Survey

Polysemy—Evidence from Linguistics, Behavioral Science, and Contextualized Language Models

Janosch Haber
School of Electronic Engineering and
Computer Science
Queen Mary University of London,
Chattermill
janoschhaber@gmail.com

Massimo Poesio
Queen Mary University of London,
Utrecht University

Polysemy is the type of lexical ambiguity where a word has multiple distinct but related interpretations. In the past decade, it has been the subject of a great many studies across multiple disciplines including linguistics, psychology, neuroscience, and computational linguistics, which have made it increasingly clear that the complexity of polysemy precludes simple, universal answers, especially concerning the representation and processing of polysemous words. But fuelled by the growing availability of large, crowdsourced datasets providing substantial empirical evidence; improved behavioral methodology; and the development of contextualized language models capable of encoding the fine-grained meaning of a word within a given context, the literature on polysemy recently has developed more complex theoretical analyses.

In this survey we discuss these recent contributions to the investigation of polysemy against the backdrop of a long legacy of research across multiple decades and disciplines. Our aim is to bring together different perspectives to achieve a more complete picture of the heterogeneity and complexity of the phenomenon of polysemy. Specifically, we highlight evidence supporting a range of hybrid models of the mental processing of polysemes. These hybrid models combine elements from different previous theoretical approaches to explain patterns and idiosyncrasies in the processing of polysemous that the best known models so far have failed to account for. Our literature review finds that (i) traditional analyses of polysemy can be limited in their generalizability by loose definitions and selective materials; (ii) linguistic tests provide useful evidence on individual cases, but fail to capture the full range of factors involved in the processing

Action Editor: Zhiyuan Liu. Submission received: 25 February 2023; revised version received: 11 September 2023; accepted for publication: 25 October 2023.

https://doi.org/10.1162/coli_a_00500

of polysemous sense extensions; and (iii) recent behavioral (psycho) linguistics studies, large-scale annotation efforts, and investigations leveraging contextualized language models provide accumulating evidence suggesting that polysemous sense similarity covers a wide spectrum between identity of sense and homonymy-like unrelatedness of meaning.

We hope that the interdisciplinary account of polysemy provided in this survey inspires further fundamental research on the nature of polysemy and better equips applied research to deal with the complexity surrounding the phenomenon, for example, by enabling the development of benchmarks and testing paradigms for large language models informed by a greater portion of the rich evidence on the phenomenon currently available.

1. Introduction

In the past few years, the Natural Language Processing (NLP) community has seen the emergence of so-called **contextualized language models** (Peters et al. 2018; Devlin et al. 2019; Radford et al. 2018; Raffel et al. 2020; Apidianaki 2023). These large neural networks with billions of parameters are designed not to encode the meaning of each word in a dictionary, but to produce unique encodings for a given input text—and each word within it. With this approach, contextualized language models appear to have solved the long-standing challenge of addressing word sense disambiguation in a manner that translates to actual performance gains in downstream tasks (cf. Ide and Véronis 1998; Navigli 2009; Bevilacqua et al. 2021).

Among other things, contextualized language models promise to address the issue of lexical ambiguity, generally focusing on cases of **homonymy**. Homonyms are words such as *match* that can take completely different meanings in different contexts (see, e.g., Weinreich 1964; Lyons 1977; Kempson 1977; Cruse 1986; Pinkal 1995; Klepousniotou et al. 2012):

Example 1

a. The **match** fell on the carpet and left a burn mark.
b. The **match** ended without a winner even after going into overtime.

a. The **bat** was found hibernating in an attic.
b. The **bat** was expertly crafted from a single piece of wood.

a. The **mole** dug a number of tunnels through the front yard.
b. The **mole** on her shoulder stopped bothering her after a while.

The interest in homonymy is not limited to research in computational linguistics and artificial intelligence. Homonymy and models for its mental processing has also been the research focus of a number of lexicographers, psycholinguists, and cognitive researchers in the past decades (Swinney 1979; Rodd, Gaskell, and Marslen-Wilson 2002, 2004). Homonymy, however, is only one form of lexical ambiguity. Closely related to it—but far less well understood—is the phenomenon of **polysemy**. Polysemous words also can assume different interpretations in different contexts; what distinguishes them from

homonyms is that their interpretations are closely related, and often invoke different aspects of or perspectives on the same concept (Apresjan 1974; Lyons 1977; Simpson 1994; Pinkal 1995; Cruse 1995; Pustejovsky 1995; Ravin and Leacock 2000; Nerlich and Clarke 2003; Vicente 2015; Dölling 2020). Take for example the different uses of *school* in Example 2, each of which intuitively reflects a different facet of the concept *school* rather than entirely unrelated meaning like the homonymic alternations in Example 1.

Example 2

a. The **school** has a dull brown facade. (building)

b. The **school** has prohibited light-up sneakers. (administration)

c. The **school** won last year's play-offs. (sports team)

d. The **school** is well respected among researchers. (institution)

Most if not all content words should be considered to be polysemous to some degree (Zipf 1945; Durkin and Manning 1989; Rodd, Gaskell, and Marslen-Wilson 2002; Travis 2008), rendering polysemy far more ubiquitous than homonymy (also see, e.g., Karjus et al. 2021; Brochhagen and Boleda 2022, for a recent investigation). Accumulating cognitive and linguistic evidence in recent years also suggests that the phenomenon of polysemy is far less homogeneous than assumed in earlier literature, which makes it more difficult to present clean-cut results or speak about polysemes in general terms. Eye-tracking as well as electro- and magnetoencephalography (EEG and MEG) studies, for example, have indicated that not only it is the case that the processing of homonyms differs from that of polysemes, but that different types and interpretations of polysemous sense extensions can also lead to notable differences in their mental processing (e.g., Frazier and Rayner 1990; Klepousniotou 2002; Pylkkänen, Llinás, and Murphy 2006; Klepousniotou et al. 2012; MacGregor, Bouwsema, and Klepousniotou 2015; Frisson 2015; Bruera et al. 2023). Recent studies leveraging large-scale annotated resources (see, e.g., Camacho-Collados and Pilehvar 2018; Murphy 2019; Armendariz et al. 2020; Trott and Bergen 2021) similarly find that annotators distinguish between polysemous senses less consistently than between homonymic meaning alternations.

Computational research on the interpretation of lexical ambiguity, or word sense disambiguation (WSD), had historically paid less attention to this fundamental distinction between homonymy and polysemy hypothesized by lexicographers and linguists. The sense repositories most widely used in NLP research—exemplified by WordNet (Fellbaum and Miller 1998)—usually represent all different interpretations of a word as distinct senses (Ide and Véronis 1998; Navigli 2009; Bevilacqua et al. 2021), irrespective of whether these are polysemous or homonymic sense extensions. (And thus can be argued to follow the so-called *Sense Enumeration* hypothesis [Katz and Fodor 1963], to be discussed later.) On the other end of the spectrum, a different strand of computational research was devoted to exploring what we will call the *One Representation* approach to the mental lexicon, where all interpretations are derived from a singular, common representation. This approach is most famously represented by Pustejovsky's *Generative Lexicon* (see, e.g., Pustejovsky 1995; Copestake and Briscoe 1995; Buitelaar 1998; Boleda, Schulte im Walde, and Badia 2012; Habibi, Hauer, and Kondrak 2021a). More data-driven computational research on word sense disambiguation however has provided increasing evidence for a more graded notion of sense

(Erk and McCarthy 2009; McCarthy, Apidianaki, and Erk 2016; Lau, Clark, and Lappin 2014). The most recent computational research enabled by the development of context sensitive-models (i.e., Camacho-Collados and Pilehvar 2018; Trott and Bergen 2021; Haber and Poesio 2021) provides further evidence for this graded view of sense distinctions. This recent research supports the theoretical perspective emerging from "modern" cognitive studies of polysemy such as Ortega-Andrés and Vicente (2019), who suggest a hybrid model for the mental representation of polysemes that incorporates a single lexicon entry for different senses as well as a notion of sense clustering or hierarchy to account for the complexity observed in the behavior of polysemous sense extension.

One of the key motivations for this survey is to present a comprehensive collection of arguments supporting this recent view that lexical ambiguity and especially polysemy are multi-faceted, heterogeneous phenomena—which should be reflected by the models and theories put forward to explain their representation and processing. Arguments and evidence are drawn from a wide array of disciplines in an attempt to provide a more complete picture of the heterogeneity and complexity of the phenomenon of polysemy. Our compilation is supported by thorough working definitions of polysemy proper and its various subtypes to unify terminology and facilitate interdisciplinary comparison. We then summarize seminal theories on the mental processing of polysemes, present an in-depth survey of behavioral evidence on the processing of polysemous words that highlights the differences between types of polysemous alternations, review corpus studies and large-scale datasets of empirical data on sense relatedness, and present an overview of traditional computational approaches and recent contextualized language models, exploring how these can be used to investigate polysemy. Our literature review finds that (i) traditional studies of polysemy can be limited in their generalizability by loose definitions and selective materials; (ii) linguistic tests can provide useful anecdotal evidence, but falter at capturing the full range of factors involved in the processing of polysemous sense extensions; and (iii) recent behavioral (psycho) linguistics studies, large-scale annotation efforts and investigations leveraging contextualized language models seem to provide accumulating evidence indicating that polyseme sense similarity covers a wide spectrum between identity of sense and homonymy-like unrelatedness of meaning.

We hope that this survey will lead to increased collaboration between research in lexicography, linguistics, psychology, cognitive neuroscience, and computational linguistics, and that it will provide a solid starting point for future research on polysemy and its representation. Expected benefits for computational research in particular include the ability to design new and improved benchmarks for testing large language models' capabilities of capturing complex distinctions in lexical meaning that will aid in improving their representation of ambiguity in future iterations.

2. Lexical Ambiguity: Homonymy, Polysemy, Vagueness, and Underspecification

Lexical ambiguity is a type of ambiguity that is due to word forms exhibiting multiplicity of meaning, namely, taking on different interpretations in different contexts (Lyons 1977; Pinkal 1995; Poesio 2020). Lexical ambiguity is ubiquitous in everyday language, and poses interesting questions and challenges to both (psycho-)linguists and computational linguists research: Why do we use ambiguous expressions? How do we process ambiguous words; how are they stored in our brains? And how should computational language models represent and deal with instances of lexical ambiguity?

2.1 Multiplicity of Meaning and Multiplicity of Sense

Modern investigations of multiplicity of meaning in the widest sense date back to at least Breal (1897), who noticed that many expressions in everyday interactions were ambiguous, but surprisingly rarely led to miscommunication. From among the different phenomena of ambiguity observed in natural language (see Poesio 2020, for a recent overview), this survey focuses on (one type of) lexical ambiguity, the phenomenon of a single word exhibiting multiplicity of meaning, i.e., allowing for different interpretations in different contexts.

Traditionally, phenomena of lexical ambiguity are further subdivided into homonymy and polysemy, separating *multiplicity of meaning* from *multiplicity of sense*. In an attempt to better distinguish these two phenomena, Weinreich (1964), for example, notes that "homonymy is observed in lexical items that *accidentally* carry two distinct and unrelated meanings" (see also Klepousniotou et al. 2012). This notion of two things "accidentally" being assigned the same name also is reflected in the choice of its description, with the term *homonym* being derived from Ancient Greek ὁμο- (homo-, *same*) and ὄνυμα (ónuma, *name*). Seminal, paradigmatic examples of homonyms include nouns like *match*, *bat*, and *mole* as shown in Example 1, which can invoke at least two different, arguably unrelated interpretations. Contrasting the unrelatedness of homonymic interpretations, *polysemy*[1] traditionally has come to signify words that can invoke different distinct but related interpretations (see, e.g., Ullmann 1959; Apresjan 1974; Lyons 1977; Swinney 1979; Bierwisch and Schreuder 1992; Simpson 1994; Pinkal 1995; Cruse 1995; Ravin and Leacock 2000), like *school* in Example 2. These two phenomena are set in opposition to *monosemy*, where words are assumed to be associated with just one, fixed interpretation.

Lexical ambiguity is a ubiquitous phenomenon. Durkin and Manning (1989), for example, estimate that 40% of frequent English words are polysemous, while scholars like Zipf (1945), Rodd, Gaskell, and Marslen-Wilson (2002), and Travis (2008) even argue that basically every content word can be used polysemically—a notion we will explore in Section 2.3. Estimates for homonymy are more conservative, with, for example, Dautriche (2015) suggesting that only about 4% of English words can have multiple, unrelated meanings (see, e.g., Karjus et al. 2021; Brochhagen and Boleda 2022, for a recent investigation of the main cognitive drivers thought to cause this disparity).[2] Distinguishing homonymy from polysemy based on the notion of "relatedness of meaning," however, also has been met with strong criticism (e.g., Lyons 1977; Pinkal 1995; Kilgarriff 1997): Relatedness itself is at best a vague proposition open to contextual biases, subjective judgment or "folk etymology," while determining homonymy based on historically "formally distinct items in some earlier stage of the language"

1 From Ancient Greek πολύς (polús, *many, much*) and σῆμᾰ (sêma, *mark, sign, token*).

2 A note on terminology: most linguistics literature will label a given word to be either a monoseme, polyseme, or homonym. As however many (if not all) content words—including homonyms—can have polysemous sense alternations, labeling specific words as *homonyms* or *polysemes* strictly speaking is not very meaningful, and instead their different interpretations should be considered polysemous or homonymic in their relation to one another. With that in mind, in the remainder of this survey we will use the terms *homonymic* or *polysemous* to refer to specific alternations of a given ambiguous word. To further clarify this distinction, we will refer to the different polysemous extensions of a word as different word *senses*, and the different interpretations of homonymic alternations as *meanings*. Note that this naming convention will be in conflict with some previous literature using these terms to refer to either or neither of the phenomena in particular.

(Klepousniotou 2002) suffers from unclear historical derivations, and begs the question how far back one should go in tracing the history of words (Lyons 1977; Pinkal 1995). The resulting vague boundary between polysemy and homonymy is at least partially responsible for the sometimes conflicting observations about the processing of homonymic and polysemous words in previous literature. When presenting previous work, we will therefore—whenever possible—aim to clarify how authors classify specific samples.

2.2 Polysemy: Types, Classes, and other Subdivisions

Besides dispute over and unclarity in their definition, a second central aspect making it difficult to draw a clear distinction between homonymy and polysemy is the observation that the latter phenomenon is quite heterogeneous in its appearance, with the perceived similarity of polysemous sense interpretations ranging from near identity to homonymy-like unrelatedness of meaning. To address this issue, a number of researchers have attempted the definition of subtypes, classes, and other distinctions of specific polysemous alternations, each presented in conjunction with hypotheses as to how the relatedness in sense interpretations evident in this subtype affects the processing of polysemes.

One of the most well-known and commonly accepted distinctions is that polysemous alternations are considered to be either *idiosyncratic* (sometimes labeled *accidental*), or *regular*. Following the definition of Apresjan (1974), a lexeme A with senses a_1 and a_2 is an example of regular polysemy if there exists at least a second lexeme B for which its senses b_1 and b_2 are "semantically distinguished in exactly the same way as a_1 and a_2"—although later publications like Falkum (2015), Vicente and Falkum (2017), and Ortega-Andrés and Vicente (2019) note that to actually exhibit regularity in its polysemous sense alternation, it should have more than one corresponding alternative. According to Vicente and Falkum (2017), "regular polysemy is typically associated with senses generated by metonymic extensions, and irregular polysemy with senses that are derived metaphorically" (also see Apresjan 1974), Bowdle and Gentner 2005). These descriptions link the observed relation between two different sense extensions to two more well-known figures of speech: metaphoric (from Greek μεταφορά [metaphorá], *transference*) sense alternations are observed in cases where an interpretation that is more inherent to one concept is transferred to another (unrelated) concept—but still evokes a related meaning. An example for this kind of metaphoric extension is noun *mouth* in Example 3:

Example 3

a. She has a number of freckles on her nose and close to her **mouth**.

b. The river never is more than 20 feet across, except close to its **mouth**.

Metonymic extensions (from Greek μετωνυμία, metōnymía, *a change of name*), on the other hand, usually indicate sense extensions referring to different aspects or facets of the same entity. The different uses of *school* in Example 2 illustrate this kind of polysemous sense extension, with different contexts invoking different aspects of the concept *school*.

A wide range of previous research has been focused on naming and specifying some of the most frequent alternations observed in regular polysemy, including, for example, *animal/meat* alternations (see Example 4, cf. Copestake and Briscoe 1995; Frisson and Frazier 2005; Falkum 2015), *container/content* alternations (Example 5, e.g., Schumacher 2013), or *physical/information* alternations (Example 6, cf. Pustejovsky 1993; Antunes and Chaves 2003; Frisson 2015):

Example 4

a. The **chicken** pecked for some seeds in the shadow of the barn.

b. The **chicken** was seasoned deliciously and served with potato wedges.

Example 5

a. Nervously waiting for his date, he peeled the label off his **beer**.

b. He didn't remember much as he had had way too much **beer**.

c. She carefully placed the priceless **bottle** in a padded box.

d. She only had about half a **bottle**, but she was feeling tipsy already.

Example 6

a. They found the **book** wedged under a window to create some airflow.

b. After two semesters they were able to cite most of the **book**.

These alternations are not only found in English,[3] but in many other languages as well, with Srinivasan and Rabagliati (2015) presenting evidence of 27 distinct cases of English polysemy also being present in 14 different languages, "suggesting that polysemy arises from conceptual constraints rather than arbitrary, language-specific conventions" (Murphy 2021).

Pustejovsky (1995) later introduced an even more fine-grained distinction between different types of regular polysemy: While some expressions are *merely regular*, they considered others to be *inherently* polysemous. For a term to exhibit inherent polysemy, the different senses need to be "somehow inherent to the entity that the term denotes." Ortega-Andrés and Vicente (2019), for example, propose that the noun *book* could be said to have inherently polysemous interpretations, as both the *physical* (He put the *book* back in the shelf) as well as the *information* reading (He read the *book* in under two hours) are inherent to what a book is. They however also argue that this characterization of inherent polysemy is rather vague, as there is no clear definition as to when certain sense interpretations are inherent or not.

Dölling (2020) offers an alternative distinction, contrasting *metonymic* and *inherent* polysemy. Following his definition, metonymic polysemy describes "cases where one of the related senses is primary and the others are metonymically derived from it,"

3 Or German, in the case of Schumacher (2013).

while inherent (or logical) polysemy "involves senses where there are no substantial reasons for assuming that one or another of them is prior" (based on earlier observations by, e.g., Nunberg 1995; Copestake and Briscoe 1995). From their collection of systematic polysemes, Dölling identifies nouns such as *rabbit, apple, oak, beer*, and *bottle* as exhibiting metonymic extensions where "even though each of the senses are in equal measure usual, one of them is primary," and all interpretations exhibit a "normal, conventionalised use." As a rule of thumb, alternations that can be described through patterns like *animal-for-meat*, *fruit-for-pulp*, or *container-for-content* are likely to be cases of metonymic polysemy. Inherent polysemous sense extension on the other hand were identified for nouns like *book, speech, newspaper*, and *lunch*, where "neither interpretation can be viewed as more basic." Potential targets here are words where function and physical realization both are integral—a book without its physical realization or content for example would arguably not be a book.

The much discussed distinction between homonymy and polysemy, as well as the number of proposed distinctions between polysemous subtypes, variants, and alternations patterns are a poignant first indicator of the heterogeneity and complexity of the phenomena involved in lexical ambiguity and specifically in polysemic sense alternation. Anecdotal evidence reveals not just paradigmatic examples for different behaviors, but often also presents equally valid exceptions to any rules based on them. Recent literature therefore more and more demands reliable empirical data to replace the use of anecdotal examples in the investigation of polysemous alternations (see, e.g., Klepousniotou, Titone, and Romero 2008; Erk and McCarthy 2009; Schumacher 2013; Ortega-Andrés and Vicente 2019; Löhr 2021; Haber and Poesio 2021). Before we turn to these studies, we need to mention two more concepts closely tied in to the interpretation of lexical ambiguous expressions: vagueness and micro-senses, and the ambiguity advantage.

2.3 Polysemy, Vagueness, and Micro-Senses

Homonymy and polysemy are considered types of ambiguity rather than vagueness, that is, the potential interpretations of polysemous or homonymic expressions form a discrete set rather than a continuous transition. Vagueness, however, also plays a central role in the theoretical conceptualization of these phenomena and their definitions: When are two senses of a word different enough to be considered distinct interpretations? Where is the cutoff-point for the relatedness of interpretations? And how can we account for the infinite sets of (discourse and deictic) contexts impacting the meaning and interpretation of a word?

The latter question is prompted by a phenomenon which Cruse (2000) called *ways of seeing* and also *micro-senses* (also see Anderson and Ortony [1975] and Kilgarriff [1997] for a seminal account within computational linguistics). Micro-senses can best be understood looking at verbs like *run* that exhibit an extraordinary amount of sense productivity (Brugman 1988; Gilliver 2013). Consider the sentences in Example 7: While all of these uses of *run* elicit closely related interpretations, the different contexts ever so slightly change the meaning of the word.[4]

4 These are but a selection of the 645 meanings Gilliver compiled when revising the *Oxford English Dictionary*.

Example 7

a. John is **running** at least 5k every morning.
b. The bank robber is **running** from the police.
c. The dog is **running** in the park.
d. The water is **running** down the steps.
e. My nose just won't stop **running**.
f. The coffee machine is **running** all morning.

A well-known illustration of the discussion on micro-senses is the debate between Jackendoff (1989) and Fodor (1998), who over a period of time discussed specifically the verb *to keep*, with Jackendoff arguing that *keep* must surely be polysemous with its uses in phrases like *keep the change*, *keep your car in the garage*, *keep the crowd happy*, while Fodor argues that *keep* in fact only has a single meaning, and "the apparent differences in meaning are simply an artefact of the different contexts in which the verb appears" (also see Falkum and Vicente 2015). Seeing micro-sense variations as the main drive of polysemous sense extensions, some scholars take an entirely pragmatic approach in explaining phenomena of lexical ambiguity and specifically polysemy (which we will briefly discuss in Section 3.2.3), and others postulate that all content words are in fact polysemous (see, e.g., Zipf 1945; Travis 1997; Rodd, Gaskell, and Marslen-Wilson 2004). While the scope of the "multiplicity of sense" classification of polysemy allows for the inclusion of micro-senses, a definition like this blurs the line between ambiguity and vagueness, as no longer could all senses be clearly distinguished.

Tying together phenomena of vagueness and ambiguity in a single formalization, Pinkal (1985) proposed the concepts of *h-type* and *p-type* ambiguity to better classify lexical ambiguity (see also Poesio 2020). Following Pinkal's approach, an expression is h-type ambiguous if and only if its "indefinite base level is inadmissible." As a consequence, h-type ambiguous words have to be immediately disambiguated because they do not allow for an underspecified representation of their base level. P-type ambiguous words on the other hand do allow for an underspecified representation, and therefore do not require an immediate disambiguation. These formalizations of h-type and p-type ambiguity can directly be applied to homonymy and polysemy, suggesting that homonyms do not have an admissible underspecified base level, while polysemes do—allowing for underspecification in the interpretation of polysemes but not so for homonyms.

2.4 Recap

In this section we have shown that while already the distinction between homonymy and polysemy proves to be anything but clear cut, the delineation between polysemy and vagueness on the other end of the spectrum seems even more difficult to draw. This puts polysemy in a unique middle ground, highlighting the complexity of the phenomenon and stresses the importance of dedicated research under strict definitions and careful methodology. Most current investigations of polysemy find it useful to at least distinguish between metaphoric and metonymic sense extensions when categorizing their materials, while we will see that others even make explicit the specific alternation

under investigation to allow for precise observations. Limiting the scope of a study however also limits the extend to which its findings can be generalized, and as a result, the literature on polysemy oftentimes fluctuates between attempting wide coverage with limited materials and detailed studies of very specific aspects. In the next three sections, we will attempt to strike a useful balance between the two, presenting seminal and recent work on the mental processing of lexical ambiguity.

3. Polysemy and the Mental Lexicon

The organization of the human mental lexicon is a central aspect of theories on lexical ambiguity: if a single word can indeed have multiple senses, and sometimes even multiple meanings, how are these connections stored in our mental representation of those words? Or—more figuratively speaking—what is the makeup of our mental lexicon? The linguistic and psychological literature has produced a range of proposals attempting to answer these questions, commonly split into three groups: *Sense Enumeration* approaches, *One Representation* models, and *Pragmatic* approaches.

3.1 The Sense Enumeration Lexicon

One of the earliest models of the mental lexicon was offered by Katz and Fodor (1963) and Katz (1972), who included in the grammar of their natural language semantics model a dictionary in which all senses of a word were to be listed, that, taken together, constitute a word's meaning. This type of mental representation later has come to be known as a Sense Enumeration approach, or **Sense Enumeration Lexicon** (SEL). SEL approaches usually do not make a principled distinction between homonymic and polysemous interpretations, with either being listed as just another possible interpretation of a given word.[5] This approach to word meaning is inspired fairly directly by the type of lexical representation found in so-called "splitting" dictionaries (see, e.g., Jackson 2002), and it underlies one of the most widely used lexical resources in NLP, the WordNet lexical database discussed in Section 6.1.2.

As Falkum and Vicente (2015) noted, Sense Enumeration models are "*prima facie* the simplest way to deal with polysemy on theoretical grounds," explaining all variability in the semantic contribution of an expression through its "different senses stored as distinct representations." SEL approaches have not, however, received much support from the academic community. Given the previously mentioned observations that polysemy is a pervasive phenomenon and that some words can have up to hundreds of possible meanings and sense interpretations, assuming individual entries for all of them would require an immense storage complexity and cause a combinatorial explosion when processing sentences containing multiple ambiguous words. Similarly, a number of philosophical concerns have been raised concerning definitional theories in general, with scholars like Kilgarriff (1997) lamenting the difficulty in "deciding when two senses are different enough to warrant a new entry, and how to represent the information that is common to multiple different senses" and Hanks (2000) questioning whether different senses actually can be represented as disjoint classes defined by

5 Defenders of the model may distinguish between polysemy and homonymy based on whether the different senses or meanings belong to a single lexical entry—but ultimately both are stored as distinct representations (Falkum and Vicente 2015).

necessary and sufficient conditions (also see Wittgenstein [1953] for an early discussion, as well as Tuggy [1993] and Laurence and Margolis [1999]). More recently—and more specifically—Vicente and Falkum (2017) noted that semantic markers proposed to distinguish senses in SEL approaches cannot account for many of the observed polysemous alternations, and Dölling (2020) remarked that Sense Enumeration accounts "miss the generalization that can be made with regard to the underlying patterns of multiple meaning" and, as a consequence, "blur the distinction between homonymy, non-systematic polysemy and systematic polysemy, and ultimately denies the existence of the latter."[6]

3.2 One Representation Models

Nowadays, the best known proposals about polysemy in theoretical linguistics can be said to subscribe to a so-called **One Representation** model of polysemy in the mental lexicon. In One Representation models, the "senses of a polysemous expression either belong or depend on a single representation" (Falkum and Vicente 2015). One representation models often are also called underspecification accounts, since—in contrast to SEL models—they do not require the full specification of all sense interpretations, but instead postulate a single, underspecified entry accessed for all interpretations of a polyseme. The question exactly how much semantic information is stored in this representation, however, divides the community (see, e.g., Caramazza and Grober 1976; Miller and Johnson-Laird 1976; Nunberg 1979), with proposals ranging from thin semantics models containing merely a set of constraints for what interpretations a word can take on, to rich semantics approaches postulating an over-specified core representation that immediately makes available all necessary information for any possible interpretation.

3.2.1 Thin Semantics. In thin semantics models, the mental representation of a word is "impoverished" compared to the meaning it can take on within a specific context (Falkum and Vicente 2015); that is, upon encountering a (polysemous) expression, an underspecified mental concept of its meaning is activated and subsequently enriched with relevant contextual information to form a specific interpretation. Thin semantics models often propose that the mental representation of a word is merely lexical, containing only information necessary to "constrain the range of concepts that words can express" (Ortega-Andrés 2021, also see Travis 2008; Falkum 2011; Carston 2013), or even that the underspecified representation is so thin that it carries no semantic content at all. Pietroski (2005), for example, proposed that the mental representation of a word is simply a set of "instructions for how to access and assemble concepts" (Ortega-Andrés 2021), linking at or pointing to a number of concepts involved in its realization.

When taking a thin semantics stance, mental representations of polysemous words are often brought back to Nunberg's core meaning approach, where "the semantic representation of polysemous terms consists in a set of features or a common core that is shared by all senses" of that expression (Falkum and Vicente 2015). This is best explained by Jackendoff's (1989) example of the verb *keep*, for which they postulate a mental meaning representation that simply states a core definition common to all

6 Dölling's observation links back to a concept sometimes called the *polysemy fallacy* as introduced by Sandra (1998), complaining that SEL approaches "fail to distinguish between those aspects of meaning that are part of the word meaning proper, and those that result from its interaction with the context" (Falkum and Vicente 2015).

interpretations, where *X* can take on different semantic values including possession, location, or memory:

Example 8

CAUSE [STATE OF X THAT ENDURES OVER TIME]

3.2.2 Rich Semantics. Rich semantics takes the opposite approach to defining the mental representation of a polysemous word by postulating that all semantic information necessary to specify its different interpretations is available in the lexicon entry. One of the most prominent and influential rich semantics models of polysemy—and arguably the best known approach to polysemy developed in Computational Linguistics—is the so-called **Generative Lexicon** (GL) theory originally proposed by Pustejovsky (1993, 1995). Pustejovsky argues that the inherent semantic information about a word is encoded in a lexical semantic level separated from less intrinsic and more general encyclopedic knowledge. The GL theory proposes that the lexical representation of meaning consists of four structures: an argument structure, an event structure, a lexical inheritance structure, and a qualia structure (see Figure 1).

The latter is the hallmark of Pustejovsky's model, designed to contain information on the roles that a word can fulfil in its different functions. This information includes aspects of "about how the object came into being (its agentive role), what kind of object it is (formal role), what it is for (telic role), and what it is constituted of (constitutive role)" (see also Falkum and Vicente 2015). As its name implies, in the generative lexicon

$$
\begin{bmatrix}
\textit{school} & \\
\text{ARGSTR} = & \begin{bmatrix}
\textit{arg1} = x \\
\textit{d-arg1} = z \\
\textit{d-arg2} = y \\
\textit{d-arg3} = w \\
\textit{d-arg3} = v \\
\textit{d-arg4} = u \\
\dots
\end{bmatrix} \\
\text{QUALIA} = & \begin{bmatrix}
F = \textit{institution}(x) \\
A = \textit{found}(z,x) \\
T = \textit{provide\_education\_to}(x,y) \\
C = \begin{bmatrix}
\textit{building\_that\_houses}(w,x), \\
\textit{teacher\_who\_teaches\_at}(v,x), \\
\textit{pupil\_who\_attends}(y,x), \\
\textit{club\_that\_represents}(t,x), \\
\textit{staff\_that\_works\_for}(s,x), \\
\dots
\end{bmatrix}
\end{bmatrix}
\end{bmatrix}
$$

Figure 1
Hypothesized qualia structure for the noun *school* (in the educational sense) based on Pustejovsky's Generative Lexicon approach, expressed as an Attribute-Value Matrix.

word meaning is *generated* by accessing specific information from this over-specified lexical entry when encountering a target word in a specific context.

For unambiguous words, the information contained in the qualia structure decides whether a word is permissible in a given context, that is, whether it fulfils the context's selectional restrictions. When a word is polysemous, it can fulfil different selectional restrictions. According to Pustejovsky, this means that at least logical polysemous words must have complex qualia structures that allow for the selection of different roles in different contexts. In order to specify these complex qualia structures, the generative lexicon postulates a special type, the so-called **dot objects** (see also Asher and Pustejovsky 2006; Asher 2011). Dot objects are "aggregate" types that combine at least two different senses into a single, underspecified type. Usually, these senses are inherent to the realization of a concept, and could be described as facets or aspects of the complex type (cf. Cruse 2004; Frisson 2009; Paradis 2004). As an example, the noun *book* would be represented as the dot object **physical object•information**, combining its realization as a *physical object* and its *information* or *content* sense. Based on this work, Arapinis and Vieu (2015) argue that words like *book* are complex "materialised informational contents" that correspond neither to the conjunction of their disjoint aspects like *physical medium* or *information object*, nor to their disjunction. Proposing these complex representations however goes hand in hand with requiring complex mechanisms for parsing ambiguous expressions, which seems to contradict the low cognitive effort associated with processing polysemous items in behavioral studies (see Section 4.2).

3.2.3 Literalist and Pragmatic Approaches. A third seminal approach to modeling the mental processing of (ambiguous) words proposes that for each word we store a single, "concrete and semantically determined representation" (Falkum and Vicente 2015), its **literal meaning**. Once this literal meaning has been activated, a context-specific interpretation is derived either through a set of lexical rules, or through pragmatic modulation.

Among the literature explaining mostly regular polysemous alternations through a set of lexical rules applied to an initial literal interpretation (see, e.g., Gillon 1992, 1999; Kilgarriff 1992; Ostler and Atkins 1991; Asher and Lascarides 2003), one of the most well-known proposals is Pelletier's (1975), and subsequently Copestake and Briscoe's (1995) work on the "universal grinder," a model explaining *count/mass* alternations like the famous "there was *rabbit* all over the highway" through a set of derivation rules. But while they gained some attention in formal and early computational semantics literature, rule-based literalist approaches have not received much support in recent years. One of the main reasons for this is that all rule-based approaches suffer from the limitation that they can only be applied to a small subset of the observed phenomena, and that even then they can be over-productive in some cases, requiring not only a formulation of derivation rules, but also a set of idiosyncratic exceptions to them. Falkum (2015), for example, lists three theoretical arguments undermining fully rule-based approaches to polysemous sense extension as offered by Copestake and Briscoe (1995) and Pustejovsky (1995). Firstly, it seems unclear how in a sentence like "Peter enjoyed the nice weather" the (assumed) intended reading of "Peter enjoyed *being outside in* the nice weather" could be generated when there is no telic information in the lexical representation of *weather* that could be used as input in the compositional process deriving this interpretation. Secondly, Falkum argues that it is difficult to see how rule-based accounts can avoid making wrong predictions about many compositional interpretations—for example, in the VP *begin a car*, which according to the telic function

of *car* should be interpreted as *begin driving a car*. And thirdly, she questions how a lexicon-internal, rule-based approach can account for the interpretative flexibility that is involved in the construction of (metonymic) polysemy. This concern is illustrated by sample sentences like Example 9 where their idiosyncrasy makes it unlikely that only lexical rules will have been involved in generating their interpretation.

Example 9

a. Will a hamster bite if it smells **rabbit** on my hands? (*rabbit odor*)
b. [Biology teacher]: **Rabbit** is smaller than hare. (*rabbit feces*)
c. [Hunter]: This time of year I prefer using **rabbit**. (*electronic rabbit calls*)
d. Last winter, we discovered **rabbit** and fox in our garden. (*rabbit tracks*)

Instead, Falkum favors a radical pragmatic account. Radical pragmatic approaches were common in early AI models, where all meanings would be generated via general commonsense reasoning (see, e.g., Hobbs et al. 1993) and are still favored by many cognitive linguists as an alternative to postulating rule-based derivations of contextualized interpretations from a literal meaning. As we briefly mentioned before, some scholars support the notion that basically every content word can be used polysemically. As this would entail an impossibly large number of senses or derivation rules that would be needed to be stored in our mental lexicon, pragmatic approaches suggest that we only store a single, fully conceptual representation of a word, and derive any contextualized readings pragmatically in an ad-hoc fashion (see, e.g., Recanati 1998; Carston 2002). According to Traugott (2017), "a fundamental claim in cognitive linguistics is that words do not have fixed meanings. They evoke meanings and are cues to potential meaning, instructions to create meanings as words are used in context" (also see, e.g., Brugman 1988; Kilgarriff 1997; Paradis 2011). As a consequence, radical pragmatic accounts "see the role of the linguistic system as being that of providing a minimal input or clue—a *sketch* or *blueprint* of the speaker's meaning—which the pragmatic inferential system uses as evidence to yield hypotheses about occasion-specific, speaker-intended meanings" (Falkum 2015).

3.3 Hybrid Models

Falkum (2015), however, also argues that while "overall, a radical pragmatic account provides the most promising basis for a unified account of the role of polysemy in several domains, [. . .] depending on their degree of conventionalisation, some senses may be stored in our mental lexicons, [and] some may be contextually derived," continuing the school of thought championed by the likes of Pustejovsky, and Asher and Lascarides. Returning to the *count/mass* alternation in *rabbit*, the authors therefore suggest that the input to the pragmatic processing of polysemes like this is composed of a rich, pragmatic representation of context and encyclopedic information, and a highly underspecified conceptualization of the target itself, which are combined to construct a narrower, ad-hoc concept (e.g., *rabbit meat*). Some of these constructions like the *animal/meat* alternation of words like *rabbit*, *chicken*, and *lamb* may become "progressively more routinised," developing "pragmatic routines" (cf. Vega Moreno 2007) that increase

the accessibility of certain interpretations. These regularities then are proposed to give rise to the "sense of regularity" observed in metonymic polysemes.

This view introduces a last variety to the range of mental models on the processing of ambiguous expressions, which we will preliminary label **hybrid models**, particularly popular in the Cognitive Linguistics literature (e.g., Cruse 1995, 2000). Hybrid models are usually based on one of the traditional mental models of language processing, but borrow some aspects of others. Klepousniotou, Titone, and Romero (2008), for example, note that while their experiments in principle support a rich underspecification model, they also find that "high-overlap polysemous words differ from moderate- and low-overlap ambiguous words in comparison [and] there are several potential ways in which they may differ in representation," suggesting that a more structured representation of polysemous word sense might replace a fully underspecified core entry. Similarly, Asher (2011) presented a different version of hybrid model, suggesting that pragmatics are involved whenever a non-default interpretation is involved. This fall-back is intended to augment his originally over-specified One Representation approach with pragmatic aspects for context coercion, but, while now allowing for these specific cases, does raise the question of when and how the fall-back is activated.

Ortega-Andrés and Vicente (2019) and Ortega-Andrés (2021) recently proposed a hierarchical ordering within the underspecified representation of polysemous sense to allow a traditional Pustejovskyan model to account for processing differences among polysemous senses. Based on a rich underspecification account, Ortega-Andrés and Vicente extend a target's knowledge structure with multiple realizers that each specify a certain range of interpretations of the overall concept. Figure 2 shows a schematic of the hierarchical structure for polyseme *school* in Ortega-Andrés and Vicente's model. According to their hypothesis, the different interpretations that can be invoked by a

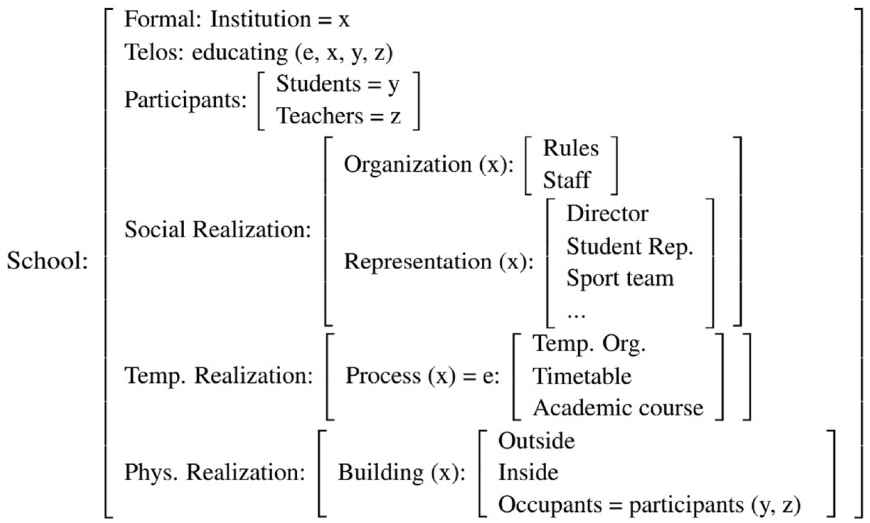

Figure 2
Schema of the knowledge structure of the polysemous realizations of word *school* according to the activation package model proposed by Ortega-Andrés and Vicente (2019). Figure reprinted with permission.

given realization (e.g., *rules* and *staff*) form so-called activation packages, groupings of interpretations that are so closely related to one another that an underspecified interpretation invoked by their realizer includes all of them simultaneously. This means that interpretations included in an activation package should allow for cost-free sense shifting, while moving to an interpretation evoked by a different realizer will lead to processing difficulties. Besides these explicit activation packages, a hierarchical representation like this however also implies an underlying notion of sense similarity which determines the representation—and consequently suggests at least different levels of similarity in the interpretation of polysemous senses.

3.4 Recap

The development of hybrid models to account for the processing of ambiguous expressions showcases a growing awareness that the observed complexity of phenomena surrounding lexical ambiguity and specifically polysemy demands a more involved representation of these expressions in our mental lexicon; Sense Enumeration approaches fall short in explaining different behavior of homonymic and polysemous items, fully underspecified One Representation models cannot account for processing differences between polysemous alternations of the same word, and pragmatic approaches still need to prove how semantic information for every possible sense extension of every single word can effectively and efficiently be stored in our language processor. Hybrid approaches now borrow from these principled schools of thought to explain nuances in observed behavior—and provide new inspiration for dedicated behavioral studies and empirically driven data annotation studies to illuminate previously ignored aspects of the phenomenon of polysemy.

4. Processing Polysemy: Behavioral Evidence

One of the main drives behind Ortega-Andrés and Vicente's model is to develop an architecture that can explain the processing differences between different polysemous sense extensions observed in recent *behavioral studies*. Behavioral studies focus on online effects as measured through, for example, reading times, eye-movements, or brain activations, investigating exactly *when* and *how* an ambiguous item is processed. Together, the answers to these two questions can provide much needed evidence to examine the different proposed models for the mental processing of ambiguous expression. If behavioral studies were to find no notable differences in the participant's processing of homonymic and polysemous samples, their distinction should be considered purely theoretical—which would support a simple Sense Enumeration approach to the mental lexicon. If polysemes appear to allow for underspecification where homonyms need disambiguation, this would favor One Representation models. And if indeed the processing of different types of polysemous alternations leads to different behavior, hybrid models are an interesting way forward.

4.1 Studies in Support of Sense Enumeration Approaches

Klein and Murphy (2001, 2002) and Foraker and Murphy (2012) are among the few authors not subscribing to an underspecified representation approach to the mental representation of polysemes, but instead present experimental evidence in favor of Sense Enumeration approaches.

Klein and Murphy (2001) introduce a range of experiments using word pairs like *shredded paper* and *liberal paper* in memory and sensicality judgment tasks to test differences in processing between phrases eliciting the same or different readings. In the first experiment, participants were shown a series of word pairs which they were asked to remember, and subsequently presented test items displayed like *daily PAPER*, for which they were asked to decide as quickly as possible whether they had seen the highlighted target word in the previously shown list. The test items all were chosen to be polysemous and divided into three conditions: a *same phrase* condition, a *consistent sense* condition, and an *inconsistent sense* condition. The authors found that if the same phrase was shown during testing, items were judged most accurately (at 79% correct),[7] followed by consistent sense samples (64% accuracy) and inconsistent ones (56%). Klein and Murphy take these results to show that when participants are presented with a different reading of a polyseme during testing than was included in the memorization list, the mismatch in representation will lead participants to not correctly link it to the memorized target, and cause recall errors.

A second experiment re-used the same materials and asked participants to rate the sensicality of a displayed phrase as quickly as possible. In this experiment, the authors measured accuracy and reaction times on the judgments for each second item. Consistent phrases again were rated more accurately than inconsistent phrases (96% accuracy vs. 87%), but reaction times only were reliable in a per-item analysis, where the target was rated more quickly if the prime was consistent. When repeated with homonymic instead of polysemous items, the authors found no significant interaction between consistency and ambiguity type, but consistency again was a reliable factor of judgment accuracy.

In a follow-up series of forced choice experiments, Klein and Murphy (2002) then tested whether participants are more likely to group together polysemous targets invoking different interpretations or unrelated words that fulfil the same conceptual or thematic role. In a first experiment, participants saw a target phrase like *wrapping PAPER* and were given two options: *liberal PAPER* and *smooth CLOTH*. In this example, the first option contains the same target word, but the modifier elicits a different interpretation than the *material* sense in the target phrase. We will call this the "polysemous" option. The other option presents a different target word which matches the taxonomic category of the target *paper*. These "non-polysemous" options were always either matching the taxonomy or the theme of target, with a phrase like *sharp SCISSORS* being an example for a thematic option for target *wrapping PAPER*. Forced to choose between the two options, participants selected the polysemous option in only 20% of the cases, independent of whether the other option was a category or thematic match.

Taken together, Klein and Murphy argue that these experiments indicate that there are no signs of significant processing differences between homonymic and polysemous targets, and that invoking different (polysemous) senses—like different homonymic meanings—requires more processing and leads to more inaccurate judgments than when invoking the same interpretation. In their view, these findings favor a Sense Enumeration approach, as inconsistent polysemous senses seem to be no more accessible after priming and priming an inconsistent sense does not facilitate the interpretation of a target in the same way as a consistent prime does. Still maintaining that polysemes can—and should be—distinguished from homonyms on a theoretical basis, Klein and

7 All test items were indeed shown in the memorization list, so the correct answer for all test items was *yes*.

Murphy (2001) therefore offer the notion that different senses of polysemous words are *related* but not *similar*.

Additional evidence for sense-enumeration approaches comes in the form of an eye-tracking study presented in Foraker and Murphy (2012), where participants read late disambiguation sentences invoking dominant or subordinate interpretations of a target polyseme. Using a subset of the target words from Klein and Murphy (2001, 2002), the authors here found that, much like traditionally shown for homonyms (see e.g., Simpson 1981; Tabossi, Colombo, and Job 1987), the biased introductions lead to significantly longer fixations on the disambiguating region and increased overall reading times when matched with an inconsistent disambiguation, while the neutral context did lead to comparable reading times for dominant interpretations but slower reading times for subordinate senses—indicating that the tested polysemous targets behaved akin to homonymic ones.

The experiments by Klein and Murphy (2001, 2002) and Foraker and Murphy (2012) however are not unchallenged, and have been strongly criticized for their methodology and materials. Taking into account for example the hypothesis of the ambiguity advantage, the function of polysemy isn't considered to be determined by the fact that senses cannot be told apart by participants—but the question is whether they do disambiguate a polysemous expression if the context does not require them to do so. Memorization and in particular forced choice tasks like employed here seem ill-suited to test for this distinction. With respect to the materials, Klepousniotou, Titone, and Romero (2008) for example report a number of homonymic targets previously mislabeled as polysemous by Klein and Murphy. To address the noise introduced by these samples, they repeated the experiments presented in Klein and Murphy (2001) using a set of highly controlled targets classified as metonymic polysemes, metaphoric polysemes, or homonyms based on the semantic overlap between their primary and secondary interpretations.

When in their experiment both the prime and target pair invoked the dominant interpretation, reaction times were significantly faster for ambiguous words with low and moderate sense overlap (i.e., homonyms and metaphoric polysemes), but high-overlap metonymic targets did not show this effect. For subordinate target pairs, reaction times were numerically (but not significantly) faster for metonymic targets than for metaphoric polysemy or homonymy. Investigating response accuracy, the authors found mirrored results: While dominant targets showed no significant word type×context effects, accuracy for low-overlap words was significantly higher for subdominant targets with matching contexts (97%) than for neutral contexts (88%) or conflicting contexts (79%). With these results contradistinctive to the results reported by Klein and Murphy, this study underlines the importance of clear definitions of what types of lexically ambiguous alternations are being investigated, and the impact of selecting appropriate samples to represent these alternations in the experiment materials.

4.2 Studies in Support of One Representation Models

As the most widely accepted model of the mental processing of polysemous words, the One Representation account has been investigated by a wide variety of (psycho-) linguistics and behavioral studies. We will here present the most seminal eye-tracking and brain activation studies focusing on the processing of lexically ambiguous items that each provide partial evidence in support of an underspecified One Representation account of polysemy in an attempt to make their central findings accessible to researchers from other domains.

4.2.1 The Immediate Partial Interpretation Hypothesis. Historically, the (psycho-) linguistics literature produced a number of different, and oftentimes conflicting, principles explaining lexical processing, including models of immediate and delayed semantic interpretation, completely-specified (maximal) and minimal commitments, and a so-called default assignment strategy where a particular option is selected based on frequency or pragmatic plausibility (cf. Frazier and Rayner [1990], and see Frisson [2009] for a comprehensive survey). Frazier and Rayner (1990) then suggested that each of these processing principles could be involved in different aspects or elements of the language comprehension process, and proposed to re-formulate the central goal of language comprehension research to "determine which class of decisions falls under which strategy, and why."

In 1990, Frazier and Rayner presented an eye-tracking study designed to investigate the validity of two general hypotheses labeled *Immediate Complete Interpretation* and *Immediate Partial Interpretation* hypotheses. According to the Immediate Complete Interpretation hypothesis, the language processor "maximizes its immediate semantic commitments by interpreting each phrase fully as the phrase is encountered." While a number of previous studies seemed to agree on their observation that semantic interpretation occurs rapidly (e.g., Crain and Steedman 1985; Marslen-Wilson and Tyler 1980; Just and Carpenter 1980), Frazier and Rayner note that this does not imply that interpretations are necessarily complete. As a result, the Immediate Partial Interpretation hypothesis proposes that "the processor may delay semantic commitments if this does not result in either i) a failure to assign any semantic value whatsoever to a word or major phrase, or ii) the need to maintain multiple incompatible values for a word, phrase or relation." The authors suggest that this *partial specification* for example could occur when commitments have to be made for interpretations involving partially compatible specifications, as when two options share some but not all features (cf. Pinkal's [1985] *Precisification Imperative*).

To test these two hypotheses, the authors presented an eye-tracking experiment involving late disambiguation of homonymic and polysemous expressions. Arguing that the different interpretations of a homonym (see Example 10) are incompatible with one another and therefore cannot be maintained without selection, Frazier and Rayner expected participants to track back when assigning a wrong interpretation to a homonym, while the overlap of the different interpretations of polysemes (see Example 11) allow for a minimal commitment or partial interpretation and therefore does not require any backtracking.

Example 10

1. Being so elegantly designed, the pitcher pleased Mary.
2. Throwing so many curve balls, the pitcher pleased Mary.
3. Of course the pitcher pleased Mary, being so elegantly designed.
4. Of course the pitcher pleased Mary, throwing so many curve balls.

Example 11

1. Lying in the rain, the newspaper was destroyed.

2. Managing advertising so poorly, the newspaper was destroyed.
3. Unfortunately the newspaper was destroyed, lying in the rain.
4. Unfortunately the newspaper was destroyed, managing advertising so poorly.

Frazier and Rayner found that late disambiguation indeed only led to increased reading times for samples containing homonymic targets. Polysemous samples exhibited reading times similar to the unambiguous controls—supporting the Immediate Partial Interpretation hypothesis. However, the authors also found that when a polysemous target was preceded by a disambiguating context, reading times were longer when it instantiated the dis-preferred reading as opposed to the preferred one, which suggests that for polysemes, like for homonyms, readers seem to commit themselves to a particular sense. Based on these observations, Frazier and Rayner suggest that when encountering a polyseme before its disambiguating context, "the shared or overlapping features of the various senses of the target are assigned as the initial semantic value." This allows for minimal semantic commitment and—as opposed to homonyms—will not generate incompatible entailments.

4.2.2 Priming and Sense Dominance. Investigating the processing of ambiguous words through priming, Klepousniotou (2002) hypothesized that processing times should differ between homonyms, where distinct senses need to be selected, and polysemes, where a "basic semantic value" suffices to continue processing. Specifically, Klepousniotou expected that polysemes should be processed faster than homonyms, and that within the polysemous targets, metonymic alternations should show larger priming effects than metaphoric ones due to the additional lexicalization involved in metaphoric alternations. To test these hypotheses, participants were presented with sentences priming either a more frequent (primary) or less frequent (secondary) reading of ambiguous targets. The priming sentence was followed either by a non-word, a target word, or a (non-primed) control item, where controls were words matched for overall corpus frequency or the dominance of the primary reading. Example 12 showcases a selection of prime and target/control pairs representative of the used homonymic, metaphoric, and metonymic materials:

Example 12

Homonymy

Prime 1: All the snow has melted now.
Prime 2: In the mountains, we refilled our canteens.
Target: spring
Frequency control: hotel
Dominance control: bridge

Metaphorical polysemy

Prime 1: My dog is happy.

Prime 2: I went to the back of the air plane.
Target: tail
Frequency control: motel
Dominance control: mouth

Mass/count metonymic polysemy

Prime 1: The hunter killed one.
Prime 2: The chef made a stew.
Target: rabbit
Frequency control: violin
Dominance control: plum

Participants were asked to judge whether the shown word is a real word of English by pressing a designated yes/no key on a keyboard, and reaction times were measured from the onset of the target. As Klepousniotou expected, reaction times were significantly faster for metaphoric and metonymic polysemes than for homonyms. Priming effects were strongest for the metonymic targets, with both target reaction times significantly lower than in the metaphor and homonymy conditions, and control reaction times much longer than for the other ambiguity types. Metaphoric polysemes not only led to significantly lower processing times than homonymic targets, but were also processed significantly faster than their respective controls. Finding that metonymic polysemes provided an even larger underspecification advantage than their metaphoric counterparts, the author suggests that her experiments support generative approaches where polysemous interpretations are constructed from a single, rich lexicon entry based on the contextual requirements.

She later however abandoned this stance in favor of a thin semantics approach when presenting another range of experiments in Klepousniotou et al. (2012). Revisiting previous work using EEG data, these experiments used *unbalanced homonymous* (*pen*), *balanced homonymous* (*panel*), *metaphorically polysemous* (*lip*), and *metonymically polysemous* words (*rabbit*) in a single-word priming delayed lexical decision task. Here, Klepousniotou et al. found that the theoretical distinction between homonymy and polysemy was reflected in the N400 component: Both balanced and unbalanced homonymous words showed priming effects with reduced N400 signals predominantly for dominant readings, while all polysemous primes lead to reduced N400 amplitudes for both readings. Taking the N400 component to reflect lexical activation and semantic processing, the authors concluded that while homonyms are processed by directly selecting their dominant reading, polysemes (both metaphoric and metonymic) facilitate the selection of any of their alternative interpretations.

This line of thinking is supported by other ERP experiments like Rodd, Gaskell, and Marslen-Wilson (2002) and Beretta, Fiorentino, and Poeppel (2005), who showed that words with more than one meaning (i.e., homonyms) were accessed more slowly than words with a single meaning (i.e., they elicited later M350 peak latencies and slower reaction times), and that words with many senses (i.e., productive polysemes) were accessed faster than words with fewer senses. Together, these studies provide important evidence for an underspecified One Representation account of polysemes. These studies however also suggest that their representation cannot be fully underspecified, as even different types of polysemes appear to be processed differently—and therefore need

to be represented in such a way that they can be distinguished and accessed by the processor.

4.2.3 Sense Frequency and Sense Shifting. Structuring his argumentation around polysemes like *book* that allow for a concrete and another, more abstract reading, Frisson (2015) showed through the results of two experiments that sense frequency had no apparent effect on sense switching costs—in contrast to the direction of switching, with especially switches from concrete to abstract interpretations leading to longer fixations on the ambiguous targets.

Arguing that both a traditional Sense Enumeration (SEL) account as well as a relevance theory (RT) approach to polyseme word sense representation would suggest a frequency bias on sense interpretations, Frisson conducted a range of reaction time and eye tracking studies. In a first experiment, Frisson presented participants with two adjective-noun pairs, asking them for binary sensicality judgments but measuring reaction times. In this setup, neither sense frequency nor the order of sense shifting within a presented pair (from abstract to concrete meaning or vice-versa) had an effect on the participants' judgment reaction times. Acknowledging the stark difference between normal reading and this sensicality judgment task, he then continued with a second experiment with full sentence stimuli resembling co-predication structures including both readings in different orderings. In this setup, he found that when a polysemous word was preceded by a neutral context, target region fixation and reading times both were short. When the ambiguous target was preceded by a disambiguating adjective, readers spent more time on the target region—without any notable differences between contexts selecting the primary or subordinate sense interpretation of the target.

SEL and RT inspired models would predict that in a neutral context, a reader assigns the most frequent sense to an ambiguous target. Given that Frisson found no difference in processing time between dominant and subordinate interpretations of polysemous targets, he argues that his experiments support neither of these approaches, but also subscribes to an underspecification account of polysemous sense representation where readers do not immediately select one of the available sense interpretations.

4.2.4 Event-related Potentials for Metonymic Polysemy. In the absence of frequency effects, Frisson did notice a higher cost associated with switching from a concrete reading to an abstract interpretation. This observation is shared by Schumacher (2013), who shows that while some metonymic extensions involve cost-free meaning selection, others "engender processing costs associated with re-conceptualization." In those cases, Schumacher suggests that targets have an "original" meaning and a set of contextually appropriate ones derived from it. Schumacher argues that different types of metonymic alternation should be distinguished: those that involve a fully underspecified representation of alternate senses in the mental lexicon, and those starting from an inherent meaning from which contextualized interpretations are derived.

Looking first at *container/content* alternations, Schumacher notes that there are "relations between ontological types—such as liquids can be contained in physical objects—whose specification is determined by encyclopedic knowledge, and these relations are made available during compositional processing to induce a meaning shift (cf. Copestake and Briscoe 1995; Dölling 1995)." Using German question-answer pair stimuli like those in Example 13 asking for the target with a specific restriction on its interpretation, Schumacher tracked participants' event related potentials (ERP) through an EEG experiment. Analyzing the grand-average ERPs, Schumacher found a more

pronounced positive deflection between 550 and 750ms and between 900 and 1100ms in the critical region of *container-for-content* than in their controls, but no statistically significant difference in the ERPs of *content-for-container* alternations and their controls. These findings were also mirrored in a pre-test as well as in a post-EEG test asking participants to rate the samples' plausibility, which revealed no differences between *content-for-container* samples and their controls, but reliably lower plausibility for *container-for-content* items than their controls.

Example 13

container-for-content (lexically based)
Was hat Heinz hastig getrunken?
Er hat **den Becher** hastig getrunken.
(What did Heinz drink quickly? He quickly drank **the cup**.)
control
Was hat Rolf wie seinen Augapfel gehütet?
Er hat **den Becher** wie seinen Augapfel gehütet.
(What did Rolf guard jealously? He jealously guarded **the cup**.)

content-for-container (based on encyclopedic knowledge)
Was hat Asterix an seinem Gürtel festgeschnallt?
Er hat **den Zaubertrank** an seinem Gürtel festgeschnallt.
(What did Asterix fasten to his belt?
He fastened **the magic potion** to his belt)
control
Was hat Miraculix vor dem Eintreffen der Römer gebraut?
Er hat **den Zaubertrank** vor dem Eintreffen der Römer gebraut.
(What did Getafix brew before the Romans arrived?
He brewed **the magic potion** before the Romans arrived.)

Schumacher proposes that the observed differences between the two alternations' acceptability scores as well as processing demands can be explained by stipulating an asymmetry between *content/container* and *container/content* alternations. She suggests that there is a close, "intrinsic" ontological relation between substances and their respective container, in that liquid substances need to be contained in something to be handled. This tighter relation manifests itself in *container-for-content* interpretation becoming encoded in the qualia structure of the content, so that a (prototypical) container becomes available for reference free of processing costs after a liquid is introduced in discourse. By contrast, the connection between a container and its content appears to be less tight (not lexicalized), so that the availability of the *content-for-container* reading relies instead on 'the application of a general lexical derivation rule.' A similar asymmetry was observed in a second EEG experiment using sample sentences containing adjective-noun pairs with matching or mismatching animacy. Here Schumacher observed an enhanced positivity over posterior electrode sites for mismatched use (e.g., *the wooden turtle*) over the literal use (e.g., *the wooden trunk*) between 550 and 750ms, while the comparison involving animacy-neutral adjectives (e.g., *grey dove* vs. *grey shirt*) registered no differences. With an even more clear distinction between primary and derived

interpretations, these findings were taken as additional support for the assumption that late positivity effects are linked to re-conceptualization, and that therefore *container-for-content* alternations require re-conceptualization, while *content-for-container* readings do not.

4.3 Recap

Schumacher's findings strikingly illustrate the heterogeneity of polysemic sense alternations: If processing differences can be found even for different relations between metonymic conceptualizations, both the investigation of these phenomena as well as its modelling need to reflect their complex and multi-faceted appearance. Behavioral studies therefore aim to focus on very specific aspects, clearly defining the types of alternations under investigation, applying suitable methodology and compiling representative materials in order to contribute insightful and robust evidence. When observing these requirements, the gross of behavioral data seems to clash with seminal Sense Enumeration and One Representation approaches, as it exhibits too many idiosyncrasies and exceptions to generalized rules as to be explained by these principled approaches. And while hybrid representation models with notions of both underspecification and some form of sense hierarchy or sense distance can fit well with the accumulating evidence, the nature of their interplay very much remains an open question that needs further investigation to meaningfully advance the field.

5. Linguistic Tests vs. Large-Scale Annotation Studies

While in linguistics literature most of the initial motivation for distinguishing homonymy from polysemy or classifying different types of polysemous alternations comes from paradigmatic or anecdotal examples, we can also find in the literature a number of linguistic tests for polysemy, starting from the so-called co-predication test. These tests were originally devised to determine identity of interpretation in ambiguous words (Zwicky and Sadock 1975); Norrick (1981) was among the first to propose that co-ordination tests like in Example 14 could be used to test for complex polysemy, that is, the activation of regular or possibly inherent polysemous sense extensions. Given that the co-ordinated structure here is acceptable even though it evokes two different senses of the target *book*, Norrick would consider this test to support a complex sense representation for the target word.

Example 14

The book was interesting$^{\text{INFO}}$ and weighed a ton$^{\text{PHYS}}$

5.1 Co-predication Tests

Co-predication is now usually defined as "a grammatical construction in which two predicates jointly apply to the same argument" (Asher 2011; Gotham 2014) and used to test for (types of) polysemy in nominals (starting with, e.g., Cruse 1986) and can be considered as a test for conflict–in type selection or referential relations (Murphy 2021). One of the most common approaches to create co-predication tests is by *conjunction reduction* (Zwicky and Sadock 1975), where two sentences with different interpretations

of an ambiguous expression are combined into a single sentence by reducing the second one into a conjunctive clause of the first:[8]

Example 15

a. The **city** has 500,000 inhabitants.

b. The **city** outlawed smoking in bars last year.

c. The **city** has 500,000 inhabitants *and* outlawed smoking in bars last year.

Co-predication however is not limited to two senses only; if a word has multiple polysemous extensions, one could in principle generate a co-predication structure containing any or all of them as well (see Example 16, adapted from Ortega-Andrés and Vicente 2019):

Example 16
Brazil is a large$^{\text{PHYS}}$ Portuguese-speaking$^{\text{CULT-LANG}}$ republic$^{\text{INST}}$ that scores low in inequality rankings$^{\text{NATION}}$ but has won the football world cup five times$^{\text{CULT-SPORT}}$

While authors such as Asher (2011) distinguish between logical and accidental polysemy by postulating that logical polysemy passes co-predication tests and accidental polysemy does not, and Ortega-Andrés and Vicente (2019) suggest the use of co-predication tests to tell apart inherent from other types of regular polysemy, linguistic tests in general are heavily context dependent, and can be made to yield inconsistent results by carefully manipulating these contexts (Geeraerts 1993; Antunes and Chaves 2003; Schumacher 2013; Falkum 2015; Murphy 2021). Consider the following examples:[9]

Example 17

a. ? Judy's **dissertation** is thought provoking though yellowed with age.

b. Judy's **dissertation** is still thought provoking though yellowed with age.

a. # They took the **door** off its hinges and walked through it.

b. The **door** was smashed in so often that it had to be bricked up.

a. ? This **book** revolutionised the western world and is full of coffee stains.

b. That **book** is wrong about nearly everything it says about biology and full of coffee stains.

a. # Mary fed and enjoyed the **lamb**.

b. Mary had fed the **lamb** herself and so she couldn't possibly enjoy it very much at dinner.

8 Example from Asher (2011).

9 Examples from Norrick (1981), Cruse (1995), and Antunes and Chaves (2003), respectively. We will use question marks (?) to indicate questionable acceptability or sentence felicity, and hashes (#) to indicate arguably unacceptable or infelicitous structures in our examples.

In each pairing, a slight modification of the predications involved in the co-predication structures, sometimes by simply adding a more descriptive context, can make an infelicitous or at least questionable sentence more acceptable—and vice versa. Dölling (2020) therefore note that "it is apparent that co-predication may not only depend on the kind of pattern connecting word meanings but also on the discourse context and the rhetorical connections between the two predications."

In addition to this observation, evidence has been accumulating that acceptability judgments are not as objective as often assumed in literature. Lau, Clark, and Lappin (2014), for example, collecting crowd-sourced acceptability annotations for textbook examples exhibiting grammatical inconsistencies, found that participants often rated acceptability differently than assumed in the original materials. When given a graded rating scale, grammaticality judgments also more resembled the results of a control study rating a shown character to be "fat" or "thin" rather than a second control rating them "male" or "female," indicating that for many annotators grammatical acceptability seem to lie on a spectrum rather than representing a binary signal[10] (also see, e.g., Keller 2000; Sorace and Keller 2005).

5.2 Corpus Studies

In order to mitigate the effects of subjective judgments, some studies use corpus-based approaches to investigating co-predication acceptability. Ježek and Vieu (2014), for example, suggest that the *variability* of co-predication contexts is the key to distinguishing inherent polysemy from context coercion. As an example, they argue that the *event* sense of *sandwich* in a sentence like "Sam finished the sandwich in one minute" is not an inherent sense interpretation, as the phrase "during the sandwich" has far fewer corpus occurrences than a related expression like "during lunch." In line with this thinking, the authors extracted from an Italian text corpus all occurrences of [V [Det N Adj] patterns that could exhibit a *physical/information* alternation like "He **picked up**$^{\text{PHYS}}$ the **interesting**$^{\text{INFO}}$ book"[11] for a number of selected target nouns. The estimated recall of this procedure was reported at about 6%, and precision varied between 0% and 80% depending on the target noun. Among the collected sentences, Ježek and Vieu (2014) only found a marginal rate of matches where the type restrictions in the V and Adj elements differed: target words *lettera* (letter), *giornale* (newspaper), and *documento* (document) showed the highest ratio of co-predication vs. same sense matches, with about 2% of matches indicating co-predication, and targets *pezzo* (piece), *prodotto* (product), and *fenomeno* (phenomenon) exhibited the lowest ratios (all below 0.3%). Based on these results, the authors concluded that the first three lemmas are more likely to be representatives of polysemy proper than the latter ones.

While this corpus-based approach presents a commendable endeavor towards attempting a large-scale investigation of polysemy, the low recall and precision of the pattern-matching setup prohibited any robust insights. There is potential in this approach, however, as polysemy doesn't require co-predication: collecting instances eliciting one reading or the other can equally well provide evidence for polysemous sense alternations and their relative frequency (see, e.g., Haber 2022), suggesting that corpus-based approaches still might contribute to the understanding of the distribution of polysemous sense in future work.

10 When assuming that gender is a binary construct, that is.
11 Note that word order is different in Italian.

5.3 Large-scale Annotation through Crowdsourcing

Instead of collecting corpus based insights, a growing number of studies have started leveraging crowdsourcing to establish large-scale datasets of layperson judgments to investigate the interpretation of ambiguous expressions beyond anecdotal evidence and singular subjective ratings. Focusing on co-predication, Murphy (2019, 2021), for example, collected annotator judgments on co-predication acceptability in a range of experiments aimed at investigating effects of sense ordering, complexity, and coherence. Among his results, Murphy observed significant effects of sense order and sentence type on acceptability ratings, with for example stimuli invoking a concrete interpretation first and co-predicating an abstract second interpretation rated to be more acceptable than if these interpretations were presented in the inverse order. Based on these findings, Murphy suggests a theory of *Incremental Semantic Complexity*, according to which the language processor overall favors the presentation of input in ascending order of semantic complexity—which becomes explicit in co-predication. The author however also finds that co-predication acceptability depends on a wide range of other factors besides complexity, with samples not normed for frequency or controlled for coherence often failing to achieve significance. He therefore argues that co-predication acceptability should not be interpreted as a surefire sign of identity of sense or inherent polysemy, but instead be seen as a complex signal illuminating some of the underlying mechanics of the language processor.

Large-scale datasets that capture nuanced word sense similarity judgments usually did so for word pairs in isolation, as often these are intended to evaluate static word sense embeddings (also see Taieb, Zesch, and Aouicha 2019). Until recently, the few exceptions to this approach included the Word Similarity in Context dataset by Huang et al. (2012), which contains 241 same-word pairs presented in different contexts, the Word in Context (WiC) dataset by Pilehvar and Camacho-Collados (2019), which contains over 7,000 sentence pairs with an overlapping English word that was annotated based on a binary classification task, and CoSimLex (Armendariz et al. 2020), which contains graded similarity judgments for related words (instead of different interpretations of the same word). Increased interest in the matter then produced a number of similar datasets in parallel: Nair, Srinivasan, and Meylan (2020) conducted an investigation of 32 polysemous and homonymic word types extracted from the Semcor corpus (Miller et al. 1993). In their annotation study, participants arranged contextualized samples in a 2D spatial arrangement task (Goldstone 1994) to produce a measure of interpretation similarity. Their results indicated that different polysemous senses are perceived significantly more similar to one another than homonymic meanings.

A year later, Trott and Bergen (2021) presented RAW-C, a dataset of "Relatedness of Ambiguous Words, in Context." To create RAW-C, 77 participants annotated a total of 112 ambiguous words, each taken to invoke two different polysemous or homonymic interpretations (38 homonyms and 74 polysemes). Using a 5-point Likert scale, annotators here rated the relatedness of an ambiguous target highlighted in a displayed pair of context sentences (see Example 18). The median relatedness for both same-sense homonyms and polysemes was 4, whereas the median relatedness for different-sense homonyms (0) was lower than that for different-sense polysemes (2), which was found to exhibit a much higher variance (see Figure 3).

Example 18

1a. He saw a fruit **bat**.

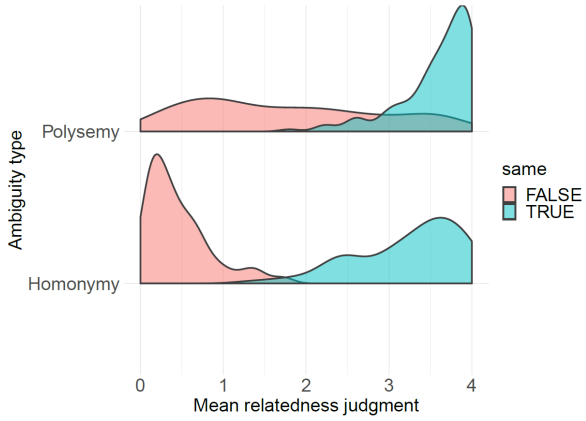

Figure 3
Mean relatedness of RAW-C judgments for sentence pairs containing lexically ambiguous words, plotted by Same Sense (True vs. False) and Ambiguity Type (Homonymy vs. Polysemy). Figure reprinted from Trott and Bergen (2021) with permission.

1b. He saw a furry **bat**.

2a. He saw a wooden **bat**.

2b. He saw a baseball **bat**.

Using a similar methodology, Haber and Poesio collected both explicit judgments of word sense similarity as well as co-predication acceptability judgments for a selection of seminal ambiguous target words, and assembled a large-scale dataset of close to 18,000 similarity and acceptability judgments for custom-made samples invoking different interpretations of ambiguous word forms (Haber and Poesio 2020, 2021; Haber 2022). Participants here were either shown (i) sentence pairs with the same or a different use of an ambiguous target word and asked to rate the target words' similarity on a continuous scale, or (ii) were shown a single sentence produced by combining the same sentences through conjunction reduction into a co-predication structure, and then asked to rate the sentence's acceptability (see Example 19 for an example of sentences used for collecting explicit similarity ratings (a and b), and combined sentences used for assessing co-predication acceptability (c)).

Example 19

a. The **newspaper** fired its editor in chief.

b. The **newspaper** got wet from the rain.

c. The newspaper fired its editor in chief and got wet from the rain.

Among his key findings, Haber reports that annotators made full use of the provided continuous rating scale in judging word sense similarity and co-predication

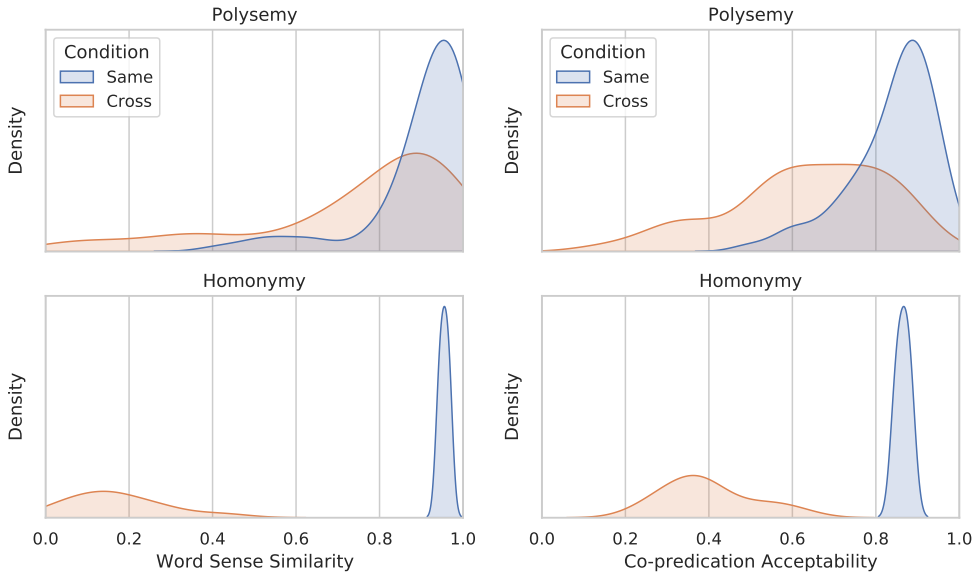

Figure 4
Normalized distributions of explicit word sense similarity ratings (left) and co-predication acceptability ratings (right) given to same-sense (blue) and cross-sense (orange) samples with polysemous (top) and homonymic (bottom) alternations.

acceptability: homonymic cross-sense samples obtained a mean similarity rating of just 0.17, significantly lower than the overall same-sense mean of 0.89. Polysemous cross-sense samples received a mean similarity score of 0.73, which is significantly lower than the same-sense mean, but significantly higher than the homonymy cross-sense mean (see Figure 4, left). The average acceptability rating for co-predication structures invoking the same sense in both predications was calculated at 0.83. For homonymic cross-sense samples, the mean acceptability was 0.41, while the mean acceptability for polysemous alternations was rated at 0.64—significantly higher than the homonym mean, but again significantly lower than the same-sense mean (see Figure 4, right). These results indicate that there seems to be a wide difference in the interpretation of polysemous senses, ranging from perceived identity in sense to an unrelatedness in meaning that is similar to that observed in homonymic items—with intermediate cases clearly present.

In addition to these overall distribution statistics, Haber also investigates potential patterns in the collected sense similarity and co-predication acceptability ratings of words with multiple sense interpretations. The left two columns in Figure 5 display the overall similarity ratings (first column) and co-predication acceptabilities (second column) of shared pairwise sense combinations for target words *newspaper* (top) and *magazine* (physical, information, organization). While *newspaper* and *magazine* appear to display a consistent pattern in their senses' similarity, many other targets in Haber's study do not. This underlines the observation of heterogeneity within different phenomena of polysemy, but also suggests that not every polysemous word allows for its own, idiosyncratic set of sense extensions, but that there is some potential for a classification or grouping of polysemous expressions.

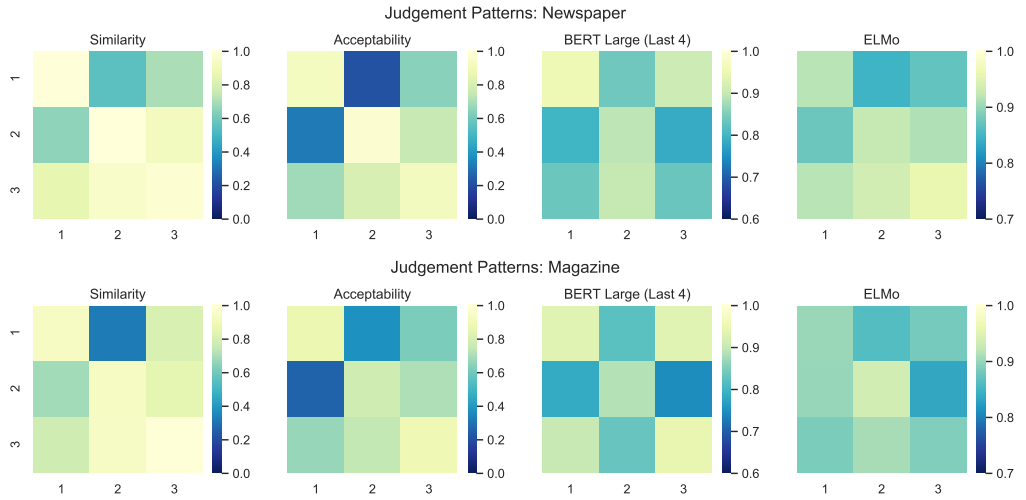

Figure 5
Similarity patterns in the sense similarity ratings for polysemes *newspaper* and *magazine*. Senses: 1-*physical*, 2-*information*, 3-*organization*. Color scales adjusted for computational measures.

At first glance, the distributions of polyseme cross-sense ratings produced by Haber appear to be almost opposite to what Trott and Bergen report in their work: Where in RAW-C the mean relatedness of polysemous cross-sense readings is low but has a significant tail to the right, here the cross-sense similarity of polysemous senses is rated relatively high with a substantial tail to the left. A key difference between these two studies is the presentation of the ambiguous targets, which in RAW-C usually are presented *after* their disambiguating context in compound noun phrases that might lead to an immediate resolution of ambiguity even in the case of polysemes, whereas in the Polyseme Sense Similarity dataset of Haber the targets are presented *before* their disambiguating context, which can allow for under-specification. These differences again highlight the subtleties in the interpretation of polysemous words and the difficulty of quantifying a phenomenon that has its assumed function in the subconscious underspecification of ambiguous terms.

5.3.1 Semantic Change Detection. Besides distinguishing different concurrent meanings of a word, a number of studies also have been investigating lexical semantic change over time. Lexical semantic change is considered to be tightly interlinked with polysemy, with Blank (1997) for example proposing polysemy to be the "synchronic, observable result of lexical semantic change" (Schlechtweg, Schulte im Walde, and Eckmann 2018, also see, e.g., Bamler and Mandt 2017; Hamilton, Leskovec, and Jurafsky 2016). The computational modeling of lexical semantic change however has been limited by the unavailability of diachronic sense references nuanced enough to track the development of specific senses. To address this, Schlechtweg, Schulte im Walde, and Eckmann (2018) recently presented the Diachronic Usage Relatedness (DURel) dataset, a resource consisting of 1,320 pairings of diachronic word uses rated for similarity. Corpus samples were selected manually based on target words either found to indicate signs of innovation through sense narrowing (Paul 2002) or reduction due to homonymy (Osman 1971). Samples were then split into two periods, one with language use recorded from

1750-1800 (EARLIER), and one with samples produced between 1850 and 1900 (LATER). Comparing ratings given to EARLIER and LATER samples, the authors found three types of words: those whose mean relatedness increased, those whose relatedness decreased, and a majority of words for which the mean relatedness remained largely unchanged. The authors however also found that their relatedness measure was prone to confusing lexical semantic change with polysemy, and words with many interpretations could not reliably be investigated through their overall mean relatedness ratings alone. So while these findings indicate a strong link between sense relatedness and the historical origin of senses, methodological limitations again prevent any clear-cut results.

5.4 Recap

The observations of similarity differences between different polysemous senses made in this section are difficult to reconcile with a fully under-specified mental representation of polysemous words: if all of the senses stored in an under-specified entry allow for co-activation and cost-free sense switching, finding evidence of perceived differences in meaning indicates that the mental representations of these senses are likely more structured than assumed by semantically thin One Representation models. On the other hand, some polysemous sense extensions are perceived as identical in meaning. This observation suggests that in the processing of some polysemous senses, no distinction is made between their different interpretations—even though the invoked senses are not identical. While this is in line with the assumptions of One Representation models, it is a challenging finding for Sense Enumeration approaches, which in these cases will struggle to specify the necessary contrast and selection criteria to warrant separate entries for the invoked senses.

The presented data seems to fit in well with recent proposals of a more structured mental representation of polysemous sense, where word sense distance could be an underlying factor in determining the similarity of sense interpretations and their co-activation. But while the increased complexity of these models appears to be better suited to capture the complexity of the phenomena, both the as of yet incomplete picture of the mental processing of polysemy and the still relatively under-developed definition of the hybrid models are central avenues of future research on the issue.

6. Computational Approaches to Lexical Ambiguity and Polysemy

> If one examines the words in a book, one at a time as through an opaque mask with a hole in it one word wide, then it is obviously impossible to determine, one at a time, the meaning of the words [. . .].
> But if one lengthens the slit in the opaque mask, until one can see not only the central word in question but also say N words on either side, then if N is large enough one can unambiguously decide the meaning of the central word [. . .].
> The practical question is: "What minimum value of N will, at least in a tolerable fraction of cases, lead to the correct choice of meaning for the central word?"
> Weaver (1949), p. 20, cited in Ide and Véronis (1998)

Word-sense disambiguation—assigning to a word the most appropriate interpretation in a context—has been identified as one of the main areas of research in NLP since the very early days of research on machine translation and information retrieval (Lesk 1986; Voorhees 1993; Ide and Véronis 1998; Sanderson 2000; Kilgarriff 2001; Mihalcea,

Chklovski, and Kilgarriff 2004; Navigli 2009; Bevilacqua et al. 2021). This section and the next two survey computational research on lexical ambiguity and word sense disambiguation, highlighting in particular how, due especially to the dominance of WordNet (Miller and Charles 1991; Fellbaum and Miller 1998) as a lexical resource, the distinction between homonymy and polysemy has not been as prominent in NLP research as in the linguistic and psychological literature discussed in the previous sections. In this section we briefly summarize early research centered around hand-coded lexical resources, focusing in particular on the sense distinctions adopted in WordNet and the other main lexical resources (for a more thorough review in particular of Word Sense Disambiguation proper, see Navigli 2009 and Bevilacqua et al. 2021). In the next two sections we cover work based on lexical representation of words learned from data—the so-called distributional representations of meaning.

6.1 Senses in Hand-Coded Lexical Resources

Word sense disambiguation was identified as a key NLP task from the very beginning of computational linguistics, motivated by applications such as Machine Translation (MT) (Weaver 1949; Bar-Hillel 1960) and Information Retrieval (IR) (Sparck-Jones 1964). Word sense disambiguation was recognized as one of the most important issues for machine translation as early as Bar-Hillel (1960), and its use became common in MT systems when the first large-scale lexical resources became available. In IR, Sparck-Jones pioneered the use of lexical semantics through her research on IR search methods taking into account synonymy (Sparck-Jones 1964). However, lexical ambiguity only became a concern later. Early work relied on hand-annotated resources such as those discussed in this section—for example, Lesk (1986) used LDOCE, whereas Voorhees (1993) used WordNet (see, e.g., Sanderson 2000 for a survey). Eventually, the field also started to investigate the usefulness of differentiating homonyms from polysemes (Stokoe 2005).

6.1.1 Machine Readable Dictionaries. The very early work on lexical semantics and lexical disambiguation in NLP (Sparck-Jones 1964; Amsler 1980; Lesk 1986) was supported by the creation of the first **machine readable dictionaries** (MRD), such as the Longman Dictionary of Contemporary English (LDOCE) and the Collins COBUILD dictionary (see Wilks, Slator, and Guthrie [1996] and Ide and Véronis [1998] for a discussion of such resources and the early research they supported). The first machine readable dictionaries were just digital versions of the printed dictionaries, but soon publishers started taking advantage of the possibilities offered by the new medium—for instance, senses in the LDOCE were categorized according to *subject codes* providing a shallow hierarchical organization that allows users to carry out topical search in addition to simple alphabetical search. The notion of sense adopted in a particular machine readable dictionary is not uniform, but tends to follow the approach followed in the printed version, so that in some MRDs all senses are listed separately, whereas in others we see a two-level organization such that the first level encodes homonymic distinctions in interpretation, whereas a second level encodes polysemous sense alternations within a single first-level interpretation. Figure 6 illustrates the first type of organization, the lexical entries for *school* in COBUILD; whereas Figure 7 shows an example of the second type of organization using the lexical entries for the wordform *bass* which is both homonymic and polysemous in LDOCE, in which polysemous sense alternations for the same hyponymic interpretation are grouped together.

While they did not introduce novel treatments of lexical ambiguity in general and/or polysemy in particular, MRDs have proven useful tools for research in polysemy

1. variable noun
   A school is a place where children are educated. You usually refer to this place as school when you are talking about the time that children spend there and the activities that they do there.
   ...a boy who was in my class at school.
2. countable noun [with singular or plural verb]
   A school is the pupils or staff at a school.
   Deirdre, the whole school's going to hate you.
3. countable noun
   A privately-run place where a particular skill or subject is taught can be referred to as a school.
   ...a riding school and equestrian centre near Chepstow.
4. variable noun & countable noun
   A university, college, or university department specialising in a particular type of subject can be referred to as a school.
   ...a lecturer in the School of Veterinary Medicine.
5. uncountable noun
   School is used to refer to university or college. [US]
   Moving rapidly through school, he graduated Phi Beta Kappa from the University of Kentucky at age 18.
6. countable noun [with singular or plural verb]
   A particular school of writers, artists, or thinkers is a group of them whose work, opinions, or theories are similar.
   ...the Chicago school of economists. [+ of]
7. countable noun [with singular or plural verb]
   A school of fish or dolphins is a large group of them moving through water together.

Figure 6
Interpretations of *school* in COBUILD.

in corpus linguistics (see, e.g., Stammers 2008) and computational linguistics (e.g. Kilgarriff 1992). And as we will see, the most widely used organization of sense distinctions in NLP, the WordNet organization, is based on the 'flat' approach of COBUILD.

6.1.2 WordNet. From the 1980s onwards, a number of research projects aimed at developing new types of lexical resources based on linguistic and psycholinguistic findings about lexical representation and semantics such as those discussed in previous sections. Among these, the best-known project is WordNet[12] (Miller et al. 1993; Miller 1995; Fellbaum and Miller 1998), a lexical resource whose organization is based on

12 https://wordnet.princeton.edu/.

1. bass[1] noun. Related topics: Music

   (a) [countable] a very low male singing voice, or a man with a voice like this

   (b) [singular] the part of a musical work that is written for a singer with a bass voice

   (c) [uncountable] the lower half of the whole range of musical notes

   (d) [countable] a bass guitar The band features Johnson on bass (=playing the bass guitar).

   (e) [countable] a double bass

   Examples from the corpus:
   The electric bass had a punchy, dynamic range that would become identified with rhythm & blues.
   ...

2. bass[2] adjective (only before noun). Related topics: Music

   (a) a bass instrument or voice produces low notes a bass drum

   Examples from the corpus:
   You need to play the bass notes slightly louder.
   ...

3. bass[3] noun (plural bass) [countable]. Related topics: Fish, Food

   (a) a fish that can be eaten and lives in both rivers and the sea

   Examples from the corpus:
   California, which has no native largemouth bass, imported fast-growing, long-living Florida-strain bass in the 1950s.
   ...

Figure 7
Interpretations of *bass* in LDOCE.

psychological findings about the structure of the mental lexicon (Miller et al. 1990). The two key characteristics of WordNet are that it consists of separate databases for each type of word (nouns, verbs, adjectives, and adverbs) and that these databases are organized "conceptually" rather than alphabetically (i.e., more like in a thesaurus than in a traditional dictionary). Another innovation is that WordNet side-stepped the

issue of how to define the concepts that are used as sense for the lexical items by identifying a concept with the set of synonyms that express that concept, or **synset**. Each synset is taken to represent a lexicalized concept, and different synsets are linked through semantic relations such as hyponymy, antonymy, and meronymy. The original English WordNet, now known as "Princeton" WordNet, grew continuously in number of lexical items and number of synsets until very recently, and versions of WordNet now exist for more than 200 languages, making the WordNet organization the de facto standard approach to sense distinctions for WSD, and WordNet the standard lexical resource at least until the recent development of BabelNet and of context-sensitive word embeddings (see next sections).

The notion of sense in WordNet is, arguably, the best example of what a sense enumeration lexicon would be like. It is very fine-grained, and no distinctions are made between homonyms and polysemes, or between different types of polysemy, as illustrated in Figure 8, where all senses of *bass* are considered at the same level, whether encoding homonymic or polysemous distinctions (Peters and Peters 2000; Mihalcea, Chklovski, and Kilgarriff 2004; Snow et al. 2007). These characteristics have been extensively discussed since WordNet senses became the gold standard for annotation or word sense disambiguation (Ide and Véronis 1998; Kilgarriff 2001; Mihalcea, Chklovski, and Kilgarriff 2004; Navigli 2009). Inter-annotator agreement (IAA) for annotating corpora with WordNet-derived senses ranges between only 67% and 78% (Fellbaum and Miller 1998; Mihalcea, Chklovski, and Kilgarriff 2004; Snyder and Palmer 2004) "depending on factors such as degree of polysemy and inter-relatedness of the senses" (Erk, McCarthy, and Gaylord 2013), which indicates a number of disagreements on sense classifications even among the expert annotators. Similar IAA levels have been found

1. bass[1] (noun): *bass* (the lowest part of the musical range)
2. bass[2] (noun): *bass, bass part* (the lowest part in polyphonic music)
3. bass[3] (noun): *bass, basso* (an adult male singer with the lowest voice)
4. bass[4] (noun): *sea bass, bass* (the lean flesh of a saltwater fish of the family Serranidae)
5. bass[5] (noun): *freshwater bass, bass* (any of various North American freshwater fish with lean flesh (especially of the genus Micropterus))
6. bass[6] (noun): *bass, bass voice, basso* (the lowest adult male singing voice)
7. bass[7] (noun): *bass, bass voice, basso* (the lowest adult male singing voice)
8. bass[8] (noun): *bass* (the member with the lowest range of a family of musical instruments)
9. bass[9] (noun): *bass* (nontechnical name for any of numerous edible marine and freshwater spiny-finned fishes)

Figure 8
(Nominal) interpretations of *bass* in WordNet.

both when using WordNet senses for the reference SemCor corpus providing the contexts used in the lexical resource (Fellbaum and Miller 1998) and creating resources for shared tasks such as SensEval (Mihalcea, Chklovski, and Kilgarriff 2004; Navigli 2009). As a result, extensive studies have been carried out to further investigate sense differentiation in WordNet (Passonneau et al. 2012b; Passonneau and Carpenter 2014) and to automatically cluster related senses into coarser grained senses, effectively creating a two-level organization between homonymic interpretations and polysemous variation within these (Snow et al. 2007; Agirre and Lopez de Lacalle 2003). A number of studies also compared discrete and graded sense assignment for examining word usage in context (Erk, McCarthy, and Gaylord 2009).

6.1.3 BabelNet. Several WordNet projects are still ongoing, both for English and for other languages (Bevilacqua et al. 2021; Navigli et al. 2021). In addition, there is a large project called BabelNet devoted to the creation of a multilingual lexical resource by organizing the WordNets for several languages (Navigli and Ponzetto 2012; Navigli et al. 2021). Being built out of existing WordNets, it adopts the basic organization of lexical entries as WordNet. This means that it follows the same "flat" approach to senses discussed above, and the same approach to senses based on synsets. The (crucial) difference is that the synsets are multilingual, including all the words expressing a particular concept in all languages. These multilingual synsets are created leveraging the second source of information integrated in BabelNet: Wikipedia. BabelNet treats each Wikipedia page as a synset whose lexicalizations are the names of the Wikipedia pages in other languages—for instance, one synset would be $\{school_{EN}, scuola_{IT}, école_{FR}, \ldots\}$. BabelNet currently covers 520 languages.

The multilingual nature of BabelNet has enabled research on polysemy across languages. For instance, Habibi, Hauer, and Kondrak (2021a) developed methods for classifying the synsets in a BabelNet entry into homonyms or polysemes based on the *One homonym per translation* hypothesis: Semantically unrelated senses of a word do not share any translation. Such methods could be used to create an alternative to CoreLex (see below) in research on polysemous senses identification. Rabinovich, Xu, and Stevenson (2020) used BabelNet to investigate polysemy cross-linguistically.

6.1.4 Verbal Lexica: PropBank, VerbNet, and FrameNet. As discussed earlier, polysemy is particularly common with verbs, so it is not surprising that much corpus based and computational work on polysemy has been motivated by evidence about polysemy in verbs (see, e.g., Schulte im Walde 2006; Rumshisky and Batiukova 2008; Peterson and Palmer 2018). However, many properties of verbal behavior that tightly correlate with verbal polysemy, such as argument structure, are not captured in WordNet, so specialized verbal lexica have been developed in NLP to capture these properties. The best known verbal lexica include the repertoire of verbal entries used for the PropBank (Palmer, Gildea, and Kingsbury 2005), VerbNet (Schuler 2005), and FrameNet (Baker, Fillmore, and Lowe 1998). The Proposition Bank,[13] or PropBank (Palmer, Gildea, and Kingsbury 2005) is a project whose original objective was to annotate the Penn Treebank with semantic roles.[14] As a byproduct, a verb lexicon was first defined. VerbNet (Schuler 2005) is a broad coverage verb lexicon designed to be compatible with WordNet, but

13 https://propbank.github.io/.

14 Since then the PropBank annotation scheme has been applied to several other corpora, in particular OntoNotes (Hovy et al. 2006).

incorporating the missing information about the syntactic realization of argument structure. VerbNet defines verb classes based on an extension to the theory of alternations developed by Levin (1993); recent work on the PropBank has been concerned with mapping verb uses in PropBank to these classes. FrameNet (Baker, Fillmore, and Lowe 1998) is a lexical resource based on Fillmore's theory of Frame Semantics (Fillmore 1985) which also provides a repertory of verb senses, used in a number of corpora including, for example, SALSA (Burchardt et al. 2006).

The creation of FrameNet, PropBank, and VerbNet spurred much research on verbal polysemy (see, e.g., Ellsworth et al. 2004; Rumshisky and Batiukova 2008; Erk, McCarthy, and Gaylord 2013; Peterson and Palmer 2018). As in the cases of nominal senses, much research has been concerned with the development of coarser sense distinctions, corresponding to groupings of polysemous senses, in order to achieve better inter-annotator agreement. Due to such coarser annotation scheme, annotator agreement often ranges at and over 90%. Erk, McCarthy, and Gaylord (2013), on the other hand, acknowledge that determining the right level of granularity for the annotation of a WSD task is an important facet of improving model performance, but argue that a theoretically more interesting approach is to explore "novel annotation tasks that allow us to probe the relatedness between dictionary senses in a flexible fashion, and to explore word meaning in context without presupposing hard boundaries between usages."

6.1.5 A Crowdsourced Lexical Resource: RezoJDM. The one substantial computational lexical resource following a more traditional distinction between homonymy and polysemy we are aware of is RezoJDM, a lexical dataset for French created using the game-with-a-purpose *Jeux de Mots* (Lafourcade 2007; Lafourcade and Le Brun 2020). Like WordNet, RezoJDM has a semantic network organization, in which nodes representing senses are related by a number of semantics relations. But unlike WordNet, the senses in WordNet have a hierarchical organization whereby homonymic sense distinctions are represented at the first level, and polysemous distinctions at a second layer. For instance, the lexically ambiguous word *frégate*, which can refer either to a type of bird or a type of ship, is associated with two first layer senses, as illustrated in Figure 9. The polysemous distinction between two related but not identical types of ship both called *frégate* is represented at the second level. RezoJDM is quite substantial—the latest version of the network consists of 2.6 million terms and 180 million relations between them—but we are not aware of research using RezoJDM to study polysemy or word sense disambiguation.

- frégate refine frégate>oiseau
- frégate refine frégate>navire
  - frégate>navire refine frégate>navire>ancient
  - frégate>navire refine frégate>navire>modern

Figure 9
Hierarchical organization of the senses of *frégate* in RezoJDM.

6.1.6 Lexica Based on Pustejovsky's Generative Lexicon Theory. Pustejovsky's Generative Lexicon theory (Pustejovsky 1995) was the foundation for a number of projects devoted to the creation of lexical resources based on the theory.

ACQUILEX was a large European project concerned with the development of a multilingual Lexical Knowledge Base, or LKB, formally modeling the Generative Lexicon as a default inheritance network built on top of typed feature structures (Copestake et al. 1994). The ACQUILEX LKB eventually grew to a respectable size (about 15,000 lexical entries), but, more importantly, the development of this lexicon led to a systematic investigation of the mechanisms proposed in generative lexicon, including the treatment of polysemy (Copestake and Briscoe 1995). We are not, however, aware of its availability.

A second resource based on Generative Lexicon theory is CORELEX (Buitelaar 1998), a (nominal) lexicon in which nouns are associated with their *underspecified semantic type*—a type indicating a particular category of systematic polysemy. CORELEX was derived from WordNet through a multi-step process in which

1. A set of 39 "basic types" are identified, including WordNet's 11 top types (e.g., **entity**, **event**) + 28 of their subtypes (e.g., **artifact**, **communication**);
2. Each synset (sense) of a noun is associated with one of these basic types;
3. A reduced set of basic types is obtained for the word by removing duplicates (e.g., the seven WordNet senses for *book* all express two basic types: **artifact** and **communication**);
4. The nouns with the same set of basic types are grouped into classes (1,648 such classes were found in the version of WordNet used by Buitelaar);
5. All classes with only one member are removed as not displaying *systematic* polysemy (leaving in this example 529 classes);
6. All classes expressing not systematic polysemy, but *homonymy* are also removed;
7. Each of the remaining classes is associated with an underspecified semantic type created from the basic types using a *type constructor* such as •. For instance, *book* is associated with the underspecified semantic type **artifact •communication**.

We are not aware of any large dataset based on the CORELEX sense distinctions, but it has been used as sense dictionary for smaller datasets used in work on polysemous sense disambiguation such as Boleda, Padó, and Utt (2012).

6.2 Word Sense Disambiguation with Hand-Identified Senses

Historically, the objective of Word Sense Disambiguation models has been to select that entry of a provided sense inventory such as the lexica discussed above which best represents the meaning of a word in a given context sentence (Ide and Véronis 1998; Navigli 2009). The most successful approaches to this type of WSD have been first knowledge-based—exploiting the structure of the provided reference knowledge resource to derive classification rules (cf. Lesk 1986; Banerjee and Pedersen 2002; Moro, Raganato, and Navigli 2014) and then supervised, that is, utilizing sense-annotated corpora such as

those discussed below for training a model (cf. Zhong and Ng 2010; Iacobacci, Pilehvar, and Navigli 2016). Much of this research has been based on WordNet senses, without therefore making a distinction between homonymic and polysemous sense disambiguation, with some exceptions (Boleda, Padó, and Utt 2012; Habibi, Hauer, and Kondrak 2021a). But the creation of the datasets used in WSD inspired much empirical research on these sense distinctions, also resulting in attempts to identify which of these senses could be clustered (Erk and McCarthy 2009; Passonneau et al. 2012b). In the following sections we will also discuss new tasks that have been introduced in recent years to properly test the more recent approaches to lexical semantics based on the distributional hypothesis, which do not assume discrete sense distinctions (Apidianaki 2023).

6.2.1 Word Sense Disambiguation Datasets. A number of datasets annotated with word senses exist, covering a number of languages, the great majority of which annotated using (some variety of) WordNet senses (Petrolito and Bond 2014; Bevilacqua et al. 2021). However, only a limited number of these are of a size that is sufficient to train modern NLP models, and these tend to be for English although a few exceptions exist (Petrolito and Bond 2014).

The best known and largest among these WordNet sense-annotated corpora is SEMCOR (Miller et al. 1993), consisting of 352 texts from the English Brown corpus (Kucera and Francis 1967) in which around 200,000 tokens were annotated with WordNet senses. SEMCOR is widely used, but has a number of limitations. For one thing, it only covers about 20% of WordNet synsets. Also, the texts are a bit dated now, so that many senses of words now common are not used. Another English dataset derived from the WordNet project is the WordNet Gloss Corpus, extracted from the definitions or glosses of WordNet's synsets. WordNets exist for more than 60 languages (Petrolito and Bond 2014); among these, those of a size suitable for training include datasets annotated as part of tree banks (corpora annotated at multiple levels) such as ANCORA and TÜBA/DZ (Petrolito and Bond 2014).

In one of these tree bank projects, ONTONOTES (Hovy et al. 2006), sense annotation is done with reference to coarser-grained clusters of senses grouping polysemous sense distinctions together. Special attention to the relatedness of WordNet senses was paid in another project, the annotation of the MASC Word Sense Corpus (Passonneau et al. 2012a), a corpus providing sense annotations for 1,000 occurrences of 100 words. This project is of particular relevance for the study of polysemy because Passonneau and colleagues used multiple annotators so as to be able to carry out statistical analyzes of the distance between these annotations (Passonneau et al. 2012b; Passonneau and Carpenter 2014)

Although only a few of these datasets are suitable for training, the situation is much better for testing, as a number of test sets in multiple languages were annotated in connection with shared tasks such as SENSEVAL and SEMEVAL (Kilgarriff 2001; Mihalcea, Chklovski, and Kilgarriff 2004). Being all based on WordNet, these datasets are not directly usable to study polysemy, but test sets were created, for example, by Boleda, Padó, and Utt (2012) and Habibi, Hauer, and Kondrak (2021a).

6.2.2 Graded Word Sense Assignment. Questioning the principled applicability of discrete sense boundaries (e.g., Tuggy 1993; Kilgarriff 1997; Cruse 2000; Hanks 2000; Kilgarriff 2001), Erk, McCarthy, and Gaylord (2009) make a case for graded word sense assignment. With graded annotations, if during annotation a word usage is assigned different sense interpretations, instead of selecting a single sense label through some form of aggregation, a graded sense assignment is established based on the distribution of

annotations. In a pilot, Erk, McCarthy, and Gaylord collected two types of graded annotations: WSsim (Word Sense Similarity) and Usim (Usage Similarity). In the first experiment, three participants rated the applicability of different WordNet sense interpretations to a given use of a target word in a context sentence. For each sense, annotators were asked to select the degree to which that sense applied to the presented use. Annotations were collected for a total of 11 lemmas presented in a grand total of 430 context sentences, most of which were randomly sampled from the SemCor (Fellbaum and Miller 1998) and SenseEval-3 (Mihalcea, Chklovski, and Kilgarriff 2004) corpora.

In the second experiment, another three participants rated the similarity between usages of the same target word displayed in two context sentences. This experiment covered 34 lemmas selected from the LEXSUB data, including the three lemmas selected from that corpus in the WSsim experiment. For each target word, 10 context sentences were sampled from the LEXSUB data, and each possible combination of context sentences was included in a list of sentence pairs (SPAIR) to be annotated. Annotators here were given the following instructions: "Your task is to rate, for each pair of sentences, how similar in meaning the two boldfaced words are on a five-point scale."

Analyzing the resulting annotations, Erk, McCarthy, and Gaylord found that in the WSsim judgments the extreme labels 1 and 5 were applied significantly more often than the intermediate values, with label 1 (lowest degree of word sense applicability) making up the lion's share of these ratings. In the Usim annotations, the authors found that annotators used intermediate labels more often than in the WSsim setting. Besides confirming that—when given the option—annotators do use graded ratings when judging word sense applicability and word use similarity, Erk, McCarthy, and Gaylord (2013) also found that the collected Usim annotations obey the triangle inequality. In Euclidean space, the "lengths of two sides of a triangle, taken together, must always be greater than the length of the third side." When checking the Usim annotations of sentence triplets with the same target lemma against this principle, they found that over 99% of comparisons did comply with the triangle inequality, indicating that the space spanned by USim annotations indeed is metric and allows for meaningful arithmetic operations.[15]

6.2.3 Using Multi-Lingual Data for Homonymy Detection. Besides approaches like these that leverage mono-lingual information to try and determine the similarity between different sense interpretations, recent studies also investigate the use of multi-lingual data to tell apart homonyms from polysemes. Based on a well-reported tendency for distinct senses of a word to translate differently in other languages (see, e.g., Resnik and Yarowsky 1999), Habibi, Hauer, and Kondrak (2021b), for example, conduct a range of experiments to test their *One Homonym Per Translation* hypothesis (Hauer and Kondrak 2020), which states that semantically unrelated senses of a word do not share any translations.[16]

In their experiments, senses are considered semantically related if their translation into Indonesian or Spanish results in the same word—two languages found to provide good coverage in BabelNet and thought to complement each other. This method correctly identifies homonyms with an F1-score of 78.6 (accuracy = 77.4) compared to an 66.7 F1-score baseline, and correctly identifies homonymic vs. polysemous senses

15 The authors also mention that this observation can be a useful filter criterion when collecting Usim annotations through crowdsourcing, as all annotations violating the triangle principle could be safely discarded.

16 With the exception of *parallel homonymy*.

with an F1-score of 74.7 (accuracy = 68.8) compared to a baseline of 35.9 F1 (accuracy = 44.3).

6.3 Recap

In this section we have seen that most work on lexical resource creation and word sense disambiguation in NLP is based on a theory of the lexicon that does not distinguish between homonymic and polysemous sense distinctions: While lexical resources making such a distinction exist, they are not widely used or have not been used as sense repositories for annotated datasets. We also saw, however, how this flat organization with a number of possibly very fine distinctions all considered equally important proved problematic both at resource creation time, resulting in disagreements between annotators such as those analyzed in Passonneau et al. (2012b), and for system evaluation (Bevilacqua et al. 2021). As a result, much research was carried out on how to use the data to cluster some of these senses. The view of the lexicon with both fine-grained distinctions between the senses grouped into clusters and coarse-grained distinctions between such clusters emerging bottom-up from such research is clearly closely related to the traditional view in which both polysemous and homonymous sense distinctions are present. In the next section, we discuss approaches to lexical semantics in which senses directly emerge from the data, without a manual specification step.

7. Distributional Semantics and Predictive Models

From the beginning of the 1990s, a new approach to sense representation became increasingly dominant in NLP: Distributional semantics. Distributional semantics is based on the Distributional Hypothesis, the assumption that "similarity in meaning results in similarity of linguistic distribution" (Harris 1954; Firth 1957; Erk 2012; Clark 2015; Lenci 2018), and aims to approximate word senses by inferring the relationships between words from large amounts of corpus data. This usually is done by abstracting words and their contexts to vectors in semantic space and measuring the similarity between the vectors of given target expressions.

7.1 Static Word Embeddings

In traditional approaches to distributional semantics, each word is assigned a single vector, resulting in an abstraction over all its contexts of use, and thus theoretically "encompassing all the word senses that are attested in the data" (Arora et al. 2018; Boleda 2020). One way to represent and investigate polysemy under these traditional approaches is by composition (Baroni, Bernardi, and Zamparelli 2014) and, in its simplest form, vector addition. Experimental investigations of composition methods for example include works like Baroni and Zamparelli (2010), Boleda et al. (2013), and Mitchell and Lapata (2010), where phrase similarity predictions derived from the best composition methods reach Spearman correlation scores with participant data of around 0.4.

In 2013, Mikolov et al. (2013a,b) presented a new approach to representing words in vector space using their distributional information: *Word Embeddings*. Observing that much of the complexity in traditional Feed-forward Neural Net Language Models (NNLM) and Recurrent Neural Net Language Models (RNNLM) stems from the non-linearity in their hidden layers, they proposed two new, log-linear approaches to processing large amounts of corpus data while deriving word representations: Continuous Bag-of-Words (CBOW) and Skip-grams. Using a sliding window determining

a target word and a context, with these techniques word embeddings are learned as input to a classifier predicting the probability of the target co-occurring within a given (past and future) context (CBOW), or the probability of a target being surrounded by the selection of context words (Skip-gram), and negative sampling was added to prevent the model from returning perfect probabilities for all proposed combinations (which would yield an initially impressive but ultimately meaningless 100% accuracy). Originally trained on the 6B token Google News corpus with a vocabulary consisting of the 1M most frequent tokens, their approach called *Word2Vec* displayed promising arithmetic features, like a relatively stable relation between the embeddings of country names and their capitals (e.g., vector(*Madrid*) - vector(*Spain*) + vector(*France*) is closer to vector(*Paris*) than to any other word), and the famous observation that vector(*King*) - vector(*Man*) + vector(*Woman*) results in a vector that is very similar to the vector representation of the word *Queen* (Mikolov, Yih, and Zweig 2013).

A year later, Pennington, Socher, and Manning (2014) presented GloVe (**Glo**bal **Ve**ctors), trained on the "non-zero entries of a global word-word co-occurrence matrix," which, according to the authors, provides the "benefit of count data while simultaneously capturing the meaningful linear substructures prevalent in recent log-bilinear prediction-based methods like Word2Vec."[17] Being competitive in embedding quality and difficult to compare in terms of training efficiency,[18] both models have been equal contenders for a range of NLP applications in academia and industry in the following years.

7.1.1 Word Sense Embeddings. A principal feature of any static word embedding approach is that words are represented by a single vector, capturing the context information of all observed uses of that word. While some have suggested that this means that this vector represents all possible meanings and senses a word can elicit, others argue that this approach merely overloads the words' representations and makes their disambiguation impossible. Instead of generating word embeddings, it therefore has been proposed to build *sense embeddings*, that is, deriving one vector for each possible interpretation of a word (for a recent survey see Camacho-Collados and Pilehvar 2018). Sense-specific representations can be generated by encoding only those contexts that invoke a given interpretation of the ambiguous target word, based on the assumption that different uses of a word are reflected by different contexts (see Pedersen and Bruce [1997] and Schütze [1998] for some of the earliest investigations of this approach, and McCarthy et al. [2004], Almuhareb and Poesio [2006], Erk and Padó [2010], and Reisinger and Mooney [2010] for more recent contributions). Notable work in this area for example includes the exploration of combining local and global contexts to produce word representations better suitable to distinguishing different readings (Huang et al. 2012), investigations non-parametrically estimating the number of senses per word type before creating word embeddings for each of these (Neelakantan et al. 2014), and improving WSD by creating individual embeddings for all senses listed in a sense inventory like BabelNet (Iacobacci, Pilehvar, and Navigli 2015).

As mentioned in Section 3.1, Kilgarriff (1997) had already voiced two principled objections to any sense-based approach: the theoretical difficulty in "deciding when two senses are different enough to warrant a new entry, and how to represent the

17 Also see `https://nlp.stanford.edu/projects/glove/`.

18 Training efficiency also is a relatively minor factor in this case as both models only need to be run once to provide their static, pre-trained word embeddings that usually are made available online.

information that is common to multiple different senses." Likewise, Hanks (2000) questions whether different senses actually can be represented as disjoint classes defined by necessary and sufficient conditions.

7.2 Polysemy in Contextualized Language Models

Static word representation approaches are limited by the fact that they might generally capture (different) word meanings, but cannot represent a specific use within a given context at test time. Static word (sense) embeddings therefore are unable to represent specific speaker meaning or the unique communicative function of a word within a given context (see, e.g., Brugman 1988; Hopper 1991; Pedersen and Bruce 1997; Schütze 1998; Paradis 2011; Frermann and Lapata 2016; Melamud, Goldberger, and Dagan 2016; Westera and Boleda 2019). To address this issue, for the past few years the NLP community has been working on a new generation of neural networks to overcome the limitation of static word embeddings, and presented a range of so-called **contextualized language models**. Contextualized language models no longer provide a dictionary of hand-coded or previously calculated word embeddings, but instead can be used to derive a representation of a specific word in a specific context based on large-scale pre-training.

7.2.1 Embeddings from Language Models (ELMo). One of the first (remarkably) successful approaches to context-specific representations was presented by Peters et al. (2018) in the form of ELMo, or Embeddings from Language Models. The underlying model is an unsupervised, bi-directional language model (biLM) pre-trained on next word prediction. Under the hood, it is made up of a character encoding layer, two LSTM (Long Short-Term Memory) layers, and a simple feedforward neural network combined with a softmax function as an output layer. After pre-training, the contextualized embedding for a target word in a given sentence can be calculated by feeding the sample sentence to the model (with parameters frozen) and extracting the different layers' outputs.

For specific downstream tasks, ELMo embeddings can be derived by concatenating hidden state representations from the forward and backward networks, multiplying the concatenated vectors with task-specific weights, and summing the result into a single output vector. Common approaches however also include simply selecting the model's top layer outputs (see, e.g., TagLM and CoVe, Peters et al. 2017; McCann et al. 2017) or the hidden state outputs of one of the inner layers. Peters et al. test their ELMo embeddings on a wide range of NLP applications by replacing previously static inputs with their contextualized encodings. Through this modification alone, they were able to report state-of-the-art results for tasks like question answering on the Stanford Questions Answering Dataset (SQuAD; Rajpurkar et al. 2016), textual entailment on the Stanford Natural Language Inference corpus (SNLI; Bowman et al. 2015), semantic role labeling on the OntoNotes benchmark (Pradhan et al. 2013), and coreference resolution on the CoNLL 2012 shared task (Pradhan et al. 2012).

Investigating the representation of ambiguous words, Peters et al. found that ELMo embeddings can be used to predict the sense of a target word using a simple 1-nearest neighbor approach. Based on the SemCor 3.0 training corpus (Miller et al. 1994), they calculated the average representation for each of the recorded senses and determined the sense of a target word by determining the most similar of these sense embeddings. Using representations from the second LSTM layer only, they reported F1 scores just slightly below the then state-of-the-art approach by Iacobacci, Pilehvar,

and Navigli (2016) on all-words fine-grained WSD (Raganato, Camacho-Collados, and Navigli 2017).

7.2.2 BERT and the Dawn of the Transformers. Since their introduction in 2017, Transformer models (Vaswani et al. 2017) have become ubiquitous in NLP research and application, effectively replacing (bi-)LM approaches due to their substantially better performance on most tasks. Offering a revised model architecture that allows them to efficiently consume immense amounts of unsupervised training data (see Figure 10 for a schematic visualization) combined with a previously unthinkable amount of model parameters, Transformer architectures especially showcase an improved capability of modeling long-range dependencies relevant for downstream tasks that require a deeper 'understanding' of the input text.

One of the most famous Transformer models is BERT, or Bidirectional Encoder Representations from Transformers (Devlin et al. 2019). Fundamentally, the BERT architecture is a stack of Transformer encoder modules consisting of multiple so-called self-attention heads. Each layer of self-attention heads is wrapped with a skip connection, and followed by layer normalization and a fully connected intermediate layer to combine and weigh outputs, turning them into the next layer's inputs. In the BASE model,

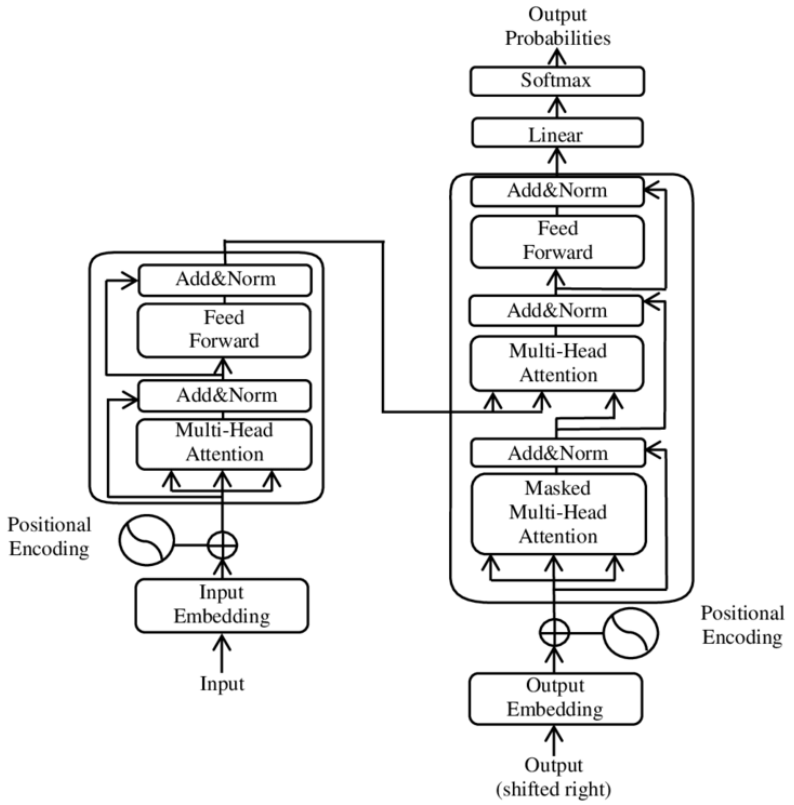

Figure 10
Schematic diagram of the Transformer architecture. Figure by Yuening Jia: DOI:10.1088/1742-6596/1314/1/012186, CC BY-SA 3.

BERT is made up of 12 layers each consisting of 12 self-attention heads, and creates hidden states of 768 dimensions (for a total of 110 million parameters). BERT Large contains 24 layers, each with 16 attention heads and hidden state representations of size 1,024, boasting a total of 340 million parameters.

While technically bi-directional, BERT can in practice be considered non-directional, as it no longer processes language input sequentially, but instead encodes entire input sequences simultaneously. In order to allow for this kind of training paradigm, Devlin et al. presented two novel pre-training strategies to replace the traditional next word prediction: masked language modeling (MLM) and next sentence prediction (NSP). In the MLM task (in linguistics literature often referred to as Cloze task [Taylor 1953]), 15% of tokens in an input sequence are replaced with [MASK]. In contrast to traditional directional language models predicting next or preceding words, the masked tokens allow the model to simultaneously consider preceding and succeeding contexts without "seeing" the target. The NSP training task is especially relevant for tasks such as question answering, where the relationship between different sentences is important.

Following an approach previously labeled Universal Language Model Fine-tuning (or ULM-Fit, see Dai and Le 2015; Howard and Ruder 2018; Radford et al. 2018), once pre-trained, BERT can be fine-tuned to a specific task relatively inexpensively by feeding it sample pairs relevant to the task at hand, like for example question-answer pairs or hypothesis-premise pairs. Fine-tuned on the GLUE (Wang et al. 2018) benchmark suite, for example, the authors report that "BERT BASE and BERT LARGE outperform all systems on all tasks by a substantial margin, obtaining 4.5% and 7.0% respective average accuracy improvement over the prior state of the art."

7.2.3 GPT, T5, and Other Variants. BERT's main competitor in the race for meaningful contextualized word embeddings is the GPT series developed by OpenAI (Radford et al. 2018, 2019; Brown et al. 2020)—recently making headlines with the releases of ChatGPT[19] and GPT-4 (OpenAI 2023). While also based on a Transformer architecture, GPT (Generative Pre-trained Transformer) models take a slightly different approach on processing input text and as a result more closely resemble traditional language models: Under the hood, GPT-2, for example, is an auto-regressive stack of Transformer decoders with between 12 and 48 layers. Its auto-regression prevents the model from using the masked word training objective applied in BERT models, but on the other hand again allows it to process in- and outputs sequentially, which—contrary to BERT—enables GPT models to also be used for text generation. GPT-2's hidden state representations range from 768 dimensions in GPT-2 Small, to 1,600 dimensions in GPT-2 Extra Large, giving the latter a total of 1.5 billion parameters—an order of magnitude more than BERT. This number however already has been put to shame by its successor GPT-3 (Brown et al. 2020), an auto-regressive language model with 175 billion parameters.[20]

Other notable mentions include Google AI's T5 (Text-To-Text Transfer Transformer Raffel et al. 2020), an 11 billion parameter model based on the novel Reformer architecture (a Transformer model designed to handle context windows of up to 1 million words, see Kitaev, Kaiser, and Levskaya [2020]); XLNet, a "generalized auto-regressive

19 https://openai.com/blog/chatgpt.

20 And likely GPT-4, for which the parameter count is undisclosed at the time of writing.

pre-training method that enables learning bi-directional contexts" and therefore "overcomes the limitations of BERT thanks to its auto-regressive formulation" (Yang et al. 2019); as well as BERT variants like RoBERTa (Robustly Optimized BERT Pre-training Approach, Zhuang et al. 2021) and ALBERT (A Lite BERT, Lan et al. 2019), and recent spin-offs like BART (a denoising autoencoder for pretraining sequence-to-sequence models, Lewis et al. 2020b) and MARGE (a Multilingual Autoencoder that Retrieves and Generates, Lewis et al. 2020a).

7.2.4 Probing Contextualized Language Models. While contextualized language models proved to be very successful in a range of downstream NLP tasks, they also provided the community with a dilemma: Due to their black-box architecture, not much is known about *how* these models achieve their remarkable performance levels. This lack of knowledge both affects model accountability and explainability, as well as "limits hypothesis-driven improvement of the architecture" (Rogers, Kovaleva, and Rumshisky 2020). The quest for insights into the inner workings of (among others) contextualized language models therefore spawned a whole new sub-field of NLP research focused on probing and explaining large neural network models.[21] As one of the first such studies, Ethayarajh (2019) investigated the vector spaces spanned by the word encodings produced by contextualized language models like ELMo, BERT, and GPT-2. He found that the word vectors of all of the tested models only occupied a narrow cone within their respective embedding spaces. Ethayarajh even found that GPT-2's last layer was "so extreme that two random words will on average have almost perfect cosine similarity."

Ethayarajh also found that the similarity between vector representations of the same word in different contexts decreased in upper layers, suggesting that "upper layers of contextualized language models produce more context-specific representations." Similar observations were made by Lin, Tan, and Frank (2019), who noticed that lower layers have the most linear word order information. Syntactic information appears to be most prominent in BERT's middle layers (Hewitt and Manning 2019), and the upper layers of BERT indicate the most task specific (Liu et al. 2019) embeddings—with semantic information spread across the entire model (Tenney, Das, and Pavlick 2019).

One of the conclusions that Ethayarajh drew from his observations is that contextualized language models do not seem to encode a fixed number of sense representations derived from the corpus data, but instead create idiosyncratic representations for each word occurrence that combine a range of different information derived from context.

Yenicelik, Schmidt, and Kilcher (2020) support this observation based on a rigorous quantitative analysis of linear separability and cluster organization in embedding vectors produced by BERT. They found that semantics here does not appear to surface as isolated clusters, but that sense embeddings form seamless structures that are tightly coupled with sentiment and syntax. Yenicelik, Schmidt, and Kilcher also found that polysemous words had a high variance in their mean standard deviation (providing support for a hypothesis initially put forward by Miller and Charles 1991) but note that other, non-polysemous words, like stop words, can have equally high variance, and that variance alone therefore is not a surefire sign of multiplicity of sense.

7.2.5 Contextualized Embeddings for Distributional Semantics. Investigating BERT as a distributional semantics model (DSM), Mickus et al. (2020) find that while BERT shows a tendency towards coherence in its contextualized word representation, it does not fully

21 Also see the BlackboxNLP Workshop Series `https://blackboxnlp.github.io/`.

live up to the expectations of a semantic vector space. In particular, they find that the target word position within a context sentence has a noticeable impact on its embedding and disturbs word sense similarity relationships. Treating BERT as a black-box model, they deliberately only use the outputs of the last layer of a vanilla, pre-trained BERT architecture, and analyze the distribution of silhouette scores (Rousseeuw 1987). Polysemous targets overall tend to have a lower cohesion score in this representation, and a lower silhouette score than monosemes—both compatible with what would be expected of a DSM.

Wilson and Marantz (2022) present a two-staged clustering approach to automatically identify the number of senses and meanings associated with an ambiguous target. They sampled 1,000 occurrences of each target word, averaged the BERT Base layer encodings of the target word within each of these samples, and subsequently reduced the resulting target word embeddings to just two dimensions with t-SNE (van der Maaten and Hinton 2008). These two-dimensional target word representations were then used as an input to a first round of density-based clustering with HDBSCAN (Campello, Moulavi, and Sander 2013) to identify sense clusters. Sampling exemplar points from each sense cluster, Wilson and Marantz then applied a second round of clustering to identify potential superstructures, which they suggest to represent different meanings of the target word. They found only a weak correlation between the BERT-derived number of senses and the number of senses reported in WordNet ($p = 0.26$), but report that a qualitative analysis of the clustering presented promising results in distinguishing polysemes from homonyms.

In addition to presenting an analysis of the collected human annotations, the studies by Nair, Srinivasan, and Meylan (2020), Trott and Bergen (2021), and Haber (2022) presented in Section 5.2 also include an investigation of how the similarity of contextualized encodings correlates with the similarity judgments assigned to ambiguous items by the annotators. Nair, Srinivasan, and Meylan for example report a strong correlation between the cosine distance of BERT sense centroids and aggregated relatedness judgments. Trott and Bergen concluded that both ELMo and BERT could differentiate same-sense and different-sense uses of an ambiguous word, but their ability to discriminate between homonymy and polysemy was "marginal at best."

In Haber's (2022) study, a Word2Vec-based baseline approach was found to neither distinguish between homonymic same-sense and cross-sense samples or homonymic and polysemous items (see Figure 11, left). BERT models on the other hand produced clearly distinct distributions for same-sense and cross-sense samples of homonymic items, while for polysemous items the two distributions were largely overlapping (Figure 11, right). This was also reflected in the strong correlation between BERT Large sense similarity predictions and the collected human labels, which was 0.687 (Pearson's r, $p = 1.22E\text{-}24$) for BERT Large, but only 0.206 ($p = 0.008$) for the Word2Vec Baseline.[22]

So, while the representations for different word senses were found to be largely overlapping, the embeddings for unrelated homonymic readings can be clearly told apart. This finding not only holds on an aggregated level, but is often replicated on the word level, where encodings of homonymic uses show significant similarity differences: Figure 12 contains the full similarity map calculated for *magazine*, which—in distinction from the previously shown Figure 5—now also includes the homonymic interpretation as a type of *storage* that is not shared by related target *newspaper*. These plots clearly

22 For comparison, the correlation between the collected explicit sense similarity judgments and co-predication acceptability judgments was 0.698 ($p = 1.09E\text{-}25$).

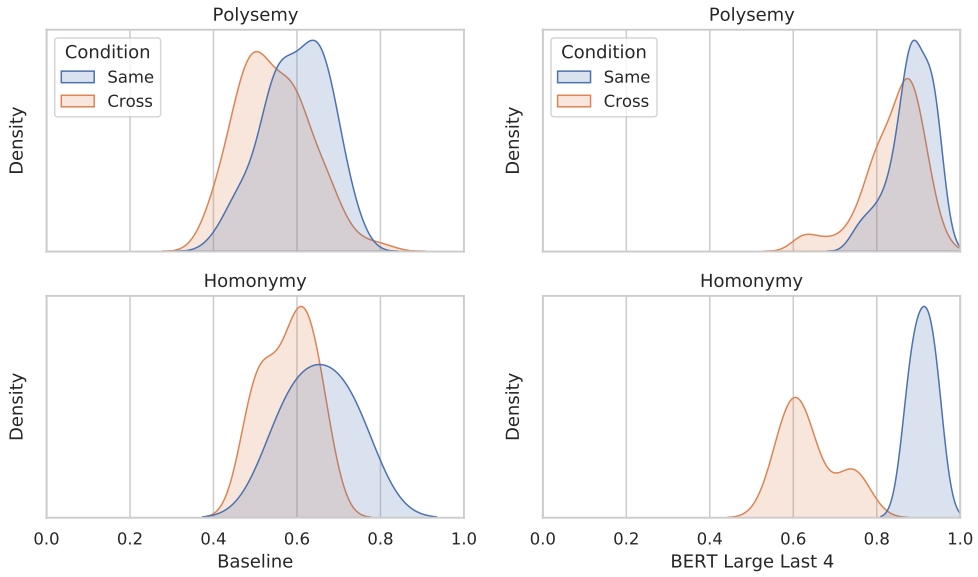

Figure 11
Distributions of embedding similarity scores obtained for same-sense (blue) and cross-sense (orange) samples with polysemous and homonymic alternations. Word2Vec Baseline on the left and BERT Large on the right. BERT Large results for summing over the last four hidden states.

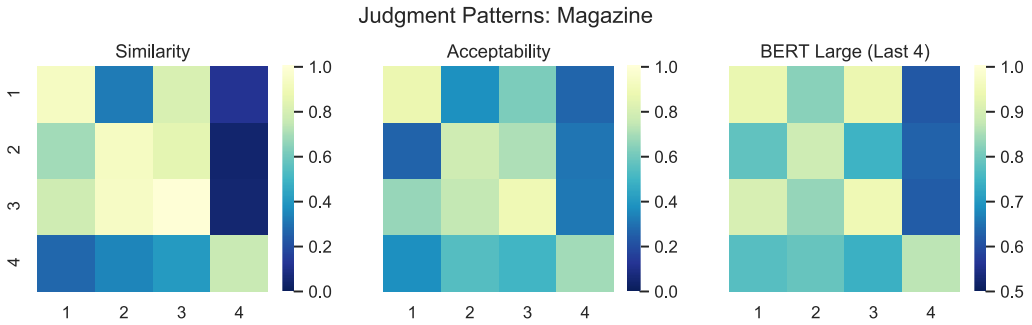

Figure 12
Similarity heat map for four different interpretations of *magazine*, including homonymic alternation 4-*storage type*. Senses: 1-*physical*, 2-*information*, 3-*organization*. Color scale adjusted for BERT.

show the low ratings given to its homonymic cross-sense samples in each of the different measures, including BERT embeddings. Using Agglomerative Hierarchical Clustering using Ward Linkage (Ward 1963) to cluster target word embeddings, Haber (2022) then find that these significant distances mean that homonymic readings can be clearly identified from their embeddings, while telling apart polysemic senses is not consistently possible.

7.2.6 Word Sense Disambiguation Revisited. Now having at their disposal a range of models capable of creating contextualized embeddings, a number of scholars started

investigating approaches to using contextualized language models and their outputs to improve WSD performance (see, e.g., Hadiwinoto, Ng, and Gan 2019; Huang et al. 2019; Song et al. 2021).

After observing that ELMo embeddings of target words in similar contexts can form clusters, and that the representations of words with multiple meanings can split into different groups, roughly representing theses different interpretations (Schuster et al. 2019), Chang and Chen (2019) for example started exploring whether contextualized embeddings are sense-informative enough to derive a sense definition given a (target, context) pair. To this end, they encoded all 79,030 meaning definitions from the Oxford dictionary, and trained a classifier to link contextualized embeddings to these definition embeddings. In the `seen` condition (target word and definitions seen during training), the retrieval precision using BERT Base embeddings ranged from 75 P@1 to 85 P@10, and in the `unseen`, zero-shot condition (target word not seen during training) the retrieval precision using BERT Large embeddings dropped to 3.5 P@1 and 15.5 P@10.

At around the same time, Wiedemann et al. (2019) introduced a simple but effective approach to WSD using a nearest neighbor classification of contextualized embeddings. Applying k-Nearest Neighbors (kNN) clustering with k set to 1, the authors simply classified test set targets based on the nearest train set embedding. While this appeared to work remarkably well for BERT embeddings of train and test samples of the SensEval-2 (Kilgarriff 2001) and SensEval-3 (Mihalcea, Chklovski, and Kilgarriff 2004) WSD tasks, outperforming the last submissions to these tasks, classification performance dropped notably when this approach was applied to the all-word tasks of SemEval2007 Task 7 (Navigli, Litkowski, and Hargraves 2007) and 17 (Pradhan et al. 2007), which both are only composed of test data. Wiedemann et al. (2019) therefore conclude that the nearest neighbor approach suffers specifically from data sparseness and appears to require reference embeddings of practically each sense to work well.

Using a similar approach, Pasini, Scozzafava, and Scarlini (2020) utilize k Means to cluster BERT's contextualized embeddings, but use the number of senses registered in BabelNet (Navigli and Ponzetto 2012) as parameter k for the clustering. While this limits the clustering to detecting only previously recorded interpretations and therefore discretizes the problem, it allows the authors to use the BabelNet definitions to automatically disambiguate the resulting clusters. Comparing with a human-annotated gold standard developed by Bennett et al. (2016), their CluBERT approach out-performs the then state-of-the art model based on the Jensen-Shannon Divergence between the predicted distribution of word use definitions and the gold standard. In a very similar vein, Levine et al. (2020) present SenseBERT, noting that they "focus on a coarse-grained variant of a word's sense, referred to as its WordNet supersense, in order to mitigate [...] brittleness of fine-grained word-sense systems caused by arbitrary sense granularity, blurriness, and general subjectiveness (Kilgarriff 1997; Schneider 2014)." This approach however limits them to identifying 45 different supersense categories, 26 of which for nouns, 15 for verbs, 3 for adjectives, and 1 for adverbs. Instead of clustering, Levine et al. opt for a self-supervised model to predict soft-label category assignments. This approach out-performed a vanilla BERT baseline on a supersense-based variant of the SemEval WSD test sets (standardized by Raganato, Camacho-Collados, and Navigli 2017) and on the WiC task (Pilehvar and Camacho-Collados 2019), but a qualitative analysis of some of the classifications revealed that the model still made consistent categorical mistakes.

Taking a different approach, Amrami and Goldberg (2019) experimented with using target substitutions in order to boost the representation of a given sense interpretation. They however had to concede that their approach did not significantly improve

performance on the word sense induction (WSI) task of SemEval 2013—as neither did their attempt at clustering samples with a dynamic cluster count as opposed to the fixed number in their previous work (Amrami and Goldberg 2018).

Blevins and Zettlemoyer (2020) finally highlight the effect of under-representation in the pre-training of large contextualized language models like BERT, specifically on their ability to perform word sense disambiguation on words that are either rare or completely unseen during training. They present an end-to-end trained bi-encoder built on top of BERT, designed to improve the performance on rare and zero-shot sentences by jointly learning contextualized word embeddings and a gloss encoder from the WSD objective alone. Applied on the English all-words WSD task introduced in Raganato, Camacho-Collados, and Navigli (2017), this model led to an overall absolute improvement of 15.6 F1 over the next-best previous system, with an 31% error reduction on less frequent senses making up for the vast majority of the improvement gain. With vanilla BERT models reaching over 94 F1 on samples labeled with the most frequent sense, Blevins and Zettlemoyer however also question the usefulness of the benchmark and stress the need for better resources to investigate performance on less frequent senses.

To us, this observation again underlines the importance of distinguishing between the purposes of WSD as an applied NLP task and investigating polysemy as a (psycho-) linguistics phenomenon: Although one could argue that the fine-grained senses recorded for example in WordNet can best be seen as polysemous sense extensions rather than homonymic meaning alternations, the fundamental issue with a framework like WordNet from a linguistics perspective is that it cannot account for a measure of distance between senses. According to growing evidence, a homonymic reading of "bank" as a landscape feature should for example be considered to be further removed from polysemous sense extensions such as "building" or "institution" related to the financial reading. This distinguishes homonymic from polysemous readings—and some polysemous sense extensions from each other—while other polysemous senses could be considered identical from a processing perspective. So while WSD is focusing on assigning the "right" sense to a given occurrence of an ambiguous item, to better understand the complexity of polysemy, linguistics research is more occupied with assessing whether an interpretation can be left under-specified, how closely related two senses are, and, as a result, *how* costly assuming an alternative reading will be with respect to the processing cost involved in correcting it.

7.2.7 Semantic Change Revisited. Investigating contextualized language models as a tool to analyze lexical semantic change, Giulianelli, Del Tredici, and Fernández (2020) showed that predicted similarity shifts correlate well with human judgments. Noting the limitations of a single word representation (Hopper 1991; Lau et al. 2012; Frermann and Lapata 2016; Hu, Li, and Liang 2019) and those of fixed word sense representations (Brugman 1988; Kilgarriff 1997; Paradis 2011) in capturing "word meaning, which is continuous in nature and modulated by context to convey ad-hoc interpretations," they suggest the use of contextualized representations to utilize case-by-case context information for a more fine-grained, seamless representation of ad-hoc sense.

As a first step of evaluation, Giulianelli, Del Tredici, and Fernández compare the similarity between BERT's embeddings with human judgments of word sense similarity. Annotators were shown pairs of target word usages within their original context, and asked to rate their similarity using a 4-point scale, ranging from *unrelated* to *identical* (also see Brown 2008; Schlechtweg, Schulte im Walde, and Eckmann 2018). Judgments from five annotators were then averaged to form a usage pair's similarity score and compared to the cosine similarities between the target's BERT embeddings. For 10 out of

the 16 tested targets, the authors determined a significant, positive correlation between human similarity scores and BERT representation similarity.

Encouraged by these results, Giulianelli, Del Tredici, and Fernández then used an unsupervised clustering of BERT embeddings[23] to create a **usage type** partitioning of contextualized representations. A qualitative analysis of the partitionings revealed that "usage types can discriminate between underlying senses of polysemous (and homonymous) words, between literal and figurative usages, and between usages that fulfil different syntactic roles."

7.3 Recap

Through their corpus-based approach to "learning" language, distributional models inherently capture our use of ambiguous expressions, and static word and sense embeddings provided a new way of representing the meaning of a word. Contextualized language models then resulted in a fundamental reshaping of the field with their ability to encode a word within its context—allowing for built-in word sense disambiguation. These models have not only led to significant improvements in a wide range of NLP tasks, but they also present an interesting research tool for the linguistics community: If contextualized language models re-produce the way we process and represent polysemous words, much could be learned from investigating their encoding of specific types and patterns of polysemous alternations. A first collection of studies has been focused on researching the correlation between large language models like BERT and the human language processor, but many open questions remain—and contextualized language models themselves are still being actively developed, with frequent improvements to their pre-training methods, word representation capabilities, or fine-tuning opportunities. We therefore expect that the next few years will yield a variety of new insights on the processing of ambiguous expressions gained from investigating their computational representations.

8. Conclusions

Lexical ambiguity has been extensively studied in lexicography, linguistics, psychology, cognitive neuroscience, and computational linguistics for a number of decades now, and the literature on the topic is vast. But particularly when investigating issues such as the distinction between homonymy and polysemy, surveying the debate and summarizing findings across disciplines is difficult: Each discipline approaches these phenomena with a different focus and has developed its own terminology to discuss them. The abundance of empirical evidence emerging in recent years however requires exactly a survey bringing different areas together to allow for a more efficient, multi-disciplinary approach to future research. In this survey, we have attempted to provide such a cross-disciplinary perspective of the current thinking and evidence on the multi-faceted phenomenon of polysemy. We concentrated in particular on a number of recent approaches to explaining the mental representation of polysemes, which were driven by large-scale empirical data and insights generated from the application of novel contextual language models.

23 In this case k Means with k maximizing the silhouette score (Rousseeuw 1987).

8.1 Recap

By providing a thorough introduction to the terminology and subtypes of ambiguity in Section 2, we hope to give researchers a common language in discussing a complex issue, making a principled distinction between word meanings and word senses, highlighting concepts such as regular polysemy, metonymic and metaphoric polysemy, under-specification and vagueness, and indicating how each of these contribute to the complexity of the phenomenon and the difficulty in providing clear-cut results.

In Section 3, we surveyed the linguistic perspective on polysemy. We sketched the development of two major (families of) approaches, namely, sense enumeration and one representation models, and provided a detailed investigation of the experiments conducted to test them in Section 4. The growing new evidence from these experiments proved to be inconsistent with either of the two main families of approaches, resulting in a range of recent hybrid approaches to explaining the mental representation of polysemes. These approaches hypothesize that some polysemous sense extensions allow for co-activation and cost-free sense shifting, while others should be clearly distinguished. These theories erode the traditional strict dichotomy between polysemy and homonymy, and introduce notions such as polysemous sense similarity or a gradedness in the interpretation of polysemous sense to explain the diverse set of observations made on the processing of polysemes. These new behavioral insights and ideology-defying hybrid models also more and more suggest polysemy to occupy a continuous scale between identity of sense and multiplicity of meaning.

Computational linguistics has been preoccupied with how lexical ambiguity affects natural language processing, information retrieval, and automatic summarization and translation at least since Weaver (1949). We surveyed seminal efforts in the creation of computational resources such as WordNet, ACQUILEX, and BABELNET and their use in applied word sense disambiguation and theoretical research, before moving to approaches based on the distributional hypothesis in Section 7. These models try to infer word meaning or word senses from the use of a given target word in corpus data, usually by generating a vector representation to locate them and investigate the relation between them within a semantic space. In recent years, this field has been dominated by a new generation of contextualized language models capable of encoding a given word within its context. We pointed out that besides resulting in much improved performance for NLP applications, these models provide an interesting new tool for the investigation of polysemy, as they might be used as a proxy for the human processing of ambiguous expressions. Both the research on contextualized language models and their correlation with the human language processor however are still relatively young and are progressing quickly, making their investigation an ongoing hot topic without much consolidated insight as of yet.

8.2 Some Key Take-Home Lessons

Comparing the different disciplines' approaches to polysemy, we hope to have conveyed the difference between the traditional view of lexical interpretation adopted in theoretical linguistics and (much of) psychology, where the representation of polysemous sense is oftentimes assumed to be under-specified and homonymic meanings are clearly separate entries; and the WordNet-inspired view most commonly found in computational linguistics, where all word meanings and senses are represented equally, at the same level of hierarchy, as in the sense enumeration view of the mental lexicon.

We also discussed, however, proposals challenging these dominant views in each of these communities, and pinpointed potential for cross-disciplinary investigations.

In Section 7.2 we discussed a recent series of large-scale, crowd-sourced experiments focused on collecting annotations of (graded) polysemous sense driven by the potential of using this data in combination with contextualized language models to tap a wholly new resource in investigating. These studies—including our own work—have indicated that contextualized language models like BERT manage to replicate to a certain degree the complex patterns found in human annotations of word sense similarity. When moving away in this way from anecdotal evidence and seminal examples, the diverse facets of polysemous sense extension become much more apparent: Even narrowly defined, regular metonymic polysemy presents an indeterminate number of different alternation types—ranging from clearly defined sense alternations to context coercion bordering on vagueness. The similarity patterns of a certain alternation can range from identical across multiple targets to completely idiosyncratic for others. Predication order, prototypicality, and frequency effects become visible.

Distilling the observations, experiments, insights, and hypotheses generated across the different disciplines surveyed here, we suggest that most evidence is pointing at a mental representation of polysemous senses that combines the capacity for under-specification with enough structure to allow for possibly the accommodation of similarity clusters, sense hierarchies, or other forms of co-activation patterns to address the wide range of different behaviors observed through the past few decades of research. These approaches combine the necessity to clearly tell apart some polysemous senses with the ability to leverage under-specification advantages for others.

8.3 The Future

In recent years, the processing of lexically ambiguous expressions has become a focus topic again. Semantic Change Detection, a nascent strand of research powered by the capacity of contextualized language models to track sense similarity through historical corpora, has started to provide a new way of explaining sense similarity through the means of exploring semantic shifts, breathing a new life into decade-old theories. Brain data on language processing becomes clearer and easier to obtain, allowing for more direct insights than behavioral studies. Crowdsourcing provides an inexpensive tool to collect hundreds or even thousands of layperson judgments for building empirical corpora dedicated to complex phenomena, and possibly training contextualized language models to detect new representatives for—or even new types of—polysemous alternations beyond those in the focus of current research. And compared to the legacy of the research surveyed in this paper, contextualized language models themselves—while seemingly omnipresent at the moment—have only been around for the blink of an eye. It is difficult to assess the potential where their development will take us—or whether a completely new approach might be just around the corner again.

We hope that the inter-disciplinary literature, experiments, and insights collected in this survey will help to identify next steps in the investigation of polysemy and provide a solid starting point for future work on this intriguing and multi-faceted phenomenon.

Acknowledgments

The authors would like to thank Gemma Boleda, Andrea Bruera, Derya Çokal, and Diane McCarthy for their invaluable input, and all reviewers of previous versions for their helpful feedback.

The research presented in this survey was funded in part by the Disagreements in

Language Interpretation (DALI) project, ERC grant 695662, with additional support by the Enrichment Scheme of The Alan Turing Institute, and the EPSRC-funded ARCIDUCA project, grant number EP/W001632/1.

References

Agirre, Eneko and Oier Lopez de Lacalle. 2003. Clustering WordNet word senses. In *Proceedings of the Conference on Recent Advances on Natural Language Processing (RANLP)*, pages 121–130. https://doi.org/10.1075/cilt.260.13agi

Almuhareb, Abdulrahman and Massimo Poesio. 2006. MSDA: Wordsense discrimination using context vectors and attributes. In *Proceedings of the 2006 Conference on ECAI 2006: 17th European Conference on Artificial Intelligence*, pages 543–547.

Amrami, Asaf and Yoav Goldberg. 2018. Word sense induction with neural biLM and symmetric patterns. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4860–4867. https://doi.org/10.18653/v1/D18-1523

Amrami, Asaf and Yoav Goldberg. 2019. Towards better substitution-based word sense induction. *ArXiv*, abs/1905.12598.

Amsler, Robert A. 1980. *The structure of the Merriam-Webster pocket dictionary*. Ph.D. thesis, University of Texas at Austin.

Anderson, Richard C. and Andrew Ortony. 1975. On putting apples into bottles—a problem of polysemy. *Cognitive Psychology*, 7(2):167–180. https://doi.org/10.1016/0010-0285(75)90008-0

Antunes, Sandra and Rui Pedro Chaves. 2003. On the licensing conditions of co-predication. In *Proceedings of the 2nd International Workshop on Generative Approaches to the Lexicon*.

Apidianaki, Marianna. 2023. From word types to tokens and back: A survey of approaches to word meaning representation and interpretation. *Computational Linguistics*, 49(2):465–523. https://doi.org/10.1162/coli_a_00474

Apresjan, Juri D. 1974. Regular polysemy. *Linguistics*, 12:5–32. https://doi.org/10.1515/ling.1974.12.142.5

Arapinis, Alexandra and Laure Vieu. 2015. A plea for complex categories in ontologies. *Applied Ontology*, 10(n° 3–4):285–296. https://doi.org/10.3233/AO-150156

Armendariz, Carlos Santos, Matthew Purver, Matej Ulčar, Senja Pollak, Nikola Ljubešić, and Mark Granroth-Wilding. 2020. CoSimLex: A resource for evaluating graded word similarity in context. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5878–5886. https://doi.org/10.18653/v1/2020.semeval-1.3

Arora, Sanjeev, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2018. Linear algebraic structure of word senses, with applications to polysemy. *Transactions of the Association for Computational Linguistics*, 6:483–495. https://doi.org/10.1162/tacl_a_00034

Asher, N. and A. Lascarides. 2003. *Logics of Conversation*. Cambridge University Press.

Asher, Nicholas. 2011. *Lexical Meaning in Context: A Web of Words*. Cambridge University Press. https://doi.org/10.1017/CBO9780511793936

Asher, Nicholas and James Pustejovsky. 2006. A type composition logic for generative lexicon. *Journal of Cognitive Science*, 6(1).

Baker, Collin F., Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of the 36th ACL*, pages 86–90. https://doi.org/10.3115/980845.980860

Bamler, Robert and Stephan Mandt. 2017. Dynamic word embeddings. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 380–389.

Banerjee, Satanjeev and Ted Pedersen. 2002. An adapted Lesk algorithm for word sense disambiguation using wordnet. In *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, pages 136–145. https://doi.org/10.1007/3-540-45715-1_11

Bar-Hillel, Yehoshua. 1960. The present status of automatic translation of languages. *Advances in Computing*, 1:91–163. https://doi.org/10.1016/S0065-2458(08)60607-5

Baroni, Marco, Raffaella Bernardi, and Roberto Zamparelli. 2014. Frege in space: A program for composition distributional semantics. *Linguistic Issues in Language Technology*, 9. https://doi.org/10.33011/lilt.v9i.1321

Baroni, Marco and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural*

Language Processing, EMNLP '10, pages 1183–1193.

Bennett, Andrew, Timothy Baldwin, Jey Han Lau, Diana McCarthy, and Francis Bond. 2016. LexSemTm: A semantic dataset based on all-words unsupervised sense distribution learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1513–1524. https://doi.org/10.18653/v1/P16-1143

Beretta, Alan, Robert Fiorentino, and David Poeppel. 2005. The effects of homonymy and polysemy on lexical access: An MEG study. *Cognitive Brain Research*, 24(1):57–65. https://doi.org/10.1016/j.cogbrainres.2004.12.006, PubMed: 15922158

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 (IJCAI-21)*, pages 4330–4338. https://doi.org/10.24963/ijcai.2021/593

Bierwisch, Manfred and Robert Schreuder. 1992. From concepts to lexical items. *Cognition*, 42(1):23–60. https://doi.org/10.1016/0010-0277(92)90039-K, PubMed: 1582158

Blank, Andreas. 1997. *Prinzipien des lexikalischen Bedeutungswandels am Beispiel der romanischen Sprachen*. Max Niemeyer Verlag. https://doi.org/10.1515/9783110931600

Blevins, Terra and Luke Zettlemoyer. 2020. Moving down the long tail of word sense disambiguation with gloss informed bi-encoders. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1006–1017. https://doi.org/10.18653/v1/2020.acl-main.95

Boleda, Gemma. 2020. Distributional semantics and linguistic theory. *Annual Review of Linguistics*, 6(1):213–234. https://doi.org/10.1146/annurev-linguistics-011619-030303

Boleda, Gemma, Marco Baroni, Louise McNally, et al. 2013. Intensionality was only alleged: On adjective-noun composition in distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013): Long papers*.

Boleda, Gemma, Sebastian Padó, and Jason Utt. 2012. Regular polysemy: A distributional model. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics*, pages 151–160.

Boleda, Gemma, Sabine Schulte im Walde, and Toni Badia. 2012. Modeling regular polysemy: A study on the semantic classification of Catalan adjectives. *Computational Linguistics*, 38(3):575–616. https://doi.org/10.1162/COLI_a_00093

Bowdle, Brian F. and Dedre Gentner. 2005. The career of metaphor. *Psychological Review*, 112(1):193–216. https://doi.org/10.1037/0033-295X.112.1.193, PubMed: 15631593

Bowman, Samuel R., Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642. https://doi.org/10.18653/v1/D15-1075

Breal, Michel. 1897. *Essai de semantique (Science des significations)*, Hachette Paris.

Brochhagen, Thomas and Gemma Boleda. 2022. When do languages use the same word for different meanings? The goldilocks principle in colexification. *Cognition*, 226:105179. https://doi.org/10.1016/j.cognition.2022.105179, PubMed: 35700657

Brown, Susan Windisch. 2008. Choosing sense distinctions for WSD: Psycholinguistic evidence. In *Proceedings of ACL-08: HLT, Short Papers*, pages 249–252. https://doi.org/10.3115/1557690.1557762

Brown, Tom, Benjamin Mann, Nick Ryder, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 33:1877–1901.

Bruera, Andrea, Yuan Tao, Andrew Anderson, Derya Cokal, Janosch Haber, and Massimo Poesio. 2023. Modeling brain representations of words' concreteness in context using GPT-2 and human ratings. *Cognitive Science*. https://doi.org/10.1111/cogs.13388, PubMed: 38103208

Brugman, Claudia. 1988. *The story of over: Polysemy, semantics, and the structure of the lexicon*. New York: Garland.

Buitelaar, Paul. 1998. *CoreLex: Systematic Polysemy and Underspecification*. Ph.D. thesis, Brandeis University.

Burchardt, Aljoscha, Katrin Erk, Anette Frank, Andrea Kowalski, Sebastian Padó, and Manfred Pinkal. 2006. The SALSA corpus: A German corpus resource for lexical semantics. In *Proceedings of the Fifth*

International Conference on Language Resources and Evaluation (LREC'06).
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(1):743–788. https://doi.org/10.1613/jair.1.11259
Campello, Ricardo J. G. B., Davoud Moulavi, and Joerg Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Advances in Knowledge Discovery and Data Mining*, pages 160–172. https://doi.org/10.1007/978-3-642-37456-2_14
Caramazza, Alfonso and Ellen Grober. 1976. Polysemy and the structure of the subjective lexicon. *Georgetown University Roundtable on Languages and Linguistics. Semantics: Theory and Application*, pages 181–206.
Carston, Robyn. 2002. *Thoughts and Utterances: The Pragmatics of Explicit Communication*. Oxford: Blackwell. https://doi.org/10.1002/9780470754603
Carston, Robyn. 2013. Word meaning, what is said, and explicature. In C. Penco and F. Domaneschi, editors, *What is Said and What is Not*. Stanford: CSLI Publications.
Chang, Ting-Yun and Yun-Nung Chen. 2019. What does this word mean? Explaining contextualized embeddings with natural language definition. 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)*, pages 6064–6070. https://doi.org/10.18653/v1/D19-1627
Clark, Stephen. 2015. *Vector Space Models of Lexical Meaning*. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118882139.ch16
Copestake, Ann and Ted Briscoe. 1995. Semi-productive polysemy and sense extension. *Journal of Semantics*, 12:15–67. https://doi.org/10.1093/jos/12.1.15
Copestake, Ann, Antonio Sanfilippo, Ted Briscoe, and Valeria de Paiva. 1994. The ACQUILEX LKB: An introduction. In Ted Briscoe, Ann Copestake, and Valeria de Paiva, editors, *Inheritance, Defaults and the Lexicon*, Studies in Natural Language Processing. Cambridge University Press, chapter 9, pages 148–163. https://doi.org/10.1017/CBO9780511663642.009
Crain, Stephen and Mark Steedman. 1985. *On Not Being Led up the Garden Path: The Use of Context by the Psychological Syntax Processor*. Studies in Natural Language Processing. Cambridge University Press. https://doi.org/10.1017/CBO9780511597855.011
Cruse, D. Alan. 2004. *Lexical 'Facets': Between Monosemy and Polysemy*. Max Niemeyer Verlag.
Cruse, David Allan. 1986. *Lexical Semantics*. Cambridge University Press.
Cruse, David Allan. 1995. Polysemy and related phenomena from a cognitive linguistic viewpoint. In Patrick Saint-Dizier and Evelyn Viegas, editors, *Computational Lexical Semantics*, Studies in Natural Language Processing. Cambridge University Press, pages 33–49. https://doi.org/10.1017/CBO9780511527227.004
Cruse, David Allan. 2000. Aspects of the micro-structure of word meanings. In Yael Ravin and Claudia Leacock, editors, *Polysemy: Theoretical and Computational Approaches*. Oxford University Press, pages 30–51. https://doi.org/10.1093/oso/9780198238423.003.0002
Dai, Andrew M. and Quoc V. Le. 2015. Semi-supervised sequence learning. In *Advances in Neural Information Processing Systems*, 9 pages.
Dautriche, Isabelle. 2015. *Weaving an ambiguous lexicon*. Ph.D. thesis, Université Sorbonne Paris Cité.
Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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 (Long and Short Papers)*, pages 4171–4186.
Durkin, K. and J. Manning. 1989. Polysemy and the subjective lexicon: Semantic relatedness and the salience of intraword senses. *Journal of Psycholinguistic Research*, 18(6):577–612. https://doi.org/10.1007/BF01067161, PubMed: 2632800
Dölling, Johannes. 1995. Ontological domains, semantic sorts and systematic ambiguity. *International Journal of Human-Computer Studies*, 43(5):785–807. https://doi.org/10.1006/ijhc.1995.1074
Dölling, Johannes. 2020. *Systematic Polysemy*. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118788516.sem099

Ellsworth, Michael, Katrin Erk, Paul Kingsbury, and Sebastian Padó. 2004. Propbank, salsa, and framenet: How design determines product. In *Proceedings of LREC*.

Erk, Katrin. 2012. Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compass*, 6(10):635–653. https://doi.org/10.1002/lnco.362

Erk, Katrin and Diana McCarthy. 2009. Graded word sense assignment. In *EMNLP 2009 - Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: A Meeting of SIGDAT, a Special Interest Group of ACL, Held in Conjunction with ACL-IJCNLP 2009*, pages 441–449. https://doi.org/10.3115/1699510.1699568

Erk, Katrin, Diana McCarthy, and Nicholas Gaylord. 2009. Investigations on word senses and word usages. In *ACL-IJCNLP 2009 - Joint Conf. of the 47th Annual Meeting of the Association for Computational Linguistics and 4th Int. Joint Conf. on Natural Language Processing of the AFNLP, Proceedings of the Conf.*, pages 10–18. https://doi.org/10.3115/1687878.1687882

Erk, Katrin, Diana McCarthy, and Nicholas Gaylord. 2013. Measuring word meaning in context. *Computational Linguistics*, 39(3):511–554. https://doi.org/10.1162/COLI_a_00142

Erk, Katrin and Sebastian Padó. 2010. Exemplar-based models for word meaning in context. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 92–97.

Ethayarajh, Kawin. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMO, and GPT-2 embeddings. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65. https://doi.org/10.18653/v1/D19-1006

Falkum, Ingrid Lossius. 2011. *The Semantics and Pragmatics of Polysemy: A Relevance-Theoretic Account*. Ph.D. thesis, UCL (University College London).

Falkum, Ingrid Lossius. 2015. The how and why of polysemy: A pragmatic account. *Lingua*, 157:83–99. https://doi.org/10.1016/j.lingua.2014.11.004

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 and George Miller. 1998. *Building Semantic Concordances*.

Fillmore, Charles J. 1985. Frames and the semantics of understanding. *Quaderni di Semantica*, IV(2).

Firth, John R. 1957. A synopsis of linguistic theory, 1930-1955. *Studies in Linguistic Analysis*.

Fodor, Jerry A. 1998. *Concepts: Where Cognitive Science Went Wrong*. Oxford University Press. https://doi.org/10.1093/0198236360.001.0001

Foraker, Stephani and Gregory L. Murphy. 2012. Polysemy in sentence comprehension: Effects of meaning dominance. *Journal of Memory and Language*, 67(4):407–425. https://doi.org/10.1016/j.jml.2012.07.010, PubMed: 23185103

Frazier, Lyn and Keith Rayner. 1990. Taking on semantic commitments: Processing multiple meanings vs. multiple senses. *Journal of Memory and Language*, 29(2):181–200. https://doi.org/10.1016/0749-596X(90)90071-7

Frermann, Lea and Mirella Lapata. 2016. A Bayesian model of diachronic meaning change. *Transactions of the Association for Computational Linguistics*, 4:31–45. https://doi.org/10.1162/tacl_a_00081

Frisson, Steven. 2009. Semantic underspecification in language processing. *Linguistics and Language Compass*, 3(1):111–127. https://doi.org/10.1111/j.1749-818X.2008.00104.x

Frisson, Steven. 2015. About bound and scary books: The processing of book polysemies. *Lingua*, 157:17–35. https://doi.org/10.1016/j.lingua.2014.07.017

Frisson, Steven and Lyn Frazier. 2005. Carving up word meaning: Portioning and grinding. *Journal of Memory and Language*, 53(2):277–291. https://doi.org/10.1016/j.jml.2005.03.004

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

Gilliver, Peter. 2013. Make, put, run: Writing and rewriting three big verbs in the OED. *Dictionaries: Journal of the Dictionary Society of North America*, 34:10–23. https://doi.org/10.1353/dic.2013.0009

Gillon, B. 1999. *The Lexical Semantics of English Count and Mass Nouns*, volume 10. Springer. https://doi.org/10.1007/978-94-017-0952-1_2

Gillon, Brendan S. 1992. Towards a common semantics for English count and mass nouns. *Linguistics and Philosophy*, 15(6):597–639. https://doi.org/10.1007/BF00628112

Giulianelli, Mario, Marco Del Tredici, and Raquel Fernández. 2020. Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973. https://doi.org/10.18653/v1/2020.acl-main.365

Goldstone, Robert L. 1994. Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General*, 123(2):178. https://doi.org/10.1037/0096-3445.123.2.178, PubMed: 8014612

Gotham, Matthew Graham Haigh. 2014. *Copredication, Quantification and Individuation*. Ph.D. thesis, UCL (University College London).

Haber, Janosch. 2022. *Word Sense Distance and Similarity Patterns in Regular Polysemy*. Ph.D. thesis, Queen Mary University of London.

Haber, Janosch and Massimo Poesio. 2020. Assessing polyseme sense similarity through co-predication acceptability and contextualised embedding distance. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 114–124.

Haber, Janosch and Massimo Poesio. 2021. Patterns of polysemy and homonymy in contextualised language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2663–2676. https://doi.org/10.18653/v1/2021.findings-emnlp.226

Habibi, Amir Ahmad, Bradley Hauer, and Grzegorz Kondrak. 2021a. Homonymy and polysemy detection with multilingual information. In *Proceedings of the 11th Global Wordnet Conference*, pages 26–35.

Habibi, Amir Ahmad, Bradley Hauer, and Grzegorz Kondrak. 2021b. Homonymy and polysemy detection with multilingual information. In *Proceedings of the 11th Global Wordnet Conference*, pages 26–35.

Hadiwinoto, Christian, Hwee Tou Ng, and Wee Chung Gan. 2019. Improved word sense disambiguation using pre-trained contextualized word representations. 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)*, pages 5297–5306. https://doi.org/10.18653/v1/D19-1533

Hamilton, William L., Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501. https://doi.org/10.18653/v1/P16-1141

Hanks, Patrick. 2000. Do word meanings exist? *Computers and the Humanities*, 34(1/2):205–215. https://doi.org/10.1023/A:1002471322828

Harris, Zellig S. 1954. Distributional structure. *Word*, 10:146–162. https://doi.org/10.1080/00437956.1954.11659520

Hauer, Bradley and Grzegorz Kondrak. 2020. One homonym per translation. In the *Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020*, pages 7895–7902. https://doi.org/10.1609/aaai.v34i05.6296

Hewitt, John and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138.

Hobbs, Jerry R., Mark E. Stickel, Douglas E. Appelt, and Paul Martin. 1993. Interpretation as abduction. *Artificial Intelligence*, 63(1):69–142. https://doi.org/10.1016/0004-3702(93)90015-4

Hopper, Paul J. 1991. On some principles of grammaticization. In *Approaches to Grammaticalization*. John Benjamins. https://doi.org/10.1075/tsl.19.1.04hop

Hovy, Eduard, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60. https://doi.org/10.3115/1614049.1614064

Howard, Jeremy and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long*

Papers), pages 328–339. https://doi.org/10.18653/v1/P18-1031

Hu, Renfen, Shen Li, and Shichen Liang. 2019. Diachronic sense modeling with deep contextualized word embeddings: An ecological view. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3899–3908. https://doi.org/10.18653/v1/P19-1379

Huang, Eric, Richard Socher, Christopher Manning, and Andrew Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 873–882.

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)*, pages 3509–3514. https://doi.org/10.18653/v1/D19-1355

Iacobacci, Ignacio, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. SensEmbed: Learning sense embeddings for word and relational similarity. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 95–105. https://doi.org/10.3115/v1/P15-1010

Iacobacci, Ignacio, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Embeddings for word sense disambiguation: An evaluation study. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 897–907. https://doi.org/10.18653/v1/P16-1085

Ide, Nancy and Jean Véronis. 1998. Introduction to the special issue on word sense disambiguation: The state of the art. *Computational Linguistics*, 24(1):1–40.

Jackendoff, Ray. 1989. What is a concept, that a person may grasp it? *Mind & Language*, 4(1–2):68–102. https://doi.org/10.1111/j.1468-0017.1989.tb00243.x

Jackendoff, Ray S. 1992. *Languages of the Mind: Essays on Mental Representation*. The MIT Press. https://doi.org/10.7551/mitpress/4129.001.0001

Jackson, Howard. 2002. *Lexicography: An Introduction*. Taylor and Francis / Routledge.

Ježek, Elisabetta and Laure Vieu. 2014. Distributional analysis of copredication: Towards distinguishing systematic polysemy from coercion. In *First Italian Conference on Computational Linguistics (CLiC-it)*, volume 1, pages 219–223.

Just, Marcel A. and Patricia A. Carpenter. 1980. A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87(4):329–354. https://doi.org/10.1037/0033-295X.87.4.329, PubMed: 7413885

Karjus, Andres, Richard A. Blythe, Simon Kirby, Tianyu Wang, and Kenny Smith. 2021. Conceptual similarity and communicative need shape colexification: An experimental study. *Cognitive Science*, 45(9):e13035. https://doi.org/10.1111/cogs.13035, PubMed: 34491584

Katz, Jerrold J. 1972. *Semantic Theory*. Harper & Row, New York.

Katz, Jerrold J. and Jerry A. Fodor. 1963. The structure of a semantic theory. *Language*, 39(2):170–210. https://doi.org/10.2307/411200

Keller, Frank. 2000. *Gradience in Grammar: Experimental and Computational Aspects of Degrees of Grammaticality*. Ph.D. thesis, University of Edinburgh. College of Science and Engineering. School of Informatics.

Kempson, Ruth M. 1977. *Semantic Theory*. Cambridge University Press.

Kilgarriff, Adam. 1992. *Polysemy*. Ph.D. thesis, University of Sussex.

Kilgarriff, Adam. 1997. I don't believe in word senses. *Computers and the Humanities*, 31(2):91–113. https://doi.org/10.1023/A:1000583911091

Kilgarriff, Adam. 2001. English lexical sample task description. In *Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 17–20.

Kitaev, Nikita, Lukasz Kaiser, and Anselm Levskaya. 2020. Reformer: The efficient transformer. *CoRR*, abs/2001.04451.

Klein, Devorah E. and Gregory L. Murphy. 2001. The representation of polysemous words. *Journal of Memory and Language*, 45(2):259–282. https://doi.org/10.1006/jmla.2001.2779

Klein, Devorah E. and Gregory L. Murphy. 2002. Paper has been my ruin: Conceptual relations of polysemous senses. *Journal of Memory and Language*, 47(4):548–570. https://doi.org/10.1016/S0749-596X(02)00020-7

Klepousniotou, Ekaterini. 2002. The Processing of lexical ambiguity: Homonymy and polysemy in the mental lexicon. *Brain and Language*, 81(1–3): 205–223. https://doi.org/10.1006/brln.2001.2518, PubMed: 12081393

Klepousniotou, Ekaterini, G. Bruce Pike, Karsten Steinhauer, and Vincent Gracco. 2012. Not all ambiguous words are created equal: An EEG investigation of homonymy and polysemy. *Brain and Language*, 123(1):11–21. https://doi.org/10.1016/j.bandl.2012.06.007, PubMed: 22819308

Klepousniotou, Ekaterini, Debra Titone, and Carolina Romero. 2008. Making sense of word senses: The comprehension of polysemy depends on sense overlap. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(6):1534–1543. https://doi.org/10.1037/a0013012, PubMed: 18980412

Kucera, Henry and W. Nelson Francis. 1967. *Computational Analysis of Present-day American English*. Brown University Press.

Lafourcade, Mathieu. 2007. Making people play for lexical acquisition with the JeuxDeMots prototype. In *Proceedings of the 7th Symposium on Natural Language Processing (SNLP)*.

Lafourcade, Mathieu and Nathalie Le Brun. 2020. Game design evaluation of GWAPs for collecting word associations. In *Proceedings of the LREC Workshop Games and Natural Language Processing*, pages 26–33.

Lan, Zhenzhong, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A lite BERT for self-supervised learning of language representations. *CoRR*, abs/1909.11942.

Lau, Jey Han, Alexander Clark, and Shalom Lappin. 2014. Measuring gradience in speakers' grammaticality judgements. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society (CogSci 2014)*.

Lau, Jey Han, Paul Cook, Diana McCarthy, David Newman, and Timothy Baldwin. 2012. Word sense induction for novel sense detection. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 591–601.

Laurence, Stephen and Eric Margolis. 1999. Concepts and cognitive science. In Eric Margolis and Stephen Laurence, editors, *Concepts: Core Readings*. MIT Press, pages 3–81.

Lenci, Alessandro. 2018. Distributional models of word meaning. *Annual Review of Linguistics*, 4(1):151–171. https://doi.org/10.1146/annurev-linguistics-030514-125254

Lesk, Michael. 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation*, pages 24–26. https://doi.org/10.1145/318723.318728

Levin, Beth. 1993. *English Verb Classes and Alternations*. University of Chicago Press.

Levine, Yoav, Barak Lenz, Or Dagan, Ori Ram, Dan Padnos, Or Sharir, Shai Shalev-Shwartz, Amnon Shashua, and Yoav Shoham. 2020. SenseBERT: Driving some sense into BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4656–4667. https://doi.org/10.18653/v1/2020.acl-main.423

Lewis, Mike, Marjan Ghazvininejad, Gargi Ghosh, Armen Aghajanyan, Sida Wang, and Luke Zettlemoyer. 2020a. Pre-training via paraphrasing. In *Advances in Neural Information Processing Systems*, pages 18470–18481.

Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020b. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880. https://doi.org/10.18653/v1/2020.acl-main.703

Lin, Yongjie, Yi Chern Tan, and Robert Frank. 2019. Open sesame: Getting inside BERT's linguistic knowledge. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 241–253. https://doi.org/10.18653/v1/W19-4825

Liu, Nelson F., Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. Linguistic knowledge and transferability of contextual representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1073–1094. https://doi.org/10.18653/v1/N19-1112

Löhr, Guido. 2021. Does polysemy support radical contextualism? On the relation between minimalism, contextualism and polysemy. *Inquiry*, 67(1):68–92. https://doi.org/10.1080/0020174X.2020.1868329

Lyons, John. 1977. *Semantics*, volume 1. Cambridge University Press. https://doi.org/10.1017/CBO9781139165693

MacGregor, Lucy J., Jennifer Bouwsema, and Ekaterini Klepousniotou. 2015. Sustained meaning activation for polysemous but not homonymous words: Evidence from EEG. *Neuropsychologia*, 68:126–138. https://doi.org/10.1016/j.neuropsychologia.2015.01.008, PubMed: 25576909

Marslen-Wilson, William and Lorraine Komisarjevsky Tyler. 1980. The temporal structure of spoken language understanding. *Cognition*, 8(1):1–71. https://doi.org/10.1016/0010-0277(80)90015-3, PubMed: 7363578

McCann, Bryan, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6297–6308.

McCarthy, Diana, Marianna Apidianaki, and Katrin Erk. 2016. Word Sense Clustering and Clusterability. *Computational Linguistics*, 42(2):245–275. https://doi.org/10.1162/COLI_a_00247

McCarthy, Diana, Rob Koeling, Julie Weeds, and John Carroll. 2004. Finding predominant word senses in untagged text. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, ACL '04, pages 279–es. https://doi.org/10.3115/1218955.1218991

Melamud, Oren, Jacob Goldberger, and Ido Dagan. 2016. context2vec: Learning generic context embedding with bidirectional LSTM. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 51–61. https://doi.org/10.18653/v1/K16-1006

Mickus, Timothee, Denis Paperno, Mathieu Constant, and Kees van Deemter. 2020. What do you mean, BERT? In *Proceedings of the Society for Computation in Linguistics 2020*, pages 279–290.

Mihalcea, Rada, Timothy Chklovski, and Adam Kilgarriff. 2004. The senseval-3 English lexical sample task. In *Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 25–28.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, pages 3111–3119.

Mikolov, Tomas, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751.

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., Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Five papers about WordNet. Unpublished Manuscript.

Miller, George A. and Walter G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28. https://doi.org/10.1080/01690969108406936

Miller, George A., Martin Chodorow, Shari Landes, Claudia Leacock, and Robert G. Thomas. 1994. Using a semantic concordance for sense identification. In *Human Language Technology: Proceedings*. https://doi.org/10.3115/1075812.1075866

Miller, George A. and Philip N. Johnson-Laird. 1976. *Language and Perception*. Harvard University Press. https://doi.org/10.4159/harvard.9780674421288

Miller, George A., Claudia Leacock, Randee Tengi, and Ross T. Bunker. 1993. A semantic concordance. In *Proceedings of the Workshop on Human Language Technology*, pages 303–308. https://doi.org/10.3115/1075671.1075742

Mitchell, Jeff and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive Science*, 34(8):1388–1429. https://doi.org/10.1111/j.1551-6709.2010.01106.x, PubMed: 21564253

Moro, Andrea, Alessandro Raganato, and Roberto Navigli. 2014. Entity linking meets word sense disambiguation: A unified approach. *Transactions of the Association for Computational Linguistics*, 2:231–244. https://doi.org/10.1162/tacl_a_00179

Murphy, Elliot. 2019. Acceptability properties of abstract senses in copredication. *Perspectives on Abstract Concepts: Cognition, Language and Communication*, 65:145. https://doi.org/10.1075/hcp.65.08mur

Murphy, Elliot. 2021. *Linguistic Representation and Processing of Copredication*. Ph.D. thesis, UCL (University College London). https://doi.org/10.31234/osf.io/yubkz

Nair, Sathvik, Mahesh Srinivasan, and Stephan Meylan. 2020. Contextualized word embeddings encode aspects of human-like word sense knowledge. In *Proceedings of the Workshop on the Cognitive Aspects of the Lexicon*, pages 129–141.

Navigli, Roberto. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys*, 41(2):1–69. https://doi.org/10.1145/1459352.1459355

Navigli, Roberto, Michele Bevilacqua, Simone Conia, Dario Montagnini, and Francesco Cecconi. 2021. Ten years of BabelNet: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21), Survey Track*, pages 4559–4567. https://doi.org/10.24963/ijcai.2021/620

Navigli, Roberto, Kenneth C. Litkowski, and Orin Hargraves. 2007. SemEval-2007 task 07: Coarse-grained English all-words task. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 30–35. https://doi.org/10.3115/1621474.1621480

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

Neelakantan, Arvind, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. 2014. Efficient non-parametric estimation of multiple embeddings per word in vector space. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1059–1069. https://doi.org/10.3115/v1/D14-1113

Nerlich, Brigitte and David D. Clarke. 2003. *Polysemy and Flexibility: Introduction and Overview*. De Gruyter Mouton. https://doi.org/10.1515/9783110895698.3

Norrick, Neal R. 1981. *Semiotic Principles in Semantic Theory*. John Benjamins. https://doi.org/10.1075/cilt.20

Nunberg, Geoffrey. 1979. The non-uniqueness of semantic solutions: Polysemy. *Linguistics and Philosophy*, 3(2):143–184. https://doi.org/10.1007/BF00126509

Nunberg, Geoffrey. 1995. Transfers of meaning. *Journal of Semantics*, 12(2):109–132. https://doi.org/10.1093/jos/12.2.109

OpenAI. 2023. GPT-4 technical report.

Ortega-Andrés, Marina. 2021. *Interpretation of Copredicative Sentences: A Rich Underspecification Account of Polysemy*. Springer International Publishing. https://doi.org/10.1007/978-3-030-56437-7_9

Ortega-Andrés, Marina and Agustín Vicente. 2019. Polysemy and co-predication. *Glossa*, 4(1). https://doi.org/10.5334/gjgl.564

Osman, Nabil. 1971. *Kleines Lexikon untergegangener Wörter: Wortuntergang seit dem Ende des 18. Jahrhunderts*, volume 487. Beck.

Ostler, Nicholas and B. T. S. Atkins. 1991. Predictable meaning shift: Some linguistic properties of lexical implication rules. In *Lexical Semantics and Knowledge Representation*, pages 87–100. https://doi.org/10.1007/3-540-55801-2_29

Palmer, Martha, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: A corpus annotated with semantic roles. *Computational Linguistics*, 31(1):71–106. https://doi.org/10.1162/0891201053630264

Paradis, Carita. 2004. Where does metonymy stop? Senses, facets, and active zones. *Metaphor and Symbol*, 19(4):245–264. https://doi.org/10.1207/s15327868ms1904_1

Paradis, Carita. 2011. Metonymization: A key mechanism in semantic change. In *Defining Metonymy in Cognitive Linguistics*. John Benjamins, pages 61–88. https://doi.org/10.1075/hcp.28.04par

Pasini, Tommaso, Federico Scozzafava, and Bianca Scarlini. 2020. CluBERT: A cluster-based approach for learning sense distributions in multiple languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4008–4018. https://doi.org/10.18653/v1/2020.acl-main.369

Passonneau, Rebecca J., Collin Baker, Christiane Fellbaum, and Nancy Ide. 2012a. The MASC word sense sentence corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*.

Passonneau, Rebecca J., Vikas Bhardwaj, Ansaf Salleb-Aouissi, and Nancy Ide. 2012b. Multiplicity and word sense: Evaluating and learning from multiply labeled word sense annotations. *Language Resources and Evaluation*, 46(2):219–252. https://doi.org/10.1007/s10579-012-9188-x

Passonneau, Rebecca J. and Bob Carpenter. 2014. The benefits of a model of annotation. *Transactions of the ACL*, 2:311–326. https://doi.org/10.1162/tacl_a_00185

Paul, Hermann. 2002. *Deutsches Wörterbuch*. Max Niemeyer Verlag. https://doi.org/10.1515/9783110929799

Pedersen, Ted and Rebecca Bruce. 1997. Distinguishing word senses in untagged text. In *Second Conference on Empirical Methods in Natural Language Processing*.

Pelletier, F. Jeffry. 1975. Non-singular reference: Some preliminaries. *Philosophia*, 5(4):451–465. https://doi.org/10.1007/BF02379268

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pages 1532–1543. https://doi.org/10.3115/v1/D14-1162

Peters, Matthew E., Waleed Ammar, Chandra Bhagavatula, and Russell Power. 2017. Semi-supervised sequence tagging with bidirectional language models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1756–1765. https://doi.org/10.18653/v1/P17-1161

Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237. https://doi.org/10.18653/v1/N18-1202

Peters, Wim and Ivonne Peters. 2000. Lexicalised systematic polysemy in WordNet. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC)*.

Peterson, Daniel and Martha Palmer. 2018. Bayesian verb sense clustering. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI)*, pages 5398–5405. https://doi.org/10.1609/aaai.v32i1.12023

Petrolito, Tommaso and Francis Bond. 2014. A survey of WordNet annotated corpora. In *Proceedings of the Seventh Global WordNet Conference*, pages 236–245.

Pietroski, Paul M. 2005. Meaning before truth. In Gerhard Preyer and Georg Peter, editors, *Contextualism in Philosophy: Knowledge, Meaning, and Truth*. Oxford University Press. https://doi.org/10.1093/oso/9780199267408.003.0010

Pilehvar, Mohammad Taher and Jose Camacho-Collados. 2019. WiC: The word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273.

Pinkal, Manfred. 1985. *Logik Und Lexikon: Die Semantik des Unbestimmten*. De Gruyter. https://doi.org/10.1515/9783110849523

Pinkal, Manfred. 1995. *Logic and Lexicon. The Semantics of the Indefinite*. Kluwer Academic Publishers. https://doi.org/10.1007/978-94-015-8445-6

Poesio, Massimo. 2020. *Ambiguity*. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118788516.sem098

Pradhan, Sameer, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards robust linguistic analysis using OntoNotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152.

Pradhan, Sameer, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40.

Pradhan, Sameer S., Eduard Hovy, Mitch Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2007. Ontonotes: A unified relational semantic representation. *International Journal of Semantic Computing*,

1(04):405–419. https://doi.org/10.1142/S1793351X07000251
Pustejovsky, James. 1993. Type coercion and lexical selection. In James Pustejovsky, editor, *Semantics and the Lexicon*. Springer, pages 73–94. https://doi.org/10.1007/978-94-011-1972-6_6
Pustejovsky, James. 1995. *The Generative Lexicon*. MIT Press.
Pylkkänen, Liina, Rodolfo Llinás, and Gregory L. Murphy. 2006. The representation of polysemy: MEG evidence. *Journal of Cognitive Neuroscience*, 18(1):97–109. https://doi.org/10.1162/089892906775250003, PubMed: 16417686
Rabinovich, Ella, Yang Xu, and Suzanne Stevenson. 2020. The typology of polysemy: A multilingual distributional framework. In *Proceedings of the Annual Conference of the Cognitive Science Society*, pages 3370–3376.
Radford, Alec, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning. *OpenAI Blog*.
Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*.
Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.
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, Long Papers*, pages 99–110. https://doi.org/10.18653/v1/E17-1010
Rajpurkar, Pranav, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392. https://doi.org/10.18653/v1/D16-1264
Ravin, Yael and Claudia Leacock. 2000. *Polysemy: An Overview*. Oxford University Press. https://doi.org/10.1093/oso/9780198238423.003.0001
Recanati, Francois. 1998. Truth-conditional pragmatics. In Asa Kasher, editor, *Pragmatics: Critical Concepts*. Routledge, pages 509–511.
Reisinger, Joseph and Raymond J. Mooney. 2010. Multi-prototype vector-space models of word meaning. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 109–117.
Resnik, Philip and David Yarowsky. 1999. Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation. *Natural Language Engineering*, 5(2):113–133. https://doi.org/10.1017/S1351324999002211
Rodd, Jennifer, Gareth Gaskell, and William Marslen-Wilson. 2002. Making sense of semantic ambiguity: Semantic competition in lexical access. *Journal of Memory and Language*, 46(2):245 – 266. https://doi.org/10.1006/jmla.2001.2810
Rodd, Jennifer M., M. Gareth Gaskell, and William D. Marslen-Wilson. 2004. Modelling the effects of semantic ambiguity in word recognition. *Cognitive Science*, 28(1):89–104. https://doi.org/10.1207/s15516709cog2801_4
Rogers, Anna, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866. https://doi.org/10.1162/tacl_a_00349
Rousseeuw, Peter J. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65. https://doi.org/10.1016/0377-0427(87)90125-7
Rumshisky, Anna and Olga Batiukova. 2008. Polysemy in verbs: Systematic relations between senses and their effect on annotation. In *Proceedings of the ACL Workshop on Human Judgments in Computational Linguistics*, pages 33–41. https://doi.org/10.3115/1611628.1611634
Sanderson, Mark. 2000. Retrieving with good sense. *Information Retrieval*, 2(1):49–69. https://doi.org/10.1023/A:1009933700147
Sandra, Dominiek. 1998. What linguists can and can't tell you about the human mind: A reply to Croft. *Cognitive Linguistics*, 9(4):361–378. https://doi.org/10.1515/cogl.1998.9.4.361

Schlechtweg, Dominik, Sabine Schulte im Walde, and Stefanie Eckmann. 2018. Diachronic usage relatedness (DURel): A framework for the annotation of lexical semantic change. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 169–174. https://doi.org/10.18653/v1/N18-2027

Schneider, Nathan. 2014. *Lexical Semantic Analysis in Natural Language*. Ph.D. thesis, Carnegie Mellon University.

Schuler, Karin Kipper. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph.D. thesis, University of Pennsylvania, Computer and Information Science.

Schulte im Walde, Sabine. 2006. Experiments on the automatic induction of German semantic verb classes. *Computational Linguistics*, 32(2):159–194. https://doi.org/10.1162/coli.2006.32.2.159

Schumacher, Petra. 2013. When combinatorial processing results in reconceptualization: Toward a new approach of compositionality. *Frontiers in Psychology*, 4:677. https://doi.org/10.3389/fpsyg.2013.00677

Schuster, Tal, Ori Ram, Regina Barzilay, and Amir Globerson. 2019. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1599–1613. https://doi.org/10.18653/v1/N19-1162

Schütze, Hinrich. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

Simpson, Greg B. 1981. Meaning dominance and semantic context in the processing of lexical ambiguity. *Journal of Verbal Learning and Verbal Behavior*, 20(1):120–136. https://doi.org/10.1016/S0022-5371(81)90356-X

Simpson, Greg B. 1994. Context and the processing of ambiguous words. *Handbook of Psycholinguistics*, 22:359–374.

Snow, Rion, Sushant Prakash, Daniel Jurafsky, and Andrew Y. Ng. 2007. Learning to merge word senses. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1005–1014.

Snyder, Benjamin and Martha Palmer. 2004. The English all-words task. In *Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 41–43.

Song, Yang, Xin Cai Ong, Hwee Tou Ng, and Qian Lin. 2021. Improved word sense disambiguation with enhanced sense representations. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4311–4320. https://doi.org/10.18653/v1/2021.findings-emnlp.365

Sorace, Antonella and Frank Keller. 2005. Gradience in linguistic data. *Lingua*, 115(11):1497–1524. https://doi.org/10.1016/j.lingua.2004.07.002

Sparck-Jones, Karen. 1964. *Synonymy and Semantic Classification*. Ph.D. thesis, University of Cambridge.

Srinivasan, Mahesh and Hugh Rabagliati. 2015. How concepts and conventions structure the lexicon: Cross-linguistic evidence from polysemy. *Lingua*, 157:124–152. https://doi.org/10.1016/j.lingua.2014.12.004

Stammers, Jonathan. 2008. Unbalanced, idle, canonical and particular: Polysemous adjectives in English dictionaries. *Lexis*, 1. https://doi.org/10.4000/lexis.771

Stokoe, Christopher. 2005. Differentiating homonymy and polysemy. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, pages 403–410. https://doi.org/10.3115/1220575.1220626

Swinney, David A. 1979. Lexical access during sentence comprehension: (Re)consideration of context effects. *Journal of Verbal Learning and Verbal Behavior*, 18(6):645–659. https://doi.org/10.1016/S0022-5371(79)90355-4

Tabossi, Patrizia, Lucia Colombo, and Remo Job. 1987. Accessing lexical ambiguity: Effects of context and dominance. *Psychological Research*, 49(2):161–167. https://doi.org/10.1007/BF00308682

Taieb, Mohamed Ali Hadj, Torsten Zesch, and Mohamed Ben Aouicha. 2019. A survey of semantic relatedness evaluation datasets and procedures. *Artificial Intelligence Review*, pages 1–42.

Taylor, Wilson L. 1953. "Cloze procedure": A new tool for measuring readability. *Journalism Quarterly*, 30(4):415–433. https://doi.org/10.1177/107769905303000401

Tenney, Ian, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP

pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601. https://doi.org/10.18653/v1/P19-1452

Traugott, Elizabeth Closs. 2017. Semantic change. *Oxford Research Encyclopedias, Linguistics*. https://doi.org/10.1093/acrefore/9780199384655.013.323

Travis, Charles. 1997. Pragmatics. In Bob Hale and Crispin Wright, editors, *A Companion to the Philosophy of Language*. Blackwell, pages 87–107.

Travis, Charles. 2008. *Occasion-sensitivity: Selected essays*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199230334.001.0001

Trott, Sean and Benjamin Bergen. 2021. RAW-C: Relatedness of ambiguous words in context (a new lexical resource for English). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7077–7087. https://doi.org/10.18653/v1/2021.acl-long.550

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

Ullmann, Stephen. 1959. *The Principles of Semantics*. Blackwell.

van der Maaten, Laurens and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605.

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*, pages 5998–6008.

Vega Moreno, Rosa E. 2007. *Creativity and Convention: The pragmatics of everyday figurative speech*. John Benjamins. https://doi.org/10.1075/pbns.156

Vicente, Agustin. 2015. The green leaves and the expert: Polysemy and truth-conditional variability. *Lingua*, 157:54–65. https://doi.org/10.1016/j.lingua.2014.04.013

Vicente, Agustin and Ingrid L. Falkum. 2017. Polysemy. *Oxford Research Encyclopedias, Linguistics*. https://doi.org/10.1093/acrefore/9780199384655.013.325

Voorhees, Ellen M. 1993. Using WordNet to disambiguate word sense for text retrieval. In *Proceedings of ACM SIGIR Conference*, pages 171–180. https://doi.org/10.1145/160688.160715

Wang, Alex, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355. https://doi.org/10.18653/v1/W18-5446

Ward, Joe H. Jr. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244. https://doi.org/10.1080/01621459.1963.10500845

Weaver, Warren. 1949. Translation. In W. N. Locke and A. D. Booth, editors, *Machine Translation of Languages: Fourteen Essays*. MIT Press.

Weinreich, Uriel. 1964. Webster's third: A critique of its semantics. *International Journal of American Linguistics*, 30(4):405–409. https://doi.org/10.1086/464799

Westera, Matthijs and Gemma Boleda. 2019. Don't blame distributional semantics if it can't do entailment. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, pages 120–133. https://doi.org/10.18653/v1/W19-0410

Wiedemann, Gregor, Steffen Remus, Avi Chawla, and Chris Biemann. 2019. Does BERT make any sense? Interpretable word sense disambiguation with contextualized embeddings. In *Proceedings of the 15th Conference on Natural Language Processing*, 10 pages.

Wilks, Yorick A., Brian M. Slator, and Louise M. Guthrie. 1996. *Electric Words*. MIT Press. https://doi.org/10.7551/mitpress/2663.001.0001

Wilson, Kyra and Alec Marantz. 2022. Contextual embeddings can distinguish homonymy from polysemy in a human-like way. In *Proceedings of the 5th International Conference on Natural Language and Speech Processing (ICNLSP 2022)*, pages 144–155.

Wittgenstein, Ludwig. 1953. *Philosophische Untersuchungen - Philosophical investigations*. Macmillan.

Yang, Zhilin, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R. Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, 11 pages.

Yenicelik, David, Florian Schmidt, and Yannic Kilcher. 2020. How does BERT capture semantics? A closer look at polysemous words. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 156–162. https://doi.org/10.18653/v1/2020.blackboxnlp-1.15

Zhong, Zhi and Hwee Tou Ng. 2010. It makes sense: A wide-coverage word sense disambiguation system for free text. In *Proceedings of the ACL 2010 System Demonstrations*, pages 78–83.

Zhuang, Liu, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227.

Zipf, George Kingsley. 1945. The meaning-frequency relationship of words. *The Journal of General Psychology*, 33(2):251–256. https://doi.org/10.1080/00221309.1945.10544509, PubMed: 21006715

Zwicky, Arnold M. and Jerrold M. Sadock. 1975. Ambiguity tests and how to fail them. In *Syntax and Semantics volume 4*. Brill, pages 1–36. https://doi.org/10.1163/9789004368828_002