves root-syntactic insights, has strong expressive power at both linguistic levels, and can potentially serve as a bridge between theoretical linguistics and natural language understanding.

Keywords: lexical ambiguity, natural language understanding, meaning representation, theoretical linguistics, category theory

1. Introduction

Lexical ambiguity is a fundamental characteristic of human language (Edmonds 2006). According to Rodd, Gaskell, and Marslen-Wilson (2004), approximately 84% of common words in English have multiple dictionary-registered readings. Given its fundamental status, lexical ambiguity has received much attention in several language-related areas of research, including linguistics (e.g., Lyons 1977; Cruse 1986; Valera 2020), psycholinguistics (e.g., Rodd, Gaskell, and Marslen-Wilson 2004; Armstrong and Plaut 2016; Rodd 2018), and natural language processing (NLP; e.g., Agirre and Edmonds 2007; Navigli 2009; Bevilacqua et al. 2021). See

DOI: HTTPS://DOI.ORG/10.4995/JCLR.2024.22348
Publisher: Universitat Politecnica de Valencia.

Cassani et al. (2023) and Haber and Poesio (2024) for recent cross-disciplinary overviews.[1]

Lexical ambiguity comes in different types. A basic distinction that is recognized in all the above areas is that between homonymy and polysemy, which respectively refer to ambiguity between unrelated meanings of a word form and ambiguity between semantically related word senses (Rodd 2018). A lesser-mentioned yet equally fundamental ambiguity type is part-of-speech (POS) ambiguity. See (1) for an illustration.

(1) a. *sharp* adj. having a thin edge *vs.* n. a musical notation (POS ambiguity)

b. *bank*$_1$ n. the higher ground along a river *vs.* (homonymy)

bank$_2$ n. a financial institution

c. *bank*$_2$ n. a company or institution *vs.* a building (polysemy)

The subscripts 1 and 2 in (1b) indicate that the two meanings of the word form *bank* are unrelated; by comparison, the two senses in (1c) are related and are both subsumed under the financial use of *bank* (i.e., *bank*$_2$). When the unrelated uses of a word form also have different pronunciations, they are no longer homophones but merely homographs. See (2) for an illustration.

(2) *bow*$_1$ /baʊ/ n. the front part of a boat or ship *vs.* (homography)

bow$_2$ /boʊ/ n. an arrow-shooting weapon

Homonymy and polysemy do not always have a clear boundary, because semantic relatedness is often a matter of degree (Geeraerts 1993; Rodd 2020; Cevoli et al. 2023). Thus, while the two readings of *bank*$_2$ in (1c) are clearly related, the relatedness of the two readings of *mouse*—'an animal' vs. 'an IT device'—is less clear. In general, semantic relatedness based on metaphorical extension (as in *mouse*) is less perceptible than that based on metonymic extension (as in *bank*$_2$). These two scenarios respectively correspond to what are called irregular/accidental and regular polysemy (Apresjan 1974; Vicente and Falkum 2017). Some common alternations in regular polysemy are animal~meat (e.g., *chicken*), container~content (e.g., *bottle*), and physical~informational (e.g., *book*) (Haber and Poesio 2024). In particular, the last alternation involves two aspects that are equally basic and inherent to the entity denoted by the word, with neither sense being an extension of the other. This type of regular polysemy is also called inherent or logical polysemy (Pustejovsky 1995; Asher 2011; Dölling 2020).

Lexical disambiguation is a long-standing challenge for NLP, where it is more widely known as word sense disambiguation (WSD). As noted in Agirre and Edmonds (2007) and Navigli (2009), WSD is an AI-complete problem, which means that it is as hard as the central problems of artificial intelligence (e.g., the Turing Test). Among the three types of lexical ambiguity, POS ambiguity is relatively easy to resolve thanks to the high performance of modern POS-taggers; thus, words are usually assumed to be already POS-tagged in the WSD task (Navigli 2009, 6). Indeed, in most WSD systems, POS-tagging is simply treated as an initial step, with the WSD

[1] Grammatical labels used in this article: CL = classifier, DISP = disposal, E = emotion, *E* = entity type, *Ev* = eventuality type, *n* = nominalizer, *p* = prepositonalizer, PASS = passive, POSS = possessive, PROG = progressive, *v* = verbalizer

algorithm being dedicated to within-POS ambiguity (Edmonds 2006, 609).

The way lexical ambiguity is dealt with in NLP is task-oriented: POS ambiguity is linked to the POS-tagging task, and the other ambiguity types are linked to the WSD task. While this strategy can successfully get the job done—which it additionally does strikingly well with recent neural language models—it does not really align well with human intuition. First, the separation of POS-tagging and WSD into two tasks presupposes a clear-cut boundary between POS ambiguity and other ambiguity types, when in reality these are often intertwined, as in (3).

(3) a. *stalk* n. part of a plant *vs.* (POS ambiguity + homonymy)
v. to follow a person
b. *ship* n. a large boat *vs.* (POS ambiguity + polysemy)
v. send to a customer

Second, while the distinction between homonymy and polysemy is as commonly known in NLP as it is in linguistics, computational research on WSD has not paid due attention to this distinction in practice (Haber and Poesio 2024, 353). In traditional knowledge-based WSD models, all senses are enumerated on the same level in the knowledge base (e.g., WordNet) regardless of relatedness degree (Falkum and Vicente 2015). Meanwhile, in currently popular neural language models, all words (or more exactly tokens) are converted to (contextualized) vector embeddings in the same way, with no representational distinction in the output for homonymous and polysemous words. Admittedly, the distinction may be partly recovered from the output semantic space (Beekhuizen, Armstrong, and Stevenson 2021; Cevoli et al. 2023) or the inner workings of the neural architecture (Cassani et al. 2023). However, without any ontology or knowledge base, the recovered (meta)information about word senses cannot be organized in a systematic, structured way. To overcome this drawback, it may be fruitful to build hybrid models that integrate the neural architecture with knowledge bases; this is indeed what is currently being explored in the field (see Section 2).

Lenci (2023, 14) concludes that human-like natural language understanding (NLU) may come from integrating data-driven representations with symbolic systems. Other scholars have expressed similar views. For instance, Bender and Koller (2020, 5185) remark that while large language models "may well end up being important components of an eventual full-scale solution to human-analogous NLU, they are not nearly-there solutions to this grand challenge." Similarly, Van Valin (2016, 2) argues that to "start down the path from NLP to NLU we have to go back to linguistics." Thus, inasmuch as human-like NLU is still the Holy Grail of AI (Lenci 2023, 1), insights from symbolic approaches to meaning representation are still relevant.

Against this backdrop, the goal of this article is to demonstrate, via the case of lexical ambiguity, how insights from theoretical linguistics can help NLP/NLU researchers achieve more human-like representations of meaning. As we will see, the linguistically motivated formal representations of lexical ambiguity are more nuanced and structured, with the aforementioned three types being subdivided and regrouped into two levels: (i) the categorization level (for homonymy and cross-POS polysemy), and (ii) the post-categorization level (for word sense polysemy within the same POS). This regrouping is more consistent with how word meanings are treated in current theoretical linguistics, where a formal tool known as root syntax (Halle and Marantz 1993, 1994, et seq.) provides a principled method to distinguish the levels and types of lexical ambiguity.

However, this linguistic tool alone is insufficient for our goal of channeling theoretical linguistic insights into NLU system building, because the expressive power of root syntax is uneven at the two levels of lexical ambiguity—it is much more expressive at the categorization level than at the post-categorization level. On the other hand, we must also recognize the ever-larger gap between modern-day linguistics and NLP/NLU. To bring us closer to the goal of human-like NLU, I use category theory (Eilenberg and Mac Lane 1945 et seq.) as a bridge, which provides a tool set that is "useful throughout science" (Spivak 2014, 3). Category theory has already been used in NLP, most representatively in a line of work by Bob Coecke and his colleagues (Coecke, Sadrzadeh, and Clark 2010 et seq.). It has been applied to lexical semantics too—for instance, in Asher (2011) and Babonnaud (2019, 2021, 2022). The current article can be considered an addition to the literature on categorical linguistics, which lies between categorical NLP and formal linguistics. Specifically, I will use the category-theoretic tool of topos (following Asher and Babonnaud) to provide a unified formal representation of lexical ambiguity that is intuitive and expressive at both levels mentioned above, which moreover preserves insights from root syntax.

The rest of this article is organized as follows. In Section 2, I describe the state of the art of WSD research. In Section 3, I revisit the task-oriented classification of lexical ambiguity and propose a more structured reclassification using the theoretical linguistic tool of root syntax. In Section 4, I further develop that reclassification into a unified formal representation using category theory. In Section 5, I make a preliminary proposal as to how the formal representations of meaning developed here can be integrated into WSD systems. Section 6 concludes.

2. Approaches to word sense disambiguation

As mentioned in Section 1, lexical disambiguation in NLP is divided into two tasks: POS-tagging and WSD. Thus, the various approaches discussed below are all focused on word senses instead of word categories. There are three main approaches to WSD: knowledge-based, supervised, and unsupervised (Agirre and Edmonds 2007; Navigli 2009; Camacho-Collados and Pilehvar 2018; Bevilacqua et al. 2021). The knowledge-based approach relies on external knowledge sources like dictionaries and thesauri and algorithmically maps each target word to one of its senses listed in the external dictionary. Popular external knowledge sources include WordNet (Miller 1995) and BabelNet (Navigli and Ponzetto 2012). Purely knowledge-based models do not use corpus data at all, which is what distinguishes them from supervised and unsupervised models (both corpus-based). Supervised models use annotated corpora (e.g., SemCor, Miller et al. 1993) in combination with machine-learning techniques to induce a classifier for each ambiguous word, while unsupervised models just use unannotated, raw corpora (e.g., the British National Corpus, Clear 1993) and induce word senses from scratch either by clustering similar contextual sentences or by clustering similar neighboring words (Navigli 2009; Lenci and Sahlgren 2023). Note that in models adopting the word sense induction

(WSI) strategy, the primary task is no longer word sense disambiguation but rather word sense discrimination.

Among the three approaches to WSD, the supervised approach has shown the best performance (Raganato, Camacho-Collados, and Navigli 2017; Pasini 2020; Bevilacqua et al. 2021). There are also hybrid models that combine the merits of the knowledge-based and the supervised approach, such as the GAS model in Luo et al. (2018) and the ARES model in Scarlini, Pasini, and Navigli (2020). Such hybrid models are reportedly the best solution to WSD so far and thus constitute a promising direction of continued research (Barba, Pasini, and Navigli 2021; Bevilacqua et al. 2021).

In both supervised and unsupervised models, word meanings are commonly represented as vectors. However, conventional vector-based word representations or embeddings (e.g., those yielded by word2vec, Mikolov et al. 2013) are not so useful in WSD due to the meaning conflation deficiency (Schütze 1998; Camacho-Collados and Pilehvar 2018)—that is, the problem that multiple meanings of a word are all conflated into a single vector. With recent advancements in NLP and especially the advent of the transformer architecture (Vaswani et al. 2017), this is not a problem anymore, because conventional static embeddings for word types are now replaced by dynamic and contextualized embeddings for word tokens. In fact, with this paradigm shift, even the WSD task itself may seem to be no longer needed, as word senses are now disambiguated to the finest degree by default.

That said, the fully token-based mode of representation is probably taking things to the other extreme, because it amounts to defining a sense for each individual occurrence of a word in a gigantic corpus and listing all such "micro-senses" (Cruse 2000) as separate dictionary entries. The lack of structure here is even worse than that in the flatly organized external dictionaries (e.g., WordNet) used in traditional knowledge-based models. While traditional WSD models merely blur the line between homonymy and polysemy, the token-based micro-senses furthermore blur the line between polysemy and vagueness (see Tuggy 1993, Agirre and Edmonds 2007, and Haber and Poesio 2024 for discussions on vagueness and word sense granularity). In short, as Lenci (2023, 11) points out, contextual embeddings "show the very same limit of former types of static distributional representations" because they still "lack the ability to represent and organize what they acquire from texts through distributional learning into proper knowledge structures."

The upshot of the above discussion is that despite their highly impressive performance from an engineering perspective, current neural language models are no more human-like in their "understanding" of lexical ambiguity than their predecessors. To move closer to human-level understanding, current models can benefit from more structured linguistic knowledge, such as the theoretical linguistic structures to be discussed in the following sections.

3. Revisiting lexical ambiguity from a linguistic perspective

In this section, I revisit the classification of lexical ambiguity in WSD from a linguistic perspective and propose an alternative classification involving two broad levels. Moreover, theoretical linguistics has a well-established tool (root syntax) to formally represent and distinguish these two levels as well as their internal nuances. In what follows, I will first promote a perspective shift (Section 3.1), then present the details of my alternative classification (Section 3.2), and lastly discuss an important further ambiguity type (i.e., semilexicality) that is naturally covered by my reclassification but systematically neglected in NLP/NLU research so far (Section 3.3).

3.1. A meaning-first approach

The task-oriented conception of lexical ambiguity in WSD reflects a form-over-meaning mindset, where the question being asked is: Given a word form like *bank*, how to pin down its meaning? This mindset is tied to the task of parsing, which is of vital importance in NLP. However, NLU is more than just text-processing. To attain human-like understanding, we must adopt a conception of language that is aligned with how linguistic knowledge is organized in the human mind. A mainstream view in current theoretical linguistics is that human language works in a meaning-over-form fashion. In Berwick and Chomsky's (2016, 101) words, "language […] is fundamentally a system of meaning. Aristotle's classic dictum that language is sound with meaning should be reversed. Language is meaning with sound (or some other externalization, or none)." To achieve a more human-like representation of lexical ambiguity, therefore, we need to pursue a meaning-first approach. In particular, we need to give homonyms separate lexical entries and treat POS ambiguity as an integral part of polysemy. In this way, we obtain an organization of lexical knowledge of the form illustrated in (4), which is essentially just how word senses are organized in dictionaries.

(4) a. *bank*$_1$ n. a mound, pile, or ridge raised above the surrounding level;
the rising ground bordering a lake, river, or sea; …
vt. to raise a bank about;
to heap or pile in a bank; …
vi. to rise in or form a bank; …
b. *bank*$_2$ n. an establishment for the custody, loan, exchange, or issue of money; …
vt. to deposit or store in a bank; …
vi. to manage a bank; …

The rationale behind this organization is that word entries are first divided into broad groups, which are usually determined by etymological origins or evolutionary pathways. We can refer to such groups as "roots" since synchronically each group of senses of this type is associated with a semantic core, which meets the definition of root in morphology. Next, word senses in the same group (i.e., sharing the same root) are organized into syntactic categories (noun, verb, etc.).

Three key factors can be identified in the above dictionary-like lexical knowledge organization: (i) the abstract root, (ii) the syntactic category, and (iii) the concrete word sense. Their interaction can be expressed by the equation in (5).

(5) root (abstract) + syntactic category (abstract) = word sense (concrete)

That is, the relationship between word roots and word senses is mediated by syntactic categories. In fact, the above rationale is a standard part of current theoretical linguistics. It is a basic idea in the branch of generative grammar known as root syntax. My alternative classification of lexical ambiguity types to be presented below is built on this formal linguistic theory.

3.2. Lexical ambiguity via root syntax

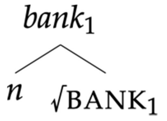
Root syntax has two influential incarnations: distributed morphology (DM; Halle and Marantz 1993, 1994; Marantz 1997) and exoskeletal syntax (XS; Borer 2005ab, 2013). A largely theory-neutral definition of the root is given in (6).

(6) The root is a purely lexical unit in formal linguistic representation that is void of categorial information. Its category is represented combinatorially.

This formal linguistic root is different from the traditional morphological root mentioned above, though the latter is the source of inspiration for the former. A crucial difference between the two is that the formal linguistic root is explicitly represented in syntactic analysis (hence the name root syntax), while the morphological root is not.

To illustrate how root syntax works, let us take the nominal use of *bank*$_1$ from (4). Its root-syntactic representation (in DM format) is given in (7). The representation can be given either as labeled bracketing or as a tiny tree diagram.

(7) *bank*$_1$ = [$_N$ *n* √BANK$_1$]



The root √BANK$_1$ in this representation has no category, so it has no concrete form or meaning either. The notation BANK$_1$ is just a mnemonic; the same root may be notated by a numerical index like 234. The *n* node is a categorizer—more exactly a nominalizer in this example. When the categorizer and the root are combined, we obtain a concrete word as a sound-meaning pair: ⟨/bæŋk/, 'a mound, pile…'⟩. Importantly, such sound-meaning pairs do not exist at the root level but only emerge or get retrieved at the post-categorization or "word" level.

The roots in root syntax are like Fellbaum's (2006) super-concepts. This status is most evident in Semitic languages. For instance, in Hebrew, the root √Š-M-N encodes a super-concept about some fatty substance, and the root √X-Š-B encodes a super-concept about some mental activity. Again, combining such roots with suitable categorizers will yield various concrete words, as in (8). These examples are taken from Arad (2005, 16).

(8)

a. √Š-M-N (about some fatty substance)		b. √X-Š-B (about some mental activity)	
šamen	adj. fat	*xašav*	v. think
šuman	n. fat	*xišev*	v. calculate
šaman	v. grow fat	*maxšava*	n. thought
hišmin	v. fatten	*maxšev*	n. computer

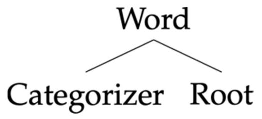
Arad (2005, 12) considers roots to be "underspecified lexical cores," which are "potentialities [that] may be incarnated in many different ways." The core idea of root syntax can thus be considered a generalization or axiomatization of the Semitic pattern, though the influence of the idea has gone far beyond Semitic languages (see, e.g., Alexiadou, Borer, and Schäfer 2014; Doron 2014). Root-level representation has been included in some NLU systems too. For instance, Black and El-Kateb (2004, 69) remark in a discussion on Arabic WordNet building that "the derivational root and form of each content word should be stored, since this way of semantically linking words is a basic expectation of a literate Arabic speaker." The pros and cons of generalizing this mode of representation to other (non-Semitic) languages await further testing.

While the above examples may give the impression that root syntax is just a restatement of POS-tagging, that is crucially not the case, because root-syntactic categorizers are not necessarily run-of-the-mill POS tags. On the one hand, the categorizer can encode more subtle grammatical information. Thus, the nominalizer can encode grammatical gender (Lowenstamm 2008) or number (De Belder 2013) information, and the verbalizer can encode eventuality type information (Cuervo 2003; Folli and Harley 2005). On the other hand, the categorizer is not necessarily a single syntactic category but may be a complex, multilayered structure. Take the causative verb *fatten*. Its root-syntactic representation is given in (9), where the root √FAT is jointly categorized by three stacked verbalizers: v_{CAUSE}, v_{GO}, and v_{BE}. Decompositional representations like this are common in root syntax.

(9) [v_{CAUSE} [v_{GO} [v_{BE} √FAT]]] (*fatten* = cause to become fat)

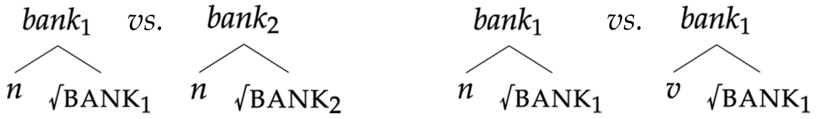
Overall, the categorizer-root combination can be considered an instantiation of the fundamental cognitive function of classification in the domain of language. There are two key components in the root categorization schema given in (10): the categorizer and the root. They together make room for two types of lexical ambiguity. We get homonymy when two identical word forms have different roots (regardless of the categorizer part) and get cross-POS polysemy when they have the same root but different categorizers.

(10) Root categorization schema



The two types of lexical ambiguity arising from the schema in (10) are illustrated in (11).

(11) a. homonymy b. cross-POS polysemy



Note that both ambiguity types in (11) reside at the categorization level, in that they both emerge in the process of root categorization. By contrast, the more subtle word sense polysemy within a single POS, such as the ambiguity between the two nominal senses of *bank*$_1$ in (4a) ('a mound, pile …' *vs.* 'the rising ground …'), only emerges after root categorization is done. We can call this higher level of lexical ambiguity the post-categorization level. Thus, we have the following two-level reclassification of the three lexical ambiguity types introduced in Section 1:

(12) a. Categorization level: homonymy/homography, cross-POS polysemy

b. Post-categorization level: intra-POS polysemy

This reclassification matches the meaning-first approach described in Section 3.1. We have given homonyms separate lexical entries (with each entry being defined by a separate root) and treated POS ambiguity as an integral part of polysemy (i.e., polysemy at the categorization level).

3.3. A further type of lexical ambiguity

Our discussion so far has focused on ambiguity in content words, which is what WSD and similar tasks in NLP are mostly (if not exclusively) about. However, content words are not the only type of lexical item[2] in human language, nor is ambiguity in content words the only kind of ambiguity in the lexicon. There is ambiguity in function or grammatical words too—and more generally, ambiguity across the content/function word boundary. This corresponds to a long-observed phenomenon in theoretical linguistics known as semilexicality (Corver and van Riemsdijk 2001) or semifunctionality (Song 2019), which is currently receiving increasing attention in the root syntax circle as well (e.g., Acedo-Matellán and Real-Puigdollers 2019; Cavirani-Pots 2020). While this type of ambiguity in the lexicon has so far been neglected in NLP/NLU, it has a natural place in the root-syntactic approach introduced in Section 3.2.

Semilexical elements are linguistic elements with both lexical content and grammatical function. Such elements abound in highly analytic languages like Chinese. See (13)–(14) for an illustration.

(13) Classifiers

a. *yī* *wèi* / *míng* / *gè* *lǎoshī* [Mandarin]

one CL$_{\text{respectful}}$ CL$_{\text{professional}}$ CL$_{\text{neutral}}$ teacher

'a teacher'

b. *yī* *zhī* / *tóu* *zhū*

one CL$_{\text{neutral}}$ CL$_{\text{pejorative}}$ pig

'a pig'

(14) Conjunctions

a. *hālì* *bōtè* *yǔ* / ?*hé* / ?*gēn* *mófǎ* *shí* [Mandarin]

Harry Potter and$_{\text{literary}}$ and$_{\text{neutral}}$ and$_{\text{colloquial}}$ magic stone

[2] The term "lexical" has two different senses in linguistics, which may cause certain confusion. It may refer either broadly to everything in the lexicon or narrowly to just content words (vis-à-vis function words). The "lexical" in "semilexicality" is used in the narrow sense. It is not exactly clear which sense the "lexical" in "lexical ambiguity" refers to in NLP, but in this article, I use "lexical ambiguity" in the broad sense, referring to ambiguity in all items stored in the lexicon. Similarly, I use "lexical item" to broadly refer to any item in the lexicon.

'Harry Potter and the Philosopher's Stone'

b. *xiǎomíng ?yǔ / hé / gēn xiǎohóng dōu zài dǎpái*
Xiaoming and$_{\text{literary}}$ and$_{\text{neutral}}$ and$_{\text{colloquial}}$ Xiaohong both PROG play poker
'Both Xiaoming and Xiaohong are playing poker.'

The classifiers in (13) all have the same grammatical function—they all serve to individuate the nominal concept and thereby prepare the ground for counting. Yet, different nouns lexically select different classifiers on the one hand, and the same noun pragmatically selects different classifiers depending on the speaker's tone or attitude on the other hand. Neither dimension of selection is grammatically based; both are determined by the lexical idiosyncrasies of each classifier instead. Similarly, the conjunctions in (14) all serve to link nouns as far as the grammar is concerned, and their alternation is determined by register (e.g., how elevated the discourse is). In short, the lexical content in classifiers and conjunctions in Chinese determines important aspects of their usage.

Semilexical elements exist in synthetic languages too, though to a lesser extent. For instance, the alternative auxiliary verbs in Italian, Dutch, and Afrikaans in (15)–(16) have all been identified as partly lexical and partly grammatical in the literature.

(15) Alternative voice auxiliaries
La pasta va / viene mangiata subito. [Italian]
the pasta PASS$_{\text{obligatory}}$ PASS$_{\text{regular}}$ eaten immediately
'Pasta must be / is eaten immediately.' (Cardinaletti and Giusti 2001, 392)

(16) Alternative aspect auxiliaries
a. *Ik heb de hele dag zitten te lezen.* [Dutch]
I have the entire day sit$_{\text{PROG}}$ to read
'I have been reading the entire day.' (Cavirani-Pots 2020, 1)
b. *Ek het gister baie (ge-)loop (en) praat.* [Afrikaans]
I have yesterday a lot walk$_{\text{PROG}}$ and talk
'I have been (walking and) talking a lot yesterday.' (Cavirani-Pots 2020, 334)

In (15), the Italian motion verbs *va* 'goes' and *viene* 'comes' are used as passive auxiliaries in place of the default *è* 'is', and in this situation they do not denote motions at all. Interestingly, the two motion verbs have slightly different connotations when used as auxiliaries—*va* gives the sentence a more obligatory-sounding tone, whereas *viene* sounds more regular and objective. In (16), Dutch *zitten* 'sit' and Afrikaans *loop* 'walk' are again motion verbs used as auxiliaries; they both denote the progressive aspect, and the sentences do not have to involve any sitting or walking at all.

Somewhat surprisingly, semilexical elements also abound in polysynthetic languages, though in a different guise. In such languages, they usually appear as affixes and are known as "lexical/field affixes" in the literature (see Song 2021a for an overview). For instance, classifiers are attested in polysynthetic languages like Halkomelem and Yurok as well, as in (17)–(18).

(17) *łíxʷ-əqən* *lisék* [Halkomelem]
three-CL$_{\text{container}}$ sack
'three sacks' (Gerdts and Hinkson 1996, 10)

(18) a. *dikwh-okwł* *bołak* [Yurok]
three-CL$_{\text{salmon}}$ salmon
'three salmon'
b. *nahks-oh* *ha'aag*
three-CL$_{\text{round}}$ rock
'three rocks' (Conathan 2004, 26–27)

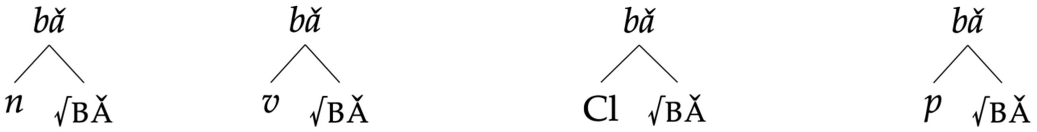
Except for their affixal status, classifiers in Halkomelem and Yurok are used in the same way as classifiers in Chinese, also in the Numeral-Classifier-Noun pattern.

Overall, semilexicality is a widespread phenomenon in human language. In root syntax—more exactly in an extension of it called generalized root syntax (Song 2019)—semilexical words are formally represented in the same way as content words, also with a categorizer and a root. It is just that now the categorizer is a functional category, not a lexical one; and the root in question has developed grammatical use for independent reasons. From this angle, semilexicality can be considered a special type of cross-POS polysemy. Take the Mandarin word form *bǎ*, for example. It is regularly used in four senses, each in a separate syntactic category, as illustrated in (19).

(19) a. *zhè liàng zìxíngchē, bǎ bù tài hǎo-shǐ* [Mandarin]
this CL$_{\text{vehicle}}$ bike handle not too good-use
'As for this bike, its handles are not quite easy to use.'
b. *nǐ-de zhízé shì bǎ mén bǎ hǎo*
you-POSS duty is DISP gate guard good
'Your duty is to guard the gate well.'
c. *yī bǎ sǎn*
one CL$_{\text{object-with-handle}}$ umbrella
'an umbrella'

The *bǎ* in (19a) is a noun meaning 'handle'. There are two occurrences of *bǎ* in (19b): the first one is a disposal preposition introducing a directly affected object (somewhat similar to a direct object marker), and the second one is a verb meaning 'to guard'. Finally, the *bǎ* in (19c) is a classifier. As we can see, the different lexical and functional senses of the same word form may well co-occur. Thus, to obtain a more complete picture of polysemy, we need to take the issue of semilexicality into account. The root-syntactic representation of the polysemy of *bǎ* is given in (20).[3]

(20) n. handle v. to guard, to hold cl. for holdable objects p. for direct objects



In the case of semilexical words, the categorizer specifies the grammatical function, while the root contributes lexical coloration of various sorts, such as encyclopedic content, speaker attitude, and register conditioning.

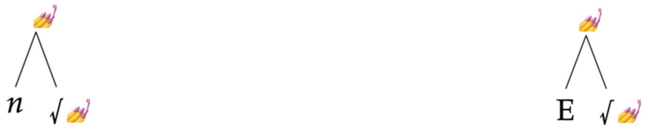
As we can see, the root-syntactic representation of lexical ambiguity is systematic and consistent for both content and semilexical words. The usefulness of root syntax does not stop here. As shown in Song (2022a), the root categorization schema may also be applied to emojis in computer-mediated communication (CMC). Specifically, the same emoji can be used either to replace a content word or to convey certain speaker emotion or attitude. These are respectively called the nonaffective and the affective use of emojis. See (21) for an illustration.

(21) a. Horses need 💅 too.

b. Sorry to say, but it's a fact 💅 (X, formerly known as Twitter)

In (21a), the emoji 💅 simply replaces the content word *manicure* (the tweet is accompanied by a video of horseshoe replacement). In (21b), the same emoji conveys a humorously nonchalant tone. The pictorial "root" of the emoji stays the same in the two different uses; what has changed is its syntactic category. Thus, the occurrence of the emoji in (21a) can be treated as a content word, while the occurrence in (21b) can be treated as a semilexical word. Such ambiguity in emoji usage is common in CMC, and it can be formally represented as in (22), where E is an emotional category.

(22) a. 💅 (n. manicure) b. 💅 (e. humorously nonchalant tone)



Thus, the root-syntactic approach to lexical ambiguity proposed here is not only more in line with human understanding but also broader in empirical coverage. It offers a unified representation for categorization-level ambiguity in content words, semilexical words, and CMC elements such as emojis. The latter two types of elements have not yet received attention in NLP/NLU research on lexical ambiguity.

[3] There is some debate among linguists as to whether *bǎ* in its disposal use is a true preposition or just some kind of light verb category. While I treat this *bǎ* as prepositional for expository convenience, nothing in my discussion hinges on this particular choice of category.

Our discussion so far has focused on the categorization level—the first of the two broad levels of lexical ambiguity defined in (12). As for lexical ambiguity at the second, post-categorizaton level, such as the subtly different nominal senses of *bank* in its financial use (e.g., the institution *vs.* the building), root syntax does not have much to say, as all categorized senses associated with a root are stored in an area of the mental lexicon called the encyclopedia, the internal structure of which is left unstudied. In the next section, I will adopt a different tool to formally represent post-categorization lexical ambiguity.

4. Lexical ambiguity representation via category theory

Recall from Section 1 that the goal of this article is to demonstrate, via the case of lexical ambiguity, that insights from theoretical linguistics are still useful in NLU development, in that they can help us achieve more human-like meaning representation. However, while the theoretical linguistic approach described in Section 3 is conceptually appealing, we are faced with two obstacles if we want to apply it to NLU system building. First, as already mentioned, the root-syntactic method is much more useful at the categorization level than at the post-categorization level. Therefore, intra-POS word sense polysemy is as much a challenge in our approach as it is in the approaches described in Sections 1–2. Second, most theoretical linguistic tools and ideas, including root syntax, have never been introduced into NLP/NLU research, so we must also consider the ever-larger gap between the two fields and find a suitable bridge. Mathematical category theory might be such a bridge. In what follows, I will first briefly introduce the application of category theory in NLP (Section 4.1), then present a category-theoretic method (via the tool of topos) to represent intra-POS polysemy (Section 4.2), and finally integrate the root-syntactic insights into this new mode of representation (Section 4.3).

4.1. Category theory in NLP

Category theory is a branch of mathematics that is dedicated to the representation and reasoning of abstract structures. In Fong and Spivak's (2019, Preface) words, it is "unmatched in its ability to organize and layer abstractions" and has the potential of "building rigorous bridges between disparate worlds, both theoretical and practical." Indeed, though it was originally invented to bridge algebra and topology (Eilenberg and Mac Lane 1945), category theory has now found applications in a range of disciplines, including physics, logic, computer science, linguistics, and philosophy.[4] In NLP, the most representative application is the approach known as categorical compositional distributional semantics or DisCoCat (Coecke, Sadrzadeh, and Clark 2010 et seq.), which is an integration of distributional and compositional semantics. Due to space limitations, here I cannot properly introduce the basics of category theory. Interested readers can consult any textbook on the subject (e.g., Mac Lane 1998; Awodey 2010; Fong and Spivak 2019). In very general terms, a category is made up of a collection of *objects* and *morphisms* between the objects, where each object has an identity morphism and morphisms can be composed (obeying certain conditions). There can be "morphisms" across categories too, which preserve structures and are called *functors*. See (23) for an illustration.

[4] For an overview of the cross-disciplinary applications of category theory, see Song (2019, 196).

(23) A functor between two categories

$$\mathcal{C} \xrightarrow{F} \mathcal{D}$$

$$\begin{tikzcd}
A \arrow[r, "f"] \arrow[dr, "h"'] \arrow[loop left, "id_A"] & B \arrow[d, "g"] \arrow[loop right, "id_B"] & & F(A) \arrow[r, "F(f)"] \arrow[dr, "F(h)"'] \arrow[loop left, "id_{F(A)}"] & F(B) \arrow[d, "F(g)"] \arrow[loop right, "id_{F(B)}"] \\
& C \arrow[loop right, "id_C"] & & & F(C) \arrow[loop right, "id_{F(C)}"]
\end{tikzcd}$$

In (23), $\mathcal{C}$ and $\mathcal{D}$ are two categories, and F is a functor between them. The two triangles live in the two categories and are connected via the functor: the $\mathcal{C}$-objects $A, B, C,$ the $\mathcal{C}$-morphisms $f, g, h,$ as well as the identity morphisms are all preserved in $\mathcal{D}$. These category-theoretic concepts are fully general; they get different concrete construals in different domains and applications.

The key idea of DisCoCat is to treat semantic interpretation as a functor from syntax to semantics, both of which are formalized as categories (see Coecke et al. 2013 for an overview). The syntax category $\mathcal{C}$ is based on Lambek's (1999) pregroup system, and the semantics category $\mathcal{FV}ect$ is based on finite-dimensional vector spaces. The functorial passage between the two is given in (24).

(24) $\mathcal{C} \overset{I}{\to} \mathcal{FV}ect$

The syntax category is made up of grammatical types (e.g., N for proper names, S for sentences), which are mapped to vector spaces via the interpretation functor. Since the functor preserves the structure of its source category in its target category, word vectors in the model can be composed into phrase or sentence vectors in a syntax-respecting way.

Lexical ambiguity has received relatively little attention in the DisCoCat framework, and the few studies dedicated to the issue are all inspired by quantum physics (Kartsaklis 2014; Piedeleu 2014; Piedeleu et al. 2015; see also Wang 2024). Leaving aside technical details, the main idea in these studies[5] is to redefine the interpretation functor and let it map ambiguous—more exactly homonymous—words to *mixed states*, which are statistical ensembles of various potential meanings. See (25) for an illustration.

(25) $\rho_{bank} = \frac{1}{2}|bank\rangle_f\langle bank|_f + \frac{1}{2}|bank\rangle_r\langle bank|_r = \begin{pmatrix} 0.48 & 0.17 \\ 0.17 & 0.52 \end{pmatrix}$ (Kartsaklis 2014, 113)

This is the mixed-state denotation of the homonymous word *bank* in a toy model. It basically says that *bank* could either mean a financial bank or a river bank, with equal probability. Crucially, a mixed state is not a vector but a specially defined operator or matrix (called a density matrix) on the vector space that the "pure states" (i.e., vectors) making up the mixed state inhabit. This reflects a key difference between DisCoCat and classical distributional models of word meanings. In the latter, all words are modeled as vectors, while in DisCoCat

[5] More exactly in all the other studies cited here except Wang (2024), which has a slight different category-theoretic foundation using (pre)sheaves. Thanks to Nicolás José Fernández-Martínez for this reference.

some words are modeled as matrices or even-higher-order tensors.

DisCoCat is first and foremost a theory of type-logical composition; it is just that its compositional ingredients, the individual word meanings, are distributionally defined. However, that is more of a matter of model choice; the semantics category in (24) may well be replaced by some non-vector-based category, such as the category of sets and relations (Piedeleu et al. 2015). The role of the lexicon in DisCoCat, as in traditional formal semantics, is just a storehouse of words serving the more central purpose of composition, and its internal structuring, including most of the lexical relations, is beyond the scope of the theory. Indeed, very little (if any) ontological information is encoded in the syntax category in (24). In short, DisCoCat and its various descendents are designed to categorify syntax and semantics but not the lexicon.

With the above clarification in place, we can now see the systematic limitations of DisCoCat in its treatment of lexical ambiguity. First, its classification of ambiguity into homonymy and polysemy is quite simplistic.[6] Thus, the classification problems mentioned in Section 1 are also problems in DisCoCat. To attain human-like NLU, we must first pin down the levels and types of ambiguity in a more fine-grained manner, as we have attempted to do with our root-syntactic classification in Section 3. Second, while homonyms are given a quantum-based modeling in DisCoCat, polysemy—especially intra-POS, regular polysemy—is still a challenge, as the subtly different senses of a polysemous word are still all encoded in a single vector or pure state. This is a nonnegligible challenge if what we want is a principled representation of humans' lexical knowledge. In fact, the challenge has been taken up in a separate line of work in categorical linguistics, which I will discuss below.

4.2. Representing intra-POS polysemy in a topos

While DisCoCat is the most representative application of category theory to linguistics, it is not the only application. Among others, there is a line of work by Asher (2011) and Babonnaud (2019, 2021, 2022) using the category-theoretic tool of topos to represent lexical knowledge, including polysemy. This alternative branch of categorical linguistics is both directly relevant to our specific case study and well aligned with our higher goal of developing human-like NLU systems. In Babonnaud's (2021, 19) words, "toposes could be the best categorical models to interpret on a unified basis a large variety of semantic frameworks with subtyping."

A topos is a special category that, in addition to the basic categorical settings, has a range of extra features that together create a quasi-set-theoretic environment. Indeed, a motivating example of toposes is just the category $\mathcal{Set}$ of sets and functions, though toposes are much more general than sets. See (26) for a more formal definition based on Bell (1988, 60).

(26) A topos is a category with

(i) *binary products*, such that for each pair of objects A, B there is an object $A \times B$;

(ii) a *terminal object* 1, to which there is a unique morphism ! from every object;

[6] While Piedeleu et al. (2015) declare that the mixed-state modeling only applies to homonyms, they in practice apply it to irregular polysemes too, such as the 'female monarch' and the 'piece on a chessboard' sense of *queen*.

(iii) a *subobject classifier* Ω, which is tied to a subobject classification configuration to be illustrated below; and

(iv) *power objects*, such that for each object A there is an object $\mathcal{P}(A)$.

In the topos of sets, binary products are cartesian products, 1 is any singleton set, Ω is the truth value set {true, false} (assuming a two-valued system), and power objects are power sets.

The products and power objects have certain canonical morphisms, which we will see below. For now, let us first focus on the subobject classifier. In (27), there are two "pullback" squares living in two toposes: a general one (27a) and one that is specialized for lexical semantics (27b).

(27) a. general subobject classification b. lexical semantic subtyping

$$\begin{tikzcd} A \arrow[r, tail, "u"] \arrow[d, "!"'] \arrow[dr, phantom, "\lrcorner", very near start] & B \arrow[d, "\chi_A"] \\ 1 \arrow[r, "\top"] & \Omega \end{tikzcd}
\qquad
\begin{tikzcd} \mathrm{Bank}_1 \arrow[r, tail, "u"] \arrow[d, "!"'] \arrow[dr, phantom, "\lrcorner", very near start] & E \arrow[d, "\chi_{\mathrm{Bank}_1}"] \\ 1 \arrow[r, "\top"] & \Omega \end{tikzcd}$$

Both pullbacks in (27) are in the subobject classification configuration, where the object in the top left corner is classified as a subobject of the object in the top right corner. What exactly a subobject means depends on what the objects in the topos are. In (27a), A is just a general subobject of B. By comparison, in the topos of lexical types in (27b), Bank$_1$ is a *subtype* of the type E for entities. For current purposes, we can abstract away from the remaining technical details in the pullbacks (see Goldblatt 1984 for a general introduction to topos theory).

Beyond the basic subtyping square, the topos tool is used in Asher (2011) in a more sophisticated way. One of Asher's main concerns is the "dot type," which he assigns to a word with inherent polysemy—a special case of intra-POS polysemy. Recall from Section 1 that this is a subtype of regular polysemy where the different senses of a word are equally basic. For instance, he assigns the noun *book* the type P·I, indicating the fact that its meaning has both a physical (P) and an informational (I) aspect. The idea of dot types can be traced back to Pustejovsky's (1995) work on his "generative lexicon." As Pustejovsky (1998, 335) explains, for each sense pair, "there is a relation which 'connects' the senses in a well-defined way"—which "must be seen as part of the definition of the semantics for the dot object […] to be well-formed."

Asher's (2011) categorical representation of dot types is rather complicated, making use of power objects in the topos environment and spanning two ontological categories. As Babonnaud (2019, 2021, 2022) points out, such complication is unnecessary, because dot types, as relations between aspects, can be treated as subtypes of product types. I adopt Babonnaud's simpler representation here, based on which we can put the dot type P·I in the following subtyping square:

(28) A dot type is a subtype of a product type (à la Babonnaud)

$$\begin{tikzcd} \mathrm{P}\cdot\mathrm{I} \arrow[r, tail, "u"] \arrow[d, "!"'] \arrow[dr, phantom, "\lrcorner", very near start] & \mathrm{P}\times\mathrm{I} \arrow[d, "\chi_{\mathrm{P\cdot I}}"] \\ 1 \arrow[r, "\top"] & \Omega \end{tikzcd}$$

Babonnaud goes one step further and writes the composite type P·I as BOOK, thereby removing the dot notation altogether. Note that while Banonnaud, like Asher, is mainly concerned with inherent polysemy, the same idea (especially under his new gestalt notation) can be applied to all cases of intra-POS polysemy, because the nature of the inter-sense relation (i.e., whether it is irregular or regular) is independent of the formal subtyping relation in the pullback square. Thus, we now have a method to systematically represent intra-POS polysemy. Recall that this was not possible in either the root-syntactic or the DisCoCat approach to lexical ambiguity.

Moreover, topos theory makes it possible to retrieve the individual aspects from a composite type and thereby to disambiguate polysemous words. The disambiguation process is different for Asher and Babonnaud due to a fundamental disagreement between them: Asher objects to the subtyping relation between P·I and $P \times I$, whereas Babonnaud supports it. Given this difference, disambiguation is easier for Babonnaud than for Asher.

To see how disambiguation works in the topos-based representation of intra-POS polysemy, we need to first recall some canonical morphisms available in the topos by definition, as in (29).

(29) a. canonical maps of products b. canonical maps of power objects

$$\begin{tikzcd} \mathrm{P} & \mathrm{P} \times \mathrm{I} \arrow[l, "\pi_1"'] \arrow[r, "\pi_2"] & \mathrm{I} \end{tikzcd}$$

$$\begin{tikzcd} \mathrm{P} \times \mathrm{I} \arrow[d, "1_{\mathrm{P}} \times \hat{f}"'] \arrow[dr, "f"] & \\ \mathrm{P} \times \mathcal{P}(\mathrm{P}) \arrow[r, "ev_{\mathrm{P}}"] & \Omega \end{tikzcd} \qquad \begin{tikzcd} \mathrm{I} \times \mathrm{P} \arrow[d, "1_{\mathrm{I}} \times \hat{g}"'] \arrow[dr, "g"] & \\ \mathrm{I} \times \mathcal{P}(\mathrm{I}) \arrow[r, "ev_{\mathrm{I}}"] & \Omega \end{tikzcd}$$

In (29a), π_1 and π_2 are the projection maps standardly defined for each product object. In (29b), f and g represent two perspectives to view the dot relation between P and I underlying the concept BOOK. The two perspectives uniquely determine two morphisms $\hat{f}: I \to \mathcal{P}(P)$ and $\hat{g}: P \to \mathcal{P}(I)$. To disambiguate P·I in Babonnaud's system, we just need to compose π_1 and π_2 with u, as in (30a). To do so in Asher's system, on the other hand, we need two additional maps from P·I to $P \times \mathcal{P}(I)$ and $\mathcal{P}(P) \times I$, as in (30b). These two "aspect" maps are specially defined in Asher (2011).

(30) a. $\mathrm{P} \cdot \mathrm{I} \xrightarrow{u} \mathrm{P} \times \mathrm{I} \xrightarrow{\pi_1} \mathrm{P},\ \mathrm{P} \cdot \mathrm{I} \xrightarrow{u} \mathrm{P} \times \mathrm{I} \xrightarrow{\pi_2} \mathrm{I}$ (à la Babonnaud)

b. $\mathrm{P} \cdot \mathrm{I} \xrightarrow{asp_{\mathrm{P}}} \mathrm{P} \times \mathcal{P}(\mathrm{I}) \xrightarrow{\pi_1} \mathrm{P},\ \mathrm{P} \cdot \mathrm{I} \xrightarrow{asp_{\mathrm{I}}} \mathcal{P}(\mathrm{P}) \times \mathrm{I} \xrightarrow{\pi_2} \mathrm{I}$ (à la Asher)

The ability to systematically disambiguate polysemous words, albeit only on the symbolic level, is a big advantage of the topos-based representation of intra-POS polysemy. Disambiguation is, to some extent, also possible in the DisCoCat representation for homonyms. However, there the disambiguation is distributionally done and hence in a continuous manner, as illustrated in (31).

(31) a. $\mathrm{Tr}\big(\rho_{\text{river bank}} \circ (|\text{fish}\rangle\langle\text{fish}|)\big) = 0.43$

b. $\mathrm{Tr}\big(\rho_{\text{river bank}} \circ (|\text{money}\rangle\langle\text{money}|)\big) = 0.06$ (based on Kartsaklis 2014, 113–114)

The two numbers in (31) measure the similarity between *river bank* and *fish*/*money*, respectively, so the word form *bank* is disambiguated, by being part of the phrase *river bank*, in the sense that its meaning is much more similar to the meaning of *fish* than to that of *money*. Note that this way of disambiguation is meant for homonymous but not polysemous words; recall from Section 4.1 that the latter cannot be assigned special "uncertain" denotations in the quantum-based modeling. By contrast, disambiguation in the topos environment is not restricted by lexical ambiguity types. The different senses of a polysemous word can be disambiguated (in a discrete manner) however subtle or close their relation is. Moreover, the various sense types in the topos ontology are neatly structured based on granularity levels (by means of the subtyping morphisms), unlike in the contextual embeddings of popular neural language models mentioned in Section 2, where all the extremely nuanced token-based sense vectors live on the same level in the output semantic space.

The topos-based representation of intra-POS polysemy is appealing, but we face a new problem now: our formal representation of lexical ambiguity is no longer methodologically unified. Part of it is done via root syntax (the categorization-level part), and part of it is done via category theory (the post-categorization part). A desirable goal is to re-unify the two levels. I will present such a reunification in the next section.

4.3. A unified categorical representation of lexical ambiguity

To re-unify the two levels of lexical ambiguity in our formal representation, we must translate our root-syntactic conception of categorization-level ambiguity into categorical terms. To see how, let us take a closer look at (27b), where the lexical type Bank_1 corresponds to the word *bank*$_1$, which in root syntax can be represented as in (32).

(32) $[_{\text{N}}\ n\ \sqrt{\text{BANK}_1}\ ]$

Thus, if we combine the root-syntactic and the category-theoretic perspective, we may take Bank_1 to be an abbreviation for the product type in (33).

(33) $\text{Bank}_1 \triangleq \text{Type}(n) \times \text{Type}(\sqrt{\text{BANK}_1}) \triangleq E \times R$

That is, we may define Bank_1 as the product of the type of the nominalizer, which by assumption is just the entity type E, and the type of roots, for which we define a new type R. However, there is a problem in this conception. If we simply equate Bank_1 with $E \times R$, we end up equating all nominal types with one another. The crux of the problem is that the distinction between roots vanishes at the type level (as is expected). Thus, to retrieve the cross-concept distinction, we need to turn to the term level. In category theory, especially in categorical logic (Crole 1993), the type *vs.* term distinction is standardly cast as an object *vs.* morphism distinction: types are objects, and terms are morphisms. In our case, we may retrieve "root terms" as morphisms from 1 to R, which are instances of *global elements* in category-theoretic terminology. Accordingly, we may retrieve the categorizer-root combination in (32) as a pair of morphisms at the term level, as in (34).

(34) $E \times 1 \xrightarrow{\langle id_E, \sqrt{\text{BANK}_1}\rangle} E \times R$

Thus, at the term level, we can treat *bank*$_1$ as the name of the morphism $\langle id_E, \sqrt{bank_1}\rangle$. That said, Bank_1 in (33) is still a type and hence an object in the topos—it is a common situation in category theory that essentially the same thing may be cast in both object and morphism terms. To preserve this insight, we can modify (33) as (35), where instead of equating Bank_1 with $E \times R$, we treat the former as a particular subtype of the latter—the subtype that is uniquely determined by *bank*$_1$.

(35) $\text{Bank}_1 \overset{u_{bank_1}}{\rightarrowtail} E \times R$

In this way, we have translated the root-syntactic representation in (32) into the topos language in full, not only giving each element in (32) a categorical counterpart, but also preserving the idea that the relation between a categorizer-root combination like [*n* √BANK1] and a word like *bank*1 is not strict identity but just one-one correspondence. In theoretical linguistic terms, this is the correspondence between a syntactic structure and its interface interpretations. A caveat here is that the subtyping relation in (35) is defined in a different way from that in (28). The supertype $E \times R$ in (35) is not based on two related word senses but directly represents the syntactic relation between a categorizer and a root. Thus, a polysemous word like *book* can be given two subtyping morphisms in the topos, one based on its semantic properties and the other based on its syntactic properties. These two morphisms are not equivalent, but they can perfectly coexist in the topos since all ingredients involved here are ontological types. I will return to this point shortly below.

In addition, we can combine the two subtyping relations in (27b) and (35) in the topos, as in (36), which expresses that the "root-tagged" entity type $E \times R$ is a subtype of the pure entity type E, and that if we concretize the root-tagging via a specific root (e.g., √BANK1), then that concrete root-tagged entity type Bank1 is also ultimately a subtype of the pure entity type E.

(36) $\mathrm{Bank}_1 \overset{u_{bank_1}}{\rightarrowtail} E \times R \overset{u}{\rightarrowtail} E$

The same translation can be extended to semilexical words, for which we simply need to replace E with some grammatical type. For example, to classify the Mandarin semilexical conjunction *gēn* 'and (colloquial)', we replace E with *Conj*, as in (37).

(37) A semilexical conjunction is a subtype of conjunction

$$\begin{tikzcd}
Conj \times R \arrow[r, tail, "u"] \arrow[d, "!"'] \arrow[dr, phantom, "\lrcorner", very near start] & Conj \arrow[d, "\chi_{Conj_R}"] \\
1 \arrow[r, "\top"] & \Omega
\end{tikzcd}$$

Then, we can similarly retrieve the root-syntactic insights about the semilexical word at both type and term levels, as in (38).

(38) a. $\mathrm{G\bar{e}n} \overset{u_{g\bar{e}n}}{\rightarrowtail} Conj \times R \overset{u}{\rightarrowtail} Conj$ (type level)

b. $g\bar{e}n \triangleq Conj \times 1 \xrightarrow{\langle id_{Conj}, \sqrt{\text{GĒN}} \rangle} Conj \times R$ (term level)

That is, the term *gēn* names the morphism $\langle id_{Conj}, \sqrt{\text{GĒN}_1} \rangle$, which in turn determines the type Gēn.

With the above translation in place, we can now represent categorization-level ambiguity in our topos in a root-syntax-conforming way. Take the Mandarin word form *bǎ* from (20), for example. The root-syntactic and category-theoretic representations for its four uses are given in Table 1.[7]

[7] Here I assume that the ontological type of the verbalizer is the eventuality type (*Ev*). I remain agnostic about the type of the prepositional categorizer and use *P* as a placeholder.

Meaning	Root-syntactic representation	Category-theoretic representation
n. handle	$[_{\mathrm{N}}\ n\ \surd\text{BĂ}\ ]$	Type: Bă $\overset{u_{bă}}{\rightarrowtail} E \times R \overset{u}{\rightarrowtail} E$ Term: $E \times 1 \xrightarrow{\langle id_E, \surd\text{BĂ}\rangle} E \times R$
v. to guard, to hold	$[_{\mathrm{V}}\ v\ \surd\text{BĂ}\ ]$	Type: Bă $\overset{u_{bă}}{\rightarrowtail} Ev \times R \overset{u}{\rightarrowtail} Ev$ Term: $Ev \times 1 \xrightarrow{\langle id_{Ev}, \surd\text{BĂ}\rangle} Ev \times R$
cl. for holdable objects	$[_{\mathrm{Cl}}\ \mathrm{Cl}\ \surd\text{BĂ}\ ]$	Type: Bă $\overset{u_{bă}}{\rightarrowtail} Cl \times R \overset{u}{\rightarrowtail} Cl$ Term: $Cl \times 1 \xrightarrow{\langle id_{Cl}, \surd\text{BĂ}\rangle} Cl \times R$
p. for direct objects	$[_{\mathrm{P}}\ p\ \surd\text{BĂ}\ ]$	Type: Bă $\overset{u_{bă}}{\rightarrowtail} P \times R \overset{u}{\rightarrowtail} P$ Term: $P \times 1 \xrightarrow{\langle id_P, \surd\text{BĂ}\rangle} P \times R$

TABLE 1. CATEGORY-THEORETIC REPRESENTATION OF CATEGORIZATION-LEVEL AMBIGUITY

Before we end this section, a clarification is in order regarding the type system used here. All the types in our discussion are *ontological* types, which are crucially different from the more familiar types in type-logical syntax or formal semantics. In these latter areas, a conjunction does not have the name-tag-like type *Conj* but is assigned a type like $t \to t \to t$ (among other possibilities). This type assignment is combinatorial in nature. Ontological and combinatorial types are two different species (see Song 2019, 24). Ontological types serve taxonomic purposes, so they are amenable to feature-based definition, which is how categories are standardly defined in linguistics. Thus, Chomsky (1970) defines four lexical categories N, V, A, P by two bivalent features [±N] and [±V]. The difference between ontological and combinatorial types is similar to the difference between generative grammar and categorial grammar in their inventories of syntactic categories, as in (39).

(39) a. N, V, A, T, Asp, Voice, C, Foc, Top, Num, D, … (generative grammar)

b. N, S, N\S, N\S/N, N\N, S\S/S, … (categorial grammar)

The categories in (39a) are taxonomically defined (by features) and more suitable for ontology building (see Song 2019 for such an ontology built with category-theoretic tools), while the categories in (39b) are combinatorially defined and more suitable for type-logical derivation. Root syntax has been given a categorical semantics in Song (2021b) too, where the categorizer-root combination is not assigned an ontological type like $E \times R$ but assigned a combinatorial type Writer E (via the tool of monad).[8] Category theory provides tools to formally represent both ontological and combinatorial types, but these should not be confused with each other.

5. Toward type-enriched word sense disambiguation

As mentioned in Sections 1–2, currently popular neural language models have achieved amazing performance, and from an engineering perspective, there is probably no need for

[8] Compositional semantic issues are beyond the scope of this article, but the categorical semantics in Song (2021b) covers both content words and semilexical words. It has been applied to emojis in Song (2022b) too.

dedicated WSD systems anymore, as token-based word embeddings are disambiguated by default. However, such disambiguation, being unstructured and undifferentiated, is far from human-like. My main aim in this article has been to explore how this situation can be improved. My position is that symbolic approaches to language can still be useful, despite the extremely minor role theoretical linguistics has played in NLP in the past decades. Indeed, the formal representations of meaning developed in Sections 3–4 can reflect both the structure and the nuances of lexical ambiguity in a coherent, rigorous way. The question, then, is how to channel them to NLU systems. While a computational implementation is beyond my scope, below I first make a preliminary proposal.

Since the topos in Section 4 is essentially a type ontology, it may be added to WSD systems as an external knowledge base, perhaps alongside other external sources like WordNet. The feasibility of this direction relies on two conditions: the feasibility of hybrid models and the interoperability of multiple knowledge bases. The former has been justified in current WSD research (see Section 2), and the latter has also received category-theoretic solutions. If we assume an ambient category of ontologies (objects) and inter-ontology connections (morphisms), then ontology merging or alignment can be formally achieved via universal categorical constructions like limits and colimits (Hitzler et al. 2005; Healy and Caudell 2006; Zimmerman et al. 2006; Antunes, Rademaker, and Abel 2019). From a different angle, we may also view a single ontology as a category (like a database schema) and connect it to other categories (either of data or of other schemata) via functors (Johnson and Rosebrugh 2010; Spivak and Kent 2012; Spivak 2014). There are also more advanced topos-theoretic treatments of ontologies beyond the subtyping pullbacks (e.g., Reformat, D'Aniello, and Gaeta 2018). In short, various ideas from existing literature are methodologically compatible with our type ontology topos in Section 4, and its formal relation with familiar knowledge sources like WordNet is an interesting issue to explore in future research.

A general procedure adopted in existing hybrid WSD models (e.g., Huang et al. 2019; Bevilacqua and Navigli 2020; Scarlini, Pasini, and Navigli 2020) as well as in general-purpose knowledge-enhanced pre-trained language models (e.g., Sun et al. 2020; Sun et al. 2021; Wang et al. 2021) is to combine word/text embeddings and knowledge embeddings in a suitable manner, usually just by concatenation. We can follow this practice and find a way to integrate the rich type information in our topos into knowledge embeddings. However, this may not be the best way to use the theoretical-linguistically informed, mathematically rigorous meaning representations developed here. Their main purpose is not to help WSD models better *perform* the task (the performance of existing models is already highly impressive) but to equip current models' output with more structure and type-ontological information, thus making it more *human-like*. Such information need not be converted to vector format but can also just be linked to the sense inventory (e.g., WordNet) by one of the ontology-merging methods mentioned above. To that end, we may leverage the graph structure of WordNet, because graphs and categories are closely related mathematically.

Given a sentence containing an ambiguous word, we can obtain the contextual embedding of the target word in the usual way via BERT (Devlin et al. 2018) or some variant of it. In a hybrid WSD model like GlossBERT (Huang et al. 2019) or EWISER (Bevilacqua and Navigli 2020), by the end of the disambiguation step, the word should also already be linked to a sense entry in WordNet. Now, with the support of our type ontology, the sense entry is furthermore linked to an object in the topos, which in turn is linked to a whole web of type-level information. If the target word is *bank* and its disambiguated sense is *bank*$_1$ (which, in WordNet, is a synset), that would be linked to Bank$_1$ in the topos (see 36), which in turn has morphisms to $E \times R$, E, and so on. If the target word is *book*, on the other hand, and if its disambiguated sense is the physical entity, its WordNet synset would be linked to P in the topos (see 30), which has morphisms to/from P·I, $P \times I$, $E \times R$, and so on. The different shapes of the local morphism webs associated with the two objects (Bank$_1$ and P) clearly indicate, among others, that *bank* is a homonym and *book* is a polyseme, for only polysemes have dot types.

In sum, the type ontology topos supplements WordNet with significant background information. While such meta-information is a natural part of humans' lexical knowledge, it is absent from even the most fine-grained vector embeddings. Of course, humans' lexical knowledge involves not only ontological but also derivational type information (see Section 4.3). The same ontology-merging strategy can be used to connect that to WordNet too. A candidate categorical ontology for derivational types is the syntax category of DisCoCat (see Section 4.1).

6. Conclusion

In this article, I have used lexical ambiguity as a case to demonstrate how insights from theoretical linguistics can be useful for building human-like NLU systems. Specifically, NLU systems require a knowledge base that is equipped with formal representations reflecting the structure of humans' cognitive system. Theoretical linguistics has many tools and insights that can facilitate this task. The specific tool I have adopted is root syntax from generative grammar, which allows us to divide the several types of lexical ambiguity into two levels: the categorization level (homonymy, cross-POS polysemy) and the post-categorization level (intra-POS polysemy). This two-level classification is more in line with human understanding compared to the task-oriented division of labor popular in NLP/NLU practice.

Despite its strong expressive power at the categorization level, root syntax is not as expressive at the post-categorization level and cannot represent intra-POS polysemy by design. To compensate for this drawback, I have further adopted category theory, in particular its topos tool, to represent intra-POS polysemy. Moreover, I have translated the root-syntactic insights at the categorization level into the categorical language, thus reaching a unified representation of lexical ambiguity at the two levels. Category theory has been applied to both linguistics and computer science, which makes it a potential bridge for integrating theoretical linguistic insights into NLU development.

Acknowledgements

Thanks to the audience at the 11th International Conference on Meaning and Knowledge Representation (Almería, July 3–5, 2024) for valuable feedback. I also want to express my gratitude to Carlos Periñán-Pascual for his patient editorial help and to the anonymous reviewers for their constructive comments. All remaining errors are my own.

References

Acedo-Matellán, Víctor, and Cristina Real-Puigdollers. 2019. "Roots into Functional Nodes: Exploring Locality and Semi-Lexicality." *The Linguistic Review* 36 (3): 411–436. https://doi.org/10.1515/tlr-2019-2019.

Agirre, Eneko, and Philip Edmonds. 2007. "Introduction." In *Word Sense Disambiguation: Algorithms and Applications*, edited by Eneko Agirre and Philip Edmonds, 1–28. New York: Springer.

Alexiadou, Artemis, Hagit Borer, and Florian Schäfer, eds. 2014. *The Syntax of Roots and the Roots of Syntax*. Oxford: Oxford University Press.

Antunes, Cauã, Alexandre Rademaker, and Mara Abel. 2019. "A Category-Theoretic Approach for the Detection of Conservativity Violations in Ontology Alignments." In *Proceedings of the XII Seminar on Ontology Research in Brazil and III Doctoral and Masters Consortium on Ontologies*, edited by João Paulo A. Almeida, Marcello Bax, Rita Berardi, and Fernanda Baião, 11–20. https://ceur-ws.org/Vol-2519/paper1.pdf.

Apresjan, Ju. D. 1974. "Regular Polysemy." *Linguistics* 12 (142): 5–32.

Arad, Maya. 2005. *Roots and Patterns: Hebrew Morpho-Syntax*. Dordrecht: Springer.

Armstrong, Blair C., and David C. Plaut. 2016. "Disparate Semantic Ambiguity Effects from Semantic Processing Dynamics Rather than Qualitative Task Differences." *Language, Cognition and Neuroscience* 31 (7): 940–966. https://doi.org/10.1080/23273798.2016.1171366.

Asher, Nicholas. 2011. *Lexical Meaning in Context: A Web of Words*. Cambridge: Cambridge University Press.

Awodey, Steve. 2010. *Category Theory*. 2nd ed. Oxford: Oxford University Press.

Babonnaud, William. 2019. "A Topos-Based Approach to Building Language Ontologies." In *Formal Grammar 24th International Conference Proceedings*, edited by Raffaella Bernardi, Greg Kobele, and Sylvian Pogodalla, 18–34. Berlin: Springer. https://doi.org/10.1007/978-3-662-59648-7_2.

Babonnaud, William. 2021. "On the Dual Interpretation of Nouns as Types and Predicates in Semantic Type Theories." In *Proceedings of the ESSLLI 2021 Workshop on Computing Semantics with Types, Frames and Related Structures*, edited by Stergios Chatzikyriakidis and Rainer Osswald, 15–24. https://aclanthology.org/2021.cstfrs-1.2.

Babonnaud, William. 2022. "Sémantique Lexicale, Compositionnalité et Coercions. Fondements Théoriques des Types Sémantiques." PhD diss., University of Lorraine.

Barba, Edoardo, Tommaso Pasini, and Roberto Navigli. 2021. "ESC: Redesigning WSD with Extractive Sense Comprehension." In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4661–4672. https://doi.org/10.18653/v1/2021.naacl-main.371.

Beekhuizen, Barend, Blair C. Armstrong, and Suzanne Stevenson. 2021. "Probing Lexical Ambiguity: Word Vectors Encode Number and Relatedness of Senses." *Cognitive Science* 45: e12943. https://doi.org/10.1111/cogs.12943.

Bell, John L. 1988. *Toposes and Local Set Theories: An Introduction*. New York: Dover Publications.

Bender, Emily M., and Alexander Koller. 2020. "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, edited by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, 5185–5198. https://aclanthology.org/2020.acl-main.463.

Berwick, Robert C., and Noam Chomsky. 2016. *Why Only Us? Language and Evolution*. Cambridge, MA: MIT Press.

Bevilacqua, Michele, and Roberto Navigli. 2020. "Breaking Through the 80% Glass Ceiling: Raising the State of the Art in Word Sense Disambiguation by Incorporating Knowledge Graph Information." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2854–2864. https://doi.org/10.18653/v1/2020.acl-main.255.

Bevilacqua, Michele, Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. "Recent Trends in Word Sense Disambiguation: A Survey." In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence,* 4330–4338. https://doi.org/10.24963/ijcai.2021/593.

Black, William J., and Sabri El-Kateb. 2004. "A Prototype English-Arabic Dictionary Based on WordNet." In *Proceedings of the Second International WordNet Conference*, edited by Petr Sojka, Karel Pala, Pavel Smrz, Christiane Fellbaum, and Piek Vossen, 67–74. Brno: Masaryk University.

Borer, Hagit. 2005a. *Structuring Sense, Volume 1: In Name Only.* Oxford: Oxford University Press.

Borer, Hagit. 2005b. *Structuring Sense, Volume 2: The Nominal Course of Events.* Oxford: Oxford University Press.

Borer, Hagit. 2013. *Structuring Sense, Volume 3: Taking Form.* Oxford: Oxford University Press.

Camacho-Collados, Jose, and Mohammad Taher Pilehvar. 2018. "From Word to Sense Embeddings: A Survey on Vector Representations of Meaning." *Journal of Artificial Intelligence Research* 63: 743–788. https://doi.org/10.1613/jair.1.11259.

Cardinaletti, Anna, and Giuliana Giusti. 2001. "'Semi-Lexical' Motion Verbs in Romance and Germanic." In *Semi-Lexical Categories: The Function of Content Words and the Content of Function Words*, edited by Norbert Corver and Henk van Riemsdijk, 371–414. Berlin: Mouton de Gruyter.

Cassani, Giovanni, Federico Bianchi, Giuseppe Attanasio, Marco Marelli, and Fritz Guenther. 2023. "Meaning Modulations and Stability in Large Language Models: An Analysis of BERT Embeddings for Psycholinguistic Research." PsyArXiv. October 11. https://doi.org/10.31234/osf.io/b45ys.

Cavirani-Pots, Cora. 2020. "Roots in Progress: Semi-Lexicality in the Dutch and Afrikaans Verbal Domain." PhD diss., KU Leuven.

Cevoli, Benedetta, Chris Watkins, Yang Gao, and Kathleen Rastle. 2023. "Shades of Meaning: Uncovering the Geometry of Ambiguous Word Representations Through Contextualised Language Models." ArXiv. April 26. https://doi.org/10.48550/arXiv.2304.13597.

Chomsky, Noam. 1970. "Remarks on nominalization." In *Reading in English Transformational Grammar*, edited by Roderick A. Jacobs and Peter S. Rosenbaum, 184–221. Waltham: Ginn.

Clear, Jeremy H. 1993. "The British National Corpus." In *The Digital Word: Text-Based Computing in the Humanities*, edited by George P. Landow and Paul Delany, 163–187. Cambridge, MA: MIT Press.

Coecke, Bob, Edward Grefenstette, and Mehrnoosh Sadrzadeh. 2013. "Lambek vs. Lambek: Functorial Vector Space Semantics and String Diagrams for Lambek Calculus." *Annals of Pure and Applied Logic* 164 (11): 1079–1100. https://doi.org/10.1016/j.apal.2013.05.009.

Coecke, Bob, Mehrnoosh Sadrzadeh, and Stephen Clark. 2010. "Mathematical Foundations for a Compositional Distributional Model of Meaning." *Linguistic Analysis* 36 (1/4): 345–384.

Conathan, Lisa. 2004. "Classifiers in Yurok, Wiyot, and Algonquian." In *Proceedings of the Thirtieth Annual Meeting of the Berkeley Linguistics Society: Special Session on the Morphology of American Indian Languages*, edited by Marc Ettlinger, Nicholas Fleisher, and Mischa Park-Doob, 22–33. Berkeley: Berkeley Linguistics Society.

Corver, Norbert, and Henk van Riemsdijk, eds. 2001. *Semi-Lexical Categories: The Function of Content Words and the Content of Function Words*. Berlin: Mouton de Gruyter.

Crole, Roy L. 1993. *Categories for Types*. Cambridge: Cambridge University Press.

Cruse, David A. 1986. *Lexical Semantics*. Cambridge: Cambridge University Press.

Cruse, David A. 2000. "Aspects of Themicro-Structure of Word Meanings." In *Polysemy: Theoretical and Computational Approaches*, edited by Yael Ravin and Claudia Leacock, 30–51. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780198238423.003.0002.

Cuervo, María Cristina. 2003. "Datives at Large." PhD diss., Massachusetts Institute of Technology.

De Belder, Marijke. 2013. "Collective Mass Affixes: When Derivation Restricts Functional Structure." *Lingua* 126: 32–50. https://doi.org/10.1016/j.lingua.2012.11.008.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, 4171–4186. https://doi.org/10.18653/v1/N19-1423.

Doron, Edit, ed. 2014. *Theoretical Linguistics* 40 (3/4)*: On the Identity of Roots*. Berlin: Mouton de Gruyter.

Dölling, Johannes. 2020. *Systematic Polysemy*. In *The Wiley Blackwell Companion to Semantics*, edited by Daniel Gutzmann, Lisa Matthewson, Cécile Meier, Hotze Rullmann, and Thomas Ede Zimmermann. https://doi.org/10.1002/9781118788516.sem099.

Edmonds, Philip. 2006. "Disambiguation, Lexical." In *Encyclopedia of Language and Linguistics*, 2nd ed., edited by Keith Brown, 607–623. Elsevier. https://doi.org/10.1016/B0-08-044854-2/00949-4.

Eilenberg, Samuel, and Saunders Mac Lane. 1945. "General Theory of Natural Equivalences." *Transactions of the American Mathematical Society* 58: 231–294.

Falkum, Ingrid Lossius, and Agustin Vicente. 2015. "Polysemy: Current Perspectives and Approaches." *Lingua* 157: 1–16. https://doi.org/10.1016/j.lingua.2015.02.002.

Fellbaum, Christiane. 2006. "WordNet(s)." In *Encyclopedia of Language and Linguistics*, 2nd ed., edited by Keith Brown, 665–670. Elsevier. https://doi.org/10.1016/B0-08-044854-2/00946-9.

Folli, Raffaella, and Heidi Harley. 2005. "Flavors of *v*." In *Aspectual Inquiries*, edited by Paula Kempchinsky and Roumyana Slabakova, 22–33. Dordrecht: Springer.

Fong, Brendan, and David I. Spivak. 2019. *An Invitation to Applied Category Theory: Seven Sketches in Compositionality*. Cambridge: Cambridge University Press.

Geeraerts, Dirk. 1993. "Vagueness's Puzzles, Polysemy's Vagaries." *Cognitive Linguistics* 4 (3): 223–272. https://doi.org/10.1515/cogl.1993.4.3.223.

Gerdts, Donna B., and Mercedes Q. Hinkson. 1996. "Salish Lexical Suffixes: A Case of Decategorialization." In *Conceptual Structure, Discourse and Language*, edited by Adele E. Goldberg, 163–176. Stanford, CA: CSLI Publications.

Goldblatt, Robert. 1984. *Topoi: The Categorial Analysis of Logic*. Revised ed. Amsterdam: Elsevier.

Haber, Janosch, and Massimo Poesio. 2024. "Polysemy—Evidence from Linguistics, Behavioral Science, and Contextualized Language Models." *Computational Linguistics* 50 (1): 351–417. https://doi.org/10.1162/coli_a_00500.

Halle, Morris, and Alec Marantz. 1993. "Distributed Morphology and the Pieces of Inflection." In *The View from Building* 20*: Essays in Linguistics in Honor of Sylvain Bromberger*, edited by Ken Hale and S. Jay Keyser, 111–176. Cambridge, MA: MIT Press.

Halle, Morris, and Alec Marantz. 1994. "Some Key Features of Distributed Morphology." In *MIT Working Articles in Linguistics* 21, edited by Andrew Carnie, Heidi Harley, and Tony Bures, 275–288. Cambridge, MA: MIT Press.

Healy, Michael John, and Thomas Preston Caudell. 2006. "Ontologies and Worlds in Category Theory: Implications for Neural Systems." *Axiomathes* 16: 165–214. https://doi.org/10.1007/s10516-005-5474-1.

Hitzler, Pascal, Markus Krötzsch, Marc Ehrig, and York Sure. 2005. "What is Ontology Merging? A Category-Theoretic Perspective Using Pushouts." In *Proceedings of the First International Workshop on Contexts and Ontologies: Theory, Practice and Applications (C&0)*, edited by Pavel Shvaiko, Jerome Euzenat, Alain Leger, Deborah L. McGuinness, and Holger Wache, 104–107. Menlo Park, CA: AAAI Press.

Huang, Luyao, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019. "GlossBERT: BERT for Word Sense Disambiguation with Gloss Knowledge." In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3509–3514. https://doi.org/10.18653/v1/D19-1355.

Johnson, Michael, and Robert Rosebrugh. 2010. "The Institutional Approach." In *Theory and Applications of Ontology: Computer Applications*, edited by Roberto Poli, Michael Healy, and Achilles Kameas, 145–60. Dordrecht: Springer.

Kartsaklis, Dimitrios. 2014. "Compositional Distributional Semantics with Compact Closed Categories and Frobenius Algebras." PhD diss., University of Oxford.

Lambek, Joachim. 1999. "Type Grammar Revisited." In *Logical Aspects of Computational Linguistics: LACL'97 Selected Articles*, edited by Alain Lecomte, François Lamarche, and Guy Perrier, 1–27. Berlin: Springer. https://doi.org/10.1007/3-540-48975-4_1.

Lenci, Alessandro. 2023. "Understanding Natural Language Understanding Systems." *Sistemi Intelligenti* 2: 227–302. https://www.rivisteweb.it/doi/10.1422/107438.

Lenci, Alessandro, and Magnus Sahlgren. 2023. *Distributional Semantics*. Cambridge: Cambridge University Press.

Lowenstamm, Jean. 2008. "On Little n, √, and Types of Nouns." In *Sounds of Silence: Empty Elements in Syntax and Phonology*, edited by Jutta Hartmann, Veronika Hegedűs, and Henk van Riemsdijk, 105–144. Amsterdam: Elsevier.

Luo, Fuli, Tianyu Liu, Qiaolin Xia, Baobao Chang, and Zhifang Sui. 2018. "Incorporating Glosses into Neural Word Sense Disambiguation." In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Volume 1*, 2473–2482. https://doi.org/10.18653/v1/P18-1230.

Lyons, John. 1977. *Semantics*. Cambridge: Cambridge University Press.

Mac Lane, Saunders. 1998. *Categories for the Working Mathematician*. 2nd ed. New York: Springer.

Marantz, Alec. 1997. "No Escape from Syntax: Don't Try Morphological Analysis in the Privacy of Your Own Lexicon." *UPenn Working Articles in Linguistics* 4 (2), 201–225.

Mikolov, Tomas, Kai Chen, Greg Corrado, Jeffrey Dean. 2013. "Efficient Estimation of Word Representations in Vector Space." ArXiv. September 7. https://doi.org/10.48550/arXiv.1301.3781.

Miller, George A. 1995. "Wordnet: A Lexical Database for English." *Communications of ACM* 38 (11): 39–41. https://doi.org/10.1145/219717.219748.

Miller, George A., Claudia Leacock, Randee Tengi, and Ross T. Bunker. 1993. "A Semantic Concordance." In *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21–24*, 303–308. https://aclanthology.org/H93-1061.

Navigli, Roberto. 2009. "Word Sense Disambiguation: A Survey." *ACM Computing Surveys (CSUR)* 41 (2): 1–69. https://doi.org/10.1145/1459352.1459355.

Navigli, Roberto, and Simone Paolo Ponzetto. 2012. "BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network." *Artificial Intelligence* 193: 217–250. https://doi.org/10.1016/j.artint.2012.07.001.

Pasini, Tommaso. 2020. "The Knowledge Acquisition Bottleneck Problem in Multilingual Word Sense Disambiguation." In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, 4936–4942. https://doi.org/10.24963/ijcai.2020/687.

Piedeleu, Robin. 2014. "Ambiguity in Categorical Models of Meaning." MSc diss., University of Oxford.

Piedeleu, Robin, Dimitri Kartsaklis, Bob Coecke, and Mehrnoosh Sadrzadeh. 2015. "Open System Categorical Quantum Semantics in Natural Language Processing." In *Proceedings of the 6th Conference on Algebra and Coalgebra in Computer Science*, edited by Larry Moss and Paweł Sobociński, 267–286. Schloss Dagstuhl: Leibniz-Zentrum für Informatik. https://doi.org/10.4230/LIPIcs.CALCO.2015.270 .

Pustejovsky, James. 1995. *The Generative Lexicon*. Cambridge, MA: MIT Press.

Pustejovsky, James. 1998. "The Semantics of Lexical Underspecification." *Folia Linguistica* 32 (3/4): 323–347.

Raganato, Alessandro, Jose Camacho-Collados, and Roberto Navigli. 2017. "Word Sense Disambiguation: A Unified Evaluation Framework and Empirical Comparison." In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Volume 1*, 99–110. https://aclanthology.org/E17-1010.

Reformat, Marek Z., Giuseppe D'Aniello, and Matteo Gaeta. 2018. "Knowledge Graphs, Category Theory and Signatures." *International Journal on Semantic Web and Information Systems* 14 (1): 23–44. https://doi.org/10.1109/WI.2018.00-49.

Rodd, Jennifer M. 2018. "Lexical Ambiguity." In *The Oxford Handbook of Psycholinguistics*, 2nd ed., edited by Shirley-Ann Rueschemeyer and M. Gareth Gaskell, 96–117. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198786825.013.5.

Rodd, Jennifer M. 2020. "Settling Into Semantic Space: An Ambiguity-Focused Account of Word-Meaning Access." *Perspectives on Psychological Science* 15 (2): 411–427. https://doi.org/10.1177/1745691619885860.

Rodd, Jennifer M., M. Gareth Gaskell, and William D. Marslen-Wilson. 2004. "Modelling the Effects of Semantic Ambiguity in Word Recognition." *Cognitive Science* 28: 89–104. https://doi.org/10.1016/j.cogsci.2003.08.002.

Scarlini, Bianca, Tommaso Pasini, and Roberto Navigli. 2020. "With More Contexts Comes Better Performance: Contextualized Sense Embeddings for All-Round Word Sense Disambiguation." In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 3528–3539. https://doi.org/10.18653/v1/2020.emnlp-main.285.

Schütze, Hinrich. 1998. "Automatic Word Sense Discrimination." *Computational Linguistics* 24 (1): 97–124.

Song, Chenchen. 2019. "On the Formal Flexibility of Syntactic Categories." PhD diss., University of Cambridge.

Song, Chenchen. 2021a. "A Typology of Semilexicality and the Locus of Grammatical Variation." Paper presented at the 9th International Conference on Formal Linguistics (ICFL9), Shanghai (Online), November 5–7.

Song, Chenchen. 2021b. "On the Semantics of Root Syntax: Challenges and Directions." In *Proceedings of the 18th International Workshop of Logic and Engineering of Natural Language Semantics 18 (LENLS18)*, 61–74.

Song, Chenchen. 2022a. "Sentence-Final Particle vs. Sentence-Final Emoji: The Syntax-Pragmatics Interface in the Era of Computer-Mediated Communication." In *Proceedings of Grapholinguistics in the 21st Century, 2022*, edited by Yannis Haralambous, 157–192. Brest: Fluxus Editions. https://doi.org/10.36824/2022-graf-song.

Song, Chenchen. 2022b. "Sentence-Final Particle vs. Sentence-Final Emoji: The Syntax-Pragmatics Interface in the Era of CMC (extended version)." Talk at SyntaxLab, Cambridge (Online), June 28.

Spivak, David I. 2014. *Category Theory for the Sciences*. Cambridge, MA: MIT Press.

Spivak, David I., and Robert E. Kent. 2012. "Ologs: A Categorical Framework for Knowledge Representation." *PLoS ONE* 7 (1): e24274. https://doi.org/10.1371/journal.pone.0024274.

Sun, Tianxiang, Yunfan Shao, Xipeng Qiu, Qipeng Guo, Yaru Hu, Xuanjing Huang, and Zheng Zhang. 2020. "CoLAKE: Contextualized Language and Knowledge Embedding." In *Proceedings of the 28th International Conference on Computational Linguistics (COLING 2020)*, 3660–3670. https://doi.org/10.18653/v1/2020.coling-main.327.

Sun, Yu, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, et al. 2021. "ERNIE 3.0: Large-Scale Knowledge Enhanced Pre-Training for Language Understanding and Generation." ArXiv. July 5, 2021. https://doi.org/10.48550/arXiv.2107.02137.

Tuggy, David. 1993. "Ambiguity, Polysemy, and Vagueness." *Cognitive Linguistics* 4 (3): 273–290. https://doi.org/10.1515/cogl.1993.4.3.273.

Valera, Salvador. 2020. "Polysemy Versus Homonymy." *Oxford Research Encyclopedia of Linguistics*. https://doi.org/10.1093/acrefore/9780199384655.013.617.

Van Valin, Robert D. 2016. "From NLP to NLU." Unpublished manuscript. https://www.ling.hhu.de/fileadmin/redaktion/Oeffentliche_Medien/Fakultaeten/Philosophische_Fakultaet/Sprache_und_Information/Van_Valin_From_NLP_to_NLU.pdf.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. "Attention is All You Need." In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, 5998–6008. Red Hook, NY: Curran Associates, Inc.

Vicente, Agustín, and Ingrid L. Falkum. 2017. "Polysemy." *Oxford Research Encyclopedia of Linguistics*. https://doi.org/10.1093/acrefore/9780199384655.013.325.

Wang, Daphne Pauline. 2024. "A Quantum-Inspired Analysis of Human Disambiguation Processes: Foundational Theory and Applications." PhD diss., University College London.

Wang, Xiaozhi, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2021. "KEPLER: A Unified Model for Knowledge Embedding and Pre-Trained Language Representation." *Transactions of the Association for Computational Linguistics* 9: 176–194. https://doi.org/10.1162/tacl_a_00360.

Zimmermann, Antoine, Markus Krötzsch, Jérôme Euzenat, and Pascal Hitzler. 2006. "Formalizing Ontology Alignment and Its Operations with Category Theory." In *Proceedings of the 4th International Conference on Formal Ontology in Information Systems (FOIS 2006)*, edited by Brandon Bennett and Christiane Fellbaum, 329–340. Baltimore, MD: IOS Press.