Psychological Review

Word Meaning Is Both Categorical and Continuous

Sean Trott and Benjamin Bergen

Online First Publication, March 9, 2023. https://dx.doi.org/10.1037/rev0000420

AMERICAN PSYCHOLOGICAL ASSOCIATION

Psychological Review

ISSN: 0033-295X

https://doi.org/10.1037/rev0000420

Word Meaning Is Both Categorical and Continuous

Sean Trott and Benjamin Bergen
Department of Cognitive Science, University of California San Diego

Most words have multiple meanings, but there are foundationally distinct accounts for this. Categorical theories posit that humans maintain discrete entries for distinct word meanings, as in a dictionary. Continuous ones eschew discrete sense representations, arguing that word meanings are best characterized as trajectories through a continuous state space. Both kinds of approach face empirical challenges. In response, we introduce two novel "hybrid" theories, which reconcile discrete sense representations with a continuous view of word meaning. We then report on two behavioral experiments, pairing them with an analytical approach relying on neural language models to test these competing accounts. The experimental results are best explained by one of the novel hybrid accounts, which posits both distinct sense representations and a continuous meaning space. This hybrid account accommodates both the dynamic, context-dependent nature of word meaning, as well as the behavioral evidence for category-like structure in human lexical knowledge. We further develop and quantify the predictive power of several computational implementations of this hybrid account. These results raise questions for future research on lexical ambiguity, such as why and when discrete sense representations might emerge in the first place. They also connect to more general questions about the role of discrete versus gradient representations in cognitive processes and suggest that at least in this case, the best explanation is one that integrates both factors: Word meaning is both categorical and continuous.

Keywords: neural language models, ambiguity, continuous state space, mental lexicon

Supplemental materials: https://doi.org/10.1037/rev0000420.supp

Words mean different things in different contexts. In some cases (approximately 7% of words in English, for instance—Rodd et al., 2004), the same sequence of characters or sounds can denote meanings that appear entirely unrelated (e.g., "river *bank*" vs. "financial *bank*"). This phenomenon is typically called *homonymy* (Valera, 2020). Far more frequent (about 84% of English words, per Rodd et al., 2004) is *polysemy*—in which related meanings (e.g., "pet *chicken*" vs. "roast *chicken*") are interpreted as corresponding to different senses of a single word (Cruse, 1986). In the limit, *all* words arguably have meanings that depend on context to some extent, even if not considered outright ambiguous (Elman, 2004; Hoffman et al., 2013; Yee & Thompson-Schill, 2016). For example, the word "runs" evokes subtly different actions in "the boy *runs*" and "the cheetah *runs*" (Elman, 2004); similarly, comprehenders might activate different sensorimotor representations of the word "lemon" in "she cut the *lemon*" and "she juggled the *lemon*" (Yee & Thompson-Schill, 2016).

Each of these phenomena—homonymy, polysemy, and context-dependence—is pervasive across the world's languages (Dautriche, 2015; Valera, 2020). Accordingly, multiplicity of meanings has driven research across many different disciplines, including linguistics (Tuggy, 1993; Valera, 2020), cognitive science and psycholinguistics (Elman, 2004; Rodd et al., 2004), lexicography (Krishnamurthy & Nicholls, 2000), natural language processing (Karidi et al., 2021; Kilgarriff, 2007; Navigli, 2009; Schneider et al., 2015), and legal studies (Schane, 2002), to name just a few. Knowing what the range of meanings is for any given word, or the different patterns that meaning-varying words in general display, is crucial for theories of language knowledge, use, and acquisition.

Yet despite widespread interest, there remains considerable disagreement about exactly how humans represent the multiplicity of word meanings. On some accounts, humans store different lexical representations for wordforms with unrelated meanings (i.e., homonyms), but not for wordforms with multiple, related senses (i.e., polysemes; e.g., Cruse, 1986); other accounts argue that humans maintain distinct representations for both homonyms and polysemes (Kempson, 1977). And still others eschew the notion of discrete

Sean Trott https://orcid.org/0000-0002-6003-3731

The authors are grateful to Ekaterina Klepousniotou and Susan Windisch Brown for making their experimental stimuli available. The authots also thank Jiangtian Li, Gregory Murphy, Seana Coulson, Marta Kutas, and Vic Ferreira for the feedback on an earlier version of this work. Finally, The authos thank members of the Language and Cognition Lab (James Michaelov, Cameron Jones, and Tyler Chang) for the valuable comments and discussion.

The experimental design, hypotheses, and analyses were preregistered on Open Science Framework in advance of data collection (https://osf.io/gj48a). Additionally, data and code to reproduce the preregistered analyses are available on Open Science Framework (https://osf.io/2s7mg/); data and code to reproduce the supplementary analyses (along with the original, preregistered analyses) are also available on GitHub (https://github.com/seantrott/trott_ph_amb). The main results presented in the article have not been published elsewhere, though the relatedness data have been published at a conference (Trott & Bergen, 2021).

Correspondence concerning this article should be addressed to Sean Trott, Department of Cognitive Science, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0515, United States. Email: sttrott@ucsd.edu

lexical representations altogether, arguing instead that word meanings are best characterized as occupying a continuous, context-sensitive state space (Elman, 2004, 2009). Importantly, these different accounts also echo more general issues in cognitive science. To what extent is human semantic knowledge constituted by discrete, symbolic representations versus gradient, subsymbolic systems (Miikkulainen & Elman, 1993)? Are concepts organized by their prototypes or exemplars (Malt, 1989)? Notably, while there have been attempts to adjudicate between a subset of these accounts, none of them is entirely consistent with current empirical evidence, and none can be strictly disconfirmed.

In the sections below, we first describe the testable predictions made by each of these competing accounts, as well as their theoretical limitations. We also introduce and elaborate on two novel "hybrid" accounts, which reconcile discrete sense representations with a continuous view of meaning, and which are designed to overcome the limitations of existing theories. We then report on two behavioral experiments able to adjudicate among them, paired with an analytical approach that relies on recent advances in neural language models (Devlin et al., 2018). The results are best explained by a hybrid account that allows for effects of both continuous (i.e., distance in state space) and categorical (i.e., sense boundaries) factors. Finally, we compare the predictive power of several computational models of the novel hybrid accounts.

The Mental Dictionary Framework

Many accounts of how word meanings are stored and represented can be grouped under the broader umbrella of the Mental Dictionary Framework. Under this view, the mental lexicon is conceptualized as a dictionary, held in long-term memory (Elman, 2004; Pinker, 1997). Each wordform maps onto a *lexical entry*, which contains information about the word's basic semantic and syntactic properties. Accordingly, ambiguous wordforms (like homonyms) map onto multiple, distinct entries, as they would in a literal dictionary. Critically, the categorical boundary between distinct word meanings is theorized to exert an influence on psychological processing above and beyond the context-dependent nature of word meaning. Put another way: there is a qualitative distinction between outright ambiguity (e.g., "delicious *port*" vs. "windy *port*") and mere underspecification (e.g., "*big* building" vs. "*big* ant").

Within this Mental Dictionary Framework, there are at least two dominant theoretical accounts. The primary distinction between these accounts is in how they treat polysemy—that is, words with multiple, related meanings—namely, whether polysemous meanings are represented differently from homonymous meanings. According to sense enumeration accounts, polysemy is represented much like homonymy: all ambiguous words map onto multiple, distinct lexical entries (Kempson, 1977). That is, just as "financial *bank*" and "river *bank*" would constitute distinct entries, so too would "pet *chicken*" and "roast *chicken*." Sense enumeration accounts are considered by some (Klepousniotou, 2002; Pustejovsky, 1995) to be uneconomical; because polysemy is extremely pervasive (Rodd et al., 2004), storing each polysemous meaning separately results in a proliferation of lexical entries. Nonetheless, the chief advantage of sense enumeration accounts is that they sidestep the difficulty of addressing irregular forms of polysemy, that is, cases in which multiple meanings are related but not in a systemic fashion (Kempson, 1977; Rice, 1992). Sense enumeration accounts make two concrete predictions about cognitive processing. First, pairs of related senses (e.g., "pet *chicken*" vs. "roast *chicken*") should be distinguishable in behavior from pairs of same-sense uses (e.g., "roast *chicken*" vs. "marinated *chicken*"). And second, because polysemy is represented in the same fashion as homonymy, the behavior of words with what are classified as related senses should *not* be distinguishable from those with homonymous senses.

Core representation accounts also view the lexicon as storing discrete entries. But unlike sense enumeration accounts, they do not view multiple related senses as separate lexical entries, instead deriving or generating meanings during online language processing from a single "core" representation (Cruse, 1986; Pustejovsky, 1995, 2002; Pustejovsky & Bouillon, 1995). For core representation accounts, the mental lexicon contains not only lexical entries, but also *rules*—much like grammar—for systematically extending word senses as a function of context. The generative lexicon (Pustejovsky, 1995) is one well-known example of a core representation account; Pustejovsky (2002) motivates this additional component by appealing both to parsimony and the underlying systematicity by which meanings are extended. In a generative lexicon, lexical entries are associated with some minimal semantic configuration—what Pustejovsky calls their *qualia structure*—which affords (or precludes) particular inferences when composed with other lexical items. For instance, the wordform "bake" would unambiguously denote a *change-of-state* process, but the interpretation of this process as *change-of-state* or *creation* would be constrained by the speaker's choice of direct object (e.g., "potato" vs. "cake"). This in turn places constraints on the interpretation of the verb. Core representation accounts thus make distinct predictions from sense enumeration accounts. Most notably, because polysemous meanings are represented in a different fashion from homonymous meanings, polysemy, and homonymy should elicit measurably distinct behavior in comprehenders. Further, under a stronger interpretation of core representation accounts, same-sense uses of a word (e.g., "wrapping *paper*" and "shredded *paper*") should not exhibit enhanced facilitation (e.g., in priming, memory, etc.) above and beyond a neutral baseline (e.g., "____ paper"); since the shared core is activated each time the wordform is encountered, even in the baseline condition, the target meaning should be equally accessible (Klein & Murphy, 2001, 2002).

Experimental research offers mixed evidence on these accounts. On the one hand, polysemy does appear to elicit distinct behavior from homonymy. In lexical decision tasks, words categorized as homonymous are recognized more slowly than those categorized as polysemous (Armstrong & Plaut, 2008; Klepousniotou, 2002; Klepousniotou & Baum, 2007; Rodd et al., 2002), possibly because the unrelated meanings compete during lexical access (Rodd et al., 2002). Homonymy may also be more challenging to learn than polysemy (Floyd & Goldberg, 2021; Rodd et al., 2012). These findings are inconsistent with predictions of the sense enumeration account, and in turn favor the core representation account.

On the other hand, senses ostensibly related through polysemy elicit distinct behavior from same-sense uses of a wordform (Klein & Murphy, 2001, 2002; Yurchenko et al., 2020). In a memory task (Klein & Murphy, 2001), subjects were worse at recognizing previously observed wordforms (e.g., "wrapping *paper*") when the repeated phrase employed them in a context evoking a polysemously related sense (e.g., "liberal *paper*") than a same-sense context (e.g., "shredded *paper*"). Similarly, in a primed sensibility judgment task, subjects were also less accurate when responding to

polysemously related senses than same-sense uses or a neutral baseline (e.g., "____ paper"; Klein & Murphy, 2001) and displayed differentiable brain responses in an electroencephalography experiment (Yurchenko et al., 2020). These findings are sometimes interpreted as evidence against the core representation account (Klein & Murphy, 2001). Indeed, they do disconfirm a strong view in which comprehenders process and represent related senses identically to same-sense uses.

Importantly, however, given that most core representation accounts argue that related senses are derived via a generative rule, it is technically possible to reconcile these accounts with the finding that related senses are processed differently from same-sense uses, since the application of this rule might increase processing time (Klein & Murphy, 2001). Some core representation accounts can also accommodate the difference in facilitation between same-sense primes and a neutral baseline—if generative rules can be primed, then it should be easier to transition between same-sense meanings (given that the rule is already primed) than between an underspecified core representation and a more specific meaning (which would require activating the rule for the first time). Thus, these particular results do not allow us to distinguish between sense enumeration accounts and a more nuanced version of core representation accounts.

This leaves considerable uncertainty. Polysemous meanings may indeed be represented separately (as in sense enumeration accounts), or at least enjoy some degree of functional separation (as more nuanced versions of both accounts would predict); but according to some evidence, this representational mechanism appears to be distinct from homonymy (as in core representation accounts). And in fact, recent work suggests that the simple distinction between polysemy and homonymy (and even between same and different senses) may be overly simplistic. Multiple studies (Bambini et al., 2016; Klepousniotou, 2002; Klepousniotou & Baum, 2007; Lopukhina et al., 2018; Yurchenko et al., 2020) have found differences in behavioral and neurophysiological responses to pairs of meanings related via different polysemy mechanisms: *metonymy* (e.g., "pet *chicken*" vs. "roast *chicken*") and *metaphor* (e.g., "polluted *atmosphere*" vs. "relaxed *atmosphere*"). Similarly, other studies (Brown, 2008; Klepousniotou et al., 2008) have found that measures of processing ease (e.g., accuracy and response time) are predicted by the *degree of overlap* or *semantic similarity* between two senses.

Importantly, current evidence as described above does not fully adjudicate between the two accounts falling under the Mental Dictionary Framework. Moreover, other approaches identify certain limitations of this framework and attempt to address them.

Challenges to the Mental Dictionary Framework

The Mental Dictionary Framework—at least as outlined above—has been challenged on several theoretical grounds. Some of these arguments relate specifically to the question of lexical ambiguity, while others concern the role of knowledge and context more generally (Elman, 2004).

Identifying Sense Boundaries Is Challenging

The Mental Dictionary Framework reifies the lexicographic concept of discrete word senses, which requires a commitment as to whether the difference in meaning conveyed by a given pair of word uses corresponds to ambiguity (i.e., distinct senses) or mere context-dependence (sometimes called *vagueness*).

This distinction may appear obvious in some cases (e.g., "river *bank*" vs. "financial *bank*" are readily interpreted as distinct senses), but in many situations, it is difficult to pin down using standard linguistic tests (Geeraerts, 1993; Hanks, 2000; Kilgarriff, 2007; Krishnamurthy & Nicholls, 2000; Tuggy, 1993). Tuggy (1993) illustrates the challenge using the verb "paint," which can describe a number of conceptually related actions, including as follows: (a) painting a portrait in oils; (b) painting a landscape with watercolors; (c) painting stripes on the parking lot; (d) applying makeup to the face; and more. (a) and (b) plausibly belong to the same-sense, but (a) and (c) may seem anomalous when used in *zeugmatic cross-reference* ("I'm painting [a portrait] and Ben is [painting stripes on the road] too"), a common test for distinguishing ambiguity from vagueness. According to that criterion, then, (a) and (c) should be considered distinct senses, suggesting that "paint" is at least partially ambiguous. Yet, even this standard test is not without limitations. Other contexts may permit a crossed reading of these two meanings ("When I'm painting [a portrait], I try to get the color on evenly, and so does Jane when she paints [stripes on the road]"; Tuggy, 1993). Further, it is not always clear whether an anomalous reading arises from lexical ambiguity per se, or because of difficulties that arise during a more general purpose process of pragmatic interpretation (Kilgarriff, 2007). Finally, as others have noted (Geeraerts, 1993; Kearns, 2006), different tests are often in conflict with one another. Two uses of "book" might have different truth conditions and accompany distinct modifiers (e.g., a cultural artifact vs. a physical object) but may still permit cross-reference ("I'm enjoying [this book] but I wish it had larger print"; Kearns, 2006, pp. 369–370).

Of course, the fact that ambiguity is sometimes hard to distinguish from context-dependence is not evidence that the distinction itself is in principle invalid. Perhaps, we simply need better tests, or the existing tests ought to be weighted in more sophisticated ways. But it does indicate that the situation is more complex than it might appear at first glance, particularly when we apply this distinction to the mental representation of word meaning—and surprisingly, there is a dearth of studies investigating the psychological reality of discrete word senses (as distinct from context-dependence) in the first place.

Homonymy and Polysemy (and Context-Dependence) Are Not Easily Distinguished

The distinction between homonymy and polysemy is also notoriously challenging to define and detect (Tuggy, 1993; Valera, 2020). These forms of lexical ambiguity are typically distinguished in one of two ways: (a) determining whether a given pair of meanings shares a common *etymon* or etymological source (polysemy) or not (homonymy); and (b) determining whether a given pair of meanings is conceptually similar or related (polysemy), or not (homonymy; Valera, 2020). Yet, both methods have limitations. Shared etymology is difficult to establish and does not entail synchronic psychological association—even if two meanings were once related, the phenomenon of *semantic drift* can lead to those meanings drifting apart over time, leading to apparent homophony (Tuggy, 1993). For example, the words *flour/flower* actually began as a borrowing from the same etymon (*flur*) from French (*fleur*, meaning both "blossom" and "the choicest part of something"); these meanings drifted in various ways (e.g., "flower of wheat" referring to the endosperm of wheat), and in

fact retained the same spelling until the 18th century—now the words are heterographic homonyms (Jurafsky, 2014).

Psychological relatedness seems preferable in principle if our goal is to establish theories about the structure of the mental lexicon. But assessing psychological relatedness raises thorny definitional and methodological questions: How exactly is "relatedness" established? Should some mechanisms or manners of conceptual relation—such as metaphor or metonymy—be weighted more heavily than others, or does any manner of relation count? Moreover, as others have noted (Klepousniotou, 2002; Valera, 2020) the very notion of relatedness—and the way it is usually measured—lies on a *continuum*, as opposed to a dichotomy. This has led some (e.g., Deane, 1988; Tuggy, 1993) to suggest that homonymy, polysemy, and underspecification ought to be considered as lying along a cline as well:

> In effect, the three types form a gradient between total semantic identity and total semantic distinctness. (Deane, 1988, p. 327)

> Ambiguity and vagueness may be seen as occupying opposite ends of a continuum with polysemy in the middle. (Tuggy, 1993, p. 1)

This continuum view is to some extent compatible with current psycholinguistic evidence. As noted earlier, processing ease (as measured by RT, accuracy, N400 effects, etc.) on several different tasks (e.g., lexical decision, primed sensibility judgments, etc.) is predicted not only by the coarse distinction between homonymy and polysemy, but also by the *degree of overlap* between two meanings (Brown, 2008; Klein & Murphy, 2002; Klepousniotou et al., 2008).

However, a cline between ambiguity and underspecification does not square easily with the Mental Dictionary Framework. Under a strong interpretation of "continuum," categories such as homonymy and polysemy are helpful descriptive abstractions, but are not viewed as psychologically real; it is challenging to reconcile this position with the Mental Dictionary Framework, in which word meanings are represented in discrete entries. This suggests that an alternative framework might be required—one that allows for greater flexibility of representation and context-dependence.

Word Meaning Is Flexible and Context-Dependent

A more general critique of the Mental Dictionary Framework is presented by Elman (2009), who argues that in general it cannot adequately address the dynamic, context-dependent nature of word meaning. Elman (2009) reviews a large body of psycholinguistic research, demonstrating that words encode detailed world knowledge, and that this knowledge appears to play an early role in sentence processing. This includes early detection of incompatible or unlikely instrument/patient pairings (e.g., *Susan used the scissors to cut the expensive wood*), the ability of discourse context to override typical verb/patient pairings (e.g., a "shopping" context renders *the lifeguard saved* ***money*** easier to process, even though the default expectation might be *saved* ***lives***), and more. In other words, "lexical representations contain a significant amount of detailed word-specific information that is available and used during online sentence processing" (Elman, 2009, p. 566).

For Elman, this raises the question of *which* information is included in these lexical representations. Overly sparse entries (e.g., a phonological representation and part-of-speech) cannot account for the early effects of lexical knowledge; but if we instead add sufficient detail to these entries to accommodate the psycholinguistic evidence, it results in a combinatorial explosion (e.g., storing all the possible instrument/patient contingencies and their respective compatibilities). Elman (2004, 2009, 2011) ends up rejecting the notion of discrete lexical entries altogether, instead advocating for a view in which word meaning is represented as *trajectories* through a continuous state space. This alternative view, which we call the Continuity of Meaning Framework, is described in more detail below.

The Continuity of Meaning Framework

In the Continuity of Meaning Framework, words are conceptualized as *cues to meaning*—eliciting context-dependent trajectories through a continuous state space, as in a recurrent neural network (Elman, 2004, 2009, 2011; Kawamoto, 1993; Li & Joanisse, 2021). In theory, the dimensions of this state space could be constituted by many different features of lexical experience, including the distributional statistics or usage patterns of a word (Li & Joanisse, 2021), as well as sensorimotor associations with that word (Elman, 2011). In this article, we focus primarily on the role of distributional patterns, but a potential role for sensorimotor correlates is considered in the General Discussion. As an additional point of clarification, we note that the *meaning* of a word under this Continuity of Meaning Framework could be conceptualized either as its *trajectory* through state space (i.e., under a nonrepresentational account) or its *location in* state space, or even as a *function* that constrains the trajectory taken through state space; we do not aim here to adjudicate between these different interpretations.

The precise trajectory elicited by a particular word token (e.g., "runs") will necessarily be contingent on the prior state of the network, which in turn is entirely dependent on context (e.g., "the boy *runs*" vs. "the cheetah *runs*"). Thus, this approach builds the role of context directly into its conception of word meaning: rather than positing discrete *senses* for two contexts of use, the difference in meaning can be captured by the different trajectories elicited by "runs" across those two sentential contexts. Accordingly, when the same wordform is encountered in contexts that differ to a greater degree, it will also elicit trajectories through the network that differ more—the distance in state space between "the boy *runs*" and "the cheetah *runs*" should be smaller than the distance between "the boy *runs*" and "the clock *runs*." This yields the theoretical benefits of word "types" without the disadvantages discussed above (Elman, 2004), in the following way. To the extent that two tokens elicit similar trajectories in state space, they behave quantitatively like a common "type" of sorts—but while also differing in subtle, context-dependent ways. This framework also reflects a larger paradigm shift toward continuous accounts of cognitive processes more generally (Spivey, 2008; Spivey & Dale, 2004, 2006); increasingly, many processes thought to consist of discrete operations carried out over symbolic representations have been modeled using a dynamical systems approach that posits no explicit representations (Beer, 2003; Chemero, 2011; Spivey, 2008).

How might this framework handle the problem of lexical ambiguity? In its strongest theoretical implementation, the notion of discrete sense categories is rejected altogether. This view—which we call pure exemplar theory—holds that discrete meaning categories for a word (i.e., "senses") is a convenient theoretical abstraction, but is not psychologically real. A "sense" is simply a label describing a stable pattern of activity within the high-dimensional state space. According to pure exemplar theory, the difference between lexical ambiguity and context-dependence is entirely a matter of

degree: all words elicit variable trajectories through state space, and although we might decide that some of these trajectories are better described in terms of multiple "sense clusters," this distinction is not assumed to be cognitively relevant—it does not influence cognitive processing above and beyond the *distance* in state space between any two contexts of use. This theory thus has an affinity to other accounts of language processing that eschew stored abstractions (Ambridge, 2020).

As a consequence, on pure exemplar theory, the difference between homonymy and polysemy is also one of degree, not kind. Homonymy corresponds to words with more distant, differentiable contexts of use, while polysemy corresponds to words whose contexts of use are closer in state space. A similar account is presented in Rodd (2020), in which these phenomena are understood from the perspective of attractor basins. Homonymous meanings correspond to distant, deep attractor basins, while polysemous meanings correspond to shallow, more connected basins.[1]

To date, pure exemplar theory cannot be strictly disconfirmed by any existing psycholinguistic research. Merely finding a difference in how comprehenders process homonymous or polysemous words, as many studies have (Armstrong & Plaut, 2008; Klepousniotou, 2002; Klepousniotou & Baum, 2007; Rodd et al., 2002), does not rule out the possibility that this difference simply reflects distances in a fundamentally continuous space; if homonymous meanings are more distant than polysemous meanings, then pure exemplar theory predicts that they should be harder to process. The same goes for finding a difference between how people process ostensibly same-sense and different-sense uses (Klein & Murphy, 2001): if same-sense uses are used in more similar contexts, then pure exemplar theory predicts that they should be easier to process than different-sense uses. Thus, pure exemplar theory currently accommodates existing data at least as well as either theory falling under the Mental Dictionary Framework. It is also more consistent with other evidence that is harder to reconcile with either account, such as the finding that the degree of overlap between two meanings—or more aptly, between two *contexts of use*—influences behavior (Brown, 2008; Klein & Murphy, 2002; Klepousniotou et al., 2008).

Hybrid Meaning Framework: Category Effects in a Continuous State Space

Although pure exemplar theory cannot be disconfirmed by current empirical evidence, there are at least two reasons to think it might not stand up to a more targeted falsification attempt.

First, outside of lexical ambiguity, there are a number of domains in which humans treat continuously varying input as falling into discrete categories that have psychological effects above and beyond continuous variation in that input (Goldstone & Hendrickson, 2010). This phenomenon, categorical perception, transforms the perceptual space "such that differences between objects that belong in different categories are accentuated, and differences between objects that fall into the same category are deemphasized" (Goldstone & Hendrickson, 2010, p. 69). Evidence for categorical perception is often demonstrated by manipulating a continuous stimulus, such as voice onset time or color hue, and asking whether behavioral or neurophysiological responses to that stimulus exhibit discontinuity. Many of these domains involve language in some way (though not all, e.g., face perception). For example, responses to continuous variation in acoustic input exhibits discontinuity dependent on the phoneme categories of a language (Liberman et al., 1957). Similarly, neurophysiological responses to variation in color hue are dependent on language-specific color categories (Mo et al., 2011; Thierry et al., 2009). This phenomenon also extends to objects, that is, whether two distinct referents are cocategorized by the lexicon of a language. English speakers distinguish *cups* from *mugs*, while Spanish speakers refer to both as *taza*. Accordingly, English speakers exhibit a sharper visual mismatch negativity effect when viewing pictures of mugs interspersed with those of cups (or vice versa), than do Spanish speakers (Boutonnet et al., 2013). While this last example involves distinct referents (i.e., cups and mugs) rather than continuous variation in a perceptual stimulus (e.g., color hue), it remains relevant to the question of lexical ambiguity. If sense categories are psychologically real, one might expect them to exert a similar influence: That is, the conceptual distance between two contexts of use should be magnified if those contexts straddle a sense boundary—and compressed if they fall within a single sense category.

Second, recent empirical evidence from an offline task (Trott & Bergen, 2021) is broadly consistent with this prediction. Trott and Bergen (2021) asked participants to rate the conceptual relatedness of the same wordform in two different contexts of use. In some cases, these contexts corresponded to the same-sense (e.g., "marinated lamb" vs. "roasted lamb"), while others corresponded to different senses (e.g., "marinated lamb" vs. "friendly lamb"). Additionally, some different-sense pairs were classified (according to dictionaries) as polysemous (e.g., "marinated lamb" vs. "friendly lamb"), while others were homonymous (e.g., "furry bat" vs. "baseball bat"). Participants' ratings were compared with a continuous measure of the distance between these contexts of use, obtained using the neural language model BERT[2], or Bidirectional Encoder Representations from Transformers (Devlin et al., 2018). As expected, more distant contexts were rated as less related (*Pearson's* $r = 0.58$). Critically, however, BERT consistently *underestimated* how related participants found same-sense pairs to be, and *overestimated* how related they found different-sense homonyms to be (Trott & Bergen, 2021). Both results point to the possibility that participants' relatedness judgments were influenced not only by continuous variation across contexts of use, but also by human sense categories. According to this interpretation, sense categories compressed the conceptual distance between same-sense pairs and amplified the distance between distance-sense pairs (particularly for homonyms).

On the other hand, there are several important limitations to this result. First, as participants' relatedness judgments were made offline, it remains unclear whether putative sense categories play an active role in shaping online word processing in context. Second, participants were explicitly asked to rate meaning similarity on a labeled scale from 1 (*totally unrelated*) to 5 (*same meaning*). This might have encouraged participants to draw on metalinguistic category knowledge to complete the task, even if such knowledge does not actually influence the course of "ordinary" language comprehension. Together, these limitations imply that we cannot yet rule out the pure exemplar theory as a viable account of the mental lexicon.

[1] Rodd (2020) does not necessarily argue for some form of the pure exemplar theory. Rather, it is the closest example of a model of lexical ambiguity in which meaning is seen as distributed feature-vectors in a continuous landscape. It is possible that Rodd's (2020) state-space model is compatible with a cognitive distinction between ambiguity and context-dependence.

[2] BERT is described in more detail in the Current Work section.

Of course, as noted in the previous section, there are also a number of limitations to both theories falling under the Mental Dictionary Framework. This raises the possibility of a hybrid account—one that reconciles the notion of discrete sense categories with a continuous, graded meaning space.

Hybrid Meaning Theory

Hybrid meaning theory posits the existence of senses (or "sense clusters"). These sense categories warp the underlying continuous context space according to which category a particular point or trajectory within that space belongs to. Specifically, contexts of use belonging to the *same*-sense category should become closer together, while contexts of use belonging to *different*-sense categories should become further apart.

Importantly, this theory requires that the cocategorization of two contexts of use depend on some factor other than distance in context space. That is, hybrid meaning theory is not merely an exaggeration of existing clumpiness. Rather, it requires that contexts of use are somehow *assigned* to distinct sense categories, which themselves are derived from a source of information or representation external to that context space—and which, in turn, warp the distance between those usage contexts. There are many possible mechanisms by which sense categories might form, including: identification of distinct sensorimotor associations for different contexts of use, distinct communicative or pragmatic contexts, and more (see the General Discussion, for a more detailed description). Importantly, the primary commitment of hybrid meaning theory is not to a specific categorization mechanism, but to the claim that sense categories impinge on a continuous meaning space and transform that space in some way.

Figure 1 presents one possible implementation of these transformations: within a sense cluster, points attract toward the centroid of that sense category, resulting in an exaggeration of conceptual distance across clusters. We call this mechanism the sense attraction account.

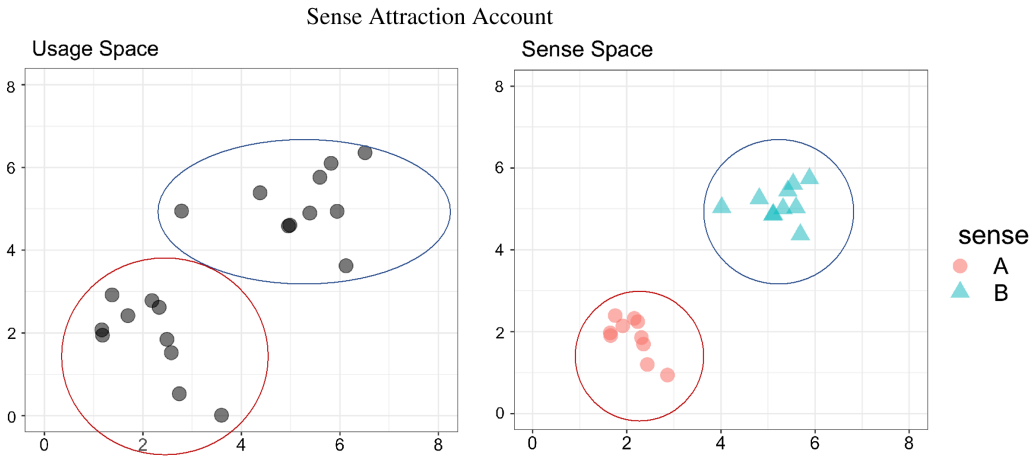
Another possible mechanism, which we call the sense distillation account, is illustrated in Figure 2. Unlike the sense attraction account, within-cluster variance is *distilled* into a single point, that is, the centroid of that cluster. Critically, this preserves the metric properties of the continuous space: Clusters with centroids that are relatively closer together will result in sense representations that are also closer in meaning space. But because within-cluster variance is removed, the sense distillation account predicts that the difficulty of transitioning between two within-sense contexts of use is not predicted by their distance in usage space—whereas the sense attraction account predicts that within-cluster variance should predict processing difficulty even for same-sense uses of a wordform.

Both possible accounts outlined above are analogous to more general cognitive mechanisms implicated in the resolution of continuously varying or ambiguous input into discrete categories, such as categorical perception (Goldstone & Hendrickson, 2010) or the Ganong effect (Ganong, 1980). As described earlier, these theoretical mechanisms have played an important role in accounting for human behavior in other domains (e.g., speech perception); we propose that an analogous mechanism could be of use in explaining human lexical knowledge.

Both implementations of hybrid meaning theory (sense compression and sense distillation) also acknowledge the importance of continuity and context-dependence, as well as the possibility that the mind carves further structure into this continuous space. While this theory has not been directly tested, its stipulation of continuous gradation in meaning allows it to accommodate existing evidence for the dynamic, flexible nature of word meaning (Elman, 2009). Further, its representation of category structure makes it consistent with evidence that discrete sense representations play a role in cognitive processing (Klein & Murphy, 2001, 2002; Yurchenko et al., 2020). It also makes a concrete prediction that differentiates it from pure

Figure 1

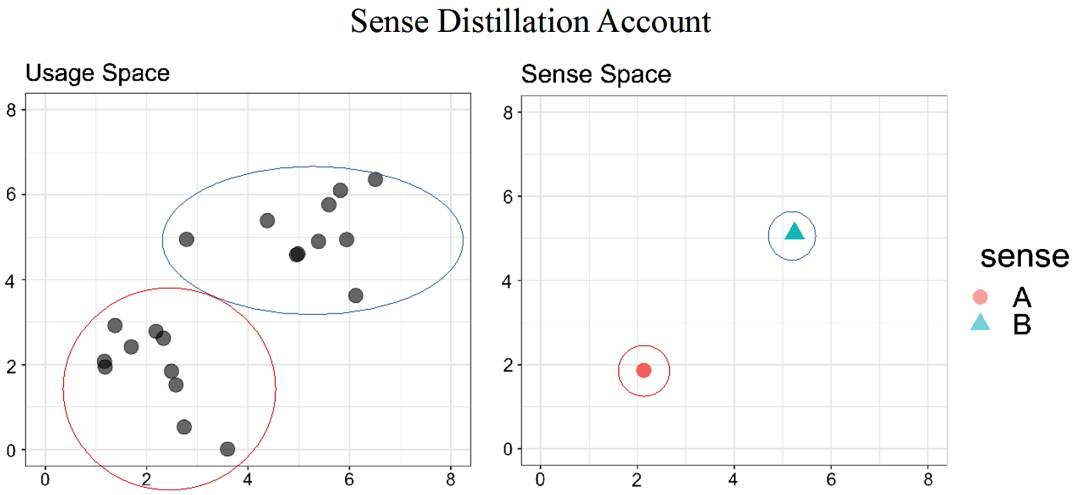
In the Sense Attraction Account, Existing Clumpiness in Usage Space Is Exaggerated



Note. The distribution of dots in Usage Space (left panel) reflects context-dependent variation in the usage of a particular token; the distribution of dots in Sense Space (right panel) preserves some of that variation but also reflects a cognitively imposed sense category. Note that the different colors and shapes are intended to reflect distinct sense categories imposed on the original continuous usage space. For within-cluster uses of a wordform, contextual distance is compressed in meaning space; for across-cluster uses of a wordform, contextual distance is amplified. See the online article for the color version of this figure.

Figure 2

In the Sense Distillation Account, Clusters Are Distilled Into Their Centroids



Note. The distribution of dots in Usage Space (left panel) reflects context-dependent variation in the usage of a particular token; the distribution of dots in Sense Space (right panel) reflects cognitively imposed sense categories. This removes within-cluster (i.e., within-sense) variance entirely, but preserves the underlying metric properties of the continuous space—that is, distant centroids will still be more distant than closer centroids. Note that the different colors and shapes are intended to reflect distinct sense categories imposed on the original continuous usage space. See the online article for the color version of this figure.

exemplar theory, as well as from both theories falling under the Mental Dictionary Framework. Namely, the ease of transitioning between two contexts of use, as in primed sensibility judgment tasks (Brown, 2008; Klein & Murphy, 2001; Klepousniotou et al., 2008) should be affected both by the distance in usage space between those contexts *and* whether or not a sense boundary separates those uses.

Nevertheless, at present, there is no reason to prefer this theory over the more parsimonious pure exemplar theory, which has not yet been disconfirmed and which also accommodates existing evidence.

Hybrid+ Theory

The hybrid meaning theory described above claims that discrete sense categories are integrated with a continuous meaning space. Yet, there is also evidence that human sense knowledge is further shaped by the kind of ambiguity at play (Armstrong & Plaut, 2008; Klepousniotou, 2002; Klepousniotou & Baum, 2007; Rodd et al., 2002; Trott & Bergen, 2021). If this is true, the hybrid meaning theory must be augmented with a categorical distinction between homonymy and polysemy—above and beyond distances in state space. We call this augmented view the hybrid+ theory.

As noted earlier, the Continuity of Meaning Framework predicts that on average, pairs of homonymous senses are likely to occupy more distant regions of sense space than pairs of related senses (Rodd, 2020). That is, homonyms and polysemes occupy a continuum of *proximity in sense space* ranging from very close to very distant. Neither pure exemplar theory nor hybrid meaning theory categorically distinguishes the two phenomena. In principle, however, it is conceivable that the human mind transforms a continuous space not only with discrete sense representations, but also in a way that reflects distinct kinds of lexical ambiguity. This transformation could have the result of differentiating homonymy and polysemy above and beyond the proximity of their sense clusters in usage space (Klepousniotou, 2002; Klepousniotou & Baum, 2007; Rodd et al., 2002; Trott & Bergen, 2021). We call this modified view hybrid+ theory, given that it posits both discrete sense representations and distinct kinds of relationships between these sense representations, all atop continuous effects of context.

As with sense categories, there are a number of reasons why a categorical difference between polysemy and homonymy could emerge. For one, various theories of lexical representation propose that they are realized through different cognitive mechanisms (Cruse, 1986), which could produce categorically distinct behavior. Additionally, polysemy is partly systematic, both within and across languages (Srinivasan & Rabagliati, 2015), which might scaffold the learning of new polysemous meanings of known wordforms (Srinivasan & Snedeker, 2011) as compared with unrelated meanings of the same wordform (e.g., "dog *bark*" and "tree *bark*").[3] In theory, this differentiation could also occur along multiple levels of granularity, distinguishing not just homonymy and polysemy but also different kinds of polysemy, such as metaphor and metonymy (Yurchenko et al., 2020). Differentiation could even occur at the level of specific semantic relations, such as animal/meat or material/product (Srinivasan & Rabagliati, 2015; Srinivasan & Snedeker, 2011).

Current Work

Thus far, we have reviewed several theories of how humans process and represent word meaning, with a particular focus on ambiguous words. The Mental Dictionary Framework views word

[3] Note, however, that there are many examples of nonsystematic polysemy, which, as mentioned in the Introduction, is sometimes cited as a motivation for sense enumeration accounts. Many languages have exceptions to rules (e.g., English does not extend "pig" to include the meat made from that animal, possibly because we have a lexical alternative), and many clusters of related meanings have irregular relations (Lehrer, 1990).

meanings as analogous to entries in a dictionary; each unique form-meaning pairing is represented in a *lexical entry*, with ambiguous words (like homonyms) corresponding to multiple lexical entries. The Mental Dictionary Framework can be further subdivided into accounts that distinguish between polysemy and homonymy (core representation accounts) and those that view all ambiguous words as mapping onto distinct lexical entries (sense enumeration accounts). Crucially, both kinds of account claim that word senses are psychologically real and constitute categorical representations in the mind.

In contrast, The Continuity of Meaning Framework views word meanings as trajectories through (or locations in) a continuous, context-sensitive state space. In pure exemplar theory, the notion of discrete sense representations is rejected altogether, along with the categorical distinction between homonymy and polysemy (Elman, 2009).

We also described two novel, "hybrid" theories falling under the Hybrid Framework. Hybrid meaning theory claims that meaning is constituted by a continuous state space, but also that existing "clumpiness" in a word's pattern of use is *exaggerated* by the mind (see Figures 1 and 2). The hybrid+ theory takes this model one step further and claims that the mind further differentiates between homonymy and polysemy in this continuous space.

To test these theories, we selected a methodological paradigm—primed sensibility judgments—that has previously been used to demonstrate categorical effects of sense boundaries (Klein & Murphy, 2001; Yurchenko et al., 2020), as well as a distinction between homonymy and polysemy (Brown, 2008; Klepousniotou et al., 2008). Specifically, processing difficulty—as indexed by response time (RT) and accuracy—is increased when the uses of an ambiguous wordform across a prime and target sentence correspond to what are classified as different *senses* (Klein & Murphy, 2001; Yurchenko et al., 2020); this effect is larger for different-sense sense pairs classified as homonyms (Brown, 2008), or with less semantic overlap (Klepousniotou et al., 2008), than for words that are closely related.

Each theory makes distinct, testable predictions about which variables should influence behavior, and which should not. This means that the predictions of each theory (with the exception of hybrid+ theory) can be falsified by finding that some variable of interest (e.g., sense boundaries) predicts behavior when the theory claims that it should not. For example, pure exemplar theory predicts that the ease of transitioning between two contexts of use (as measured by RT or accuracy) should be predicted by a continuous measure of the distance between those contexts in usage space—but not by whether those contexts of use span a sense boundary (e.g., "marinated lamb" and "friendly lamb") or belong to the same-sense (e.g., "marinated lamb" and "roasted lamb"). Conversely, both varieties of the Mental Dictionary Framework predict an effect of sense boundaries on behavior, but not a graded effect of contextual distance above and beyond this categorical effect. Hybrid meaning theory predicts both a graded effect of contextual distance *and* an effect of sense boundaries—but critically, the effect of sense boundaries should *not* be different across homonyms and polysemes (once contextual distance is accounted for). In other words, hybrid meaning theory (along with pure exemplar theory, and both mental dictionary theories) does not predict an *interaction* between sense boundary and ambiguity type. Technically, this theory is compatible with a main effect of ambiguity type (i.e., an overall difference across homonymous and polysemous stimuli), given that different words and sentence frames will be used. In order to falsify the theory's predictions, we would need to observe an interaction: a larger effect of sense boundaries for homonyms than polysemes, as observed for offline judgments in Trott and Bergen (2021). Only hybrid+ theory is compatible with this interaction effect. Accordingly, only this final theory cannot be strictly falsified via traditional null hypothesis significance testing, given that it predicts nonzero effects for all variables of interest. That said, certain patterns of results are nonetheless more compatible with alternative, simpler theories; for example, there is little reason to prefer hybrid+ theory if no graded effect of context is found, once categorical sense representations are accounted for (see Table 1).

Past work has focused primarily on adjudicating between the varieties of the Mental Dictionary Framework. Although a number of researchers have raised the possibility of homonymy and polysemy occupying a continuum (see Challenges to the Mental Dictionary Framework above), none have attempted to directly adjudicate between the Mental Dictionary Framework and pure exemplar theory, nor test the hybrid theories introduced here. That is what the current experiments aimed to do.

Measuring Continuous Contextual Distances

A critical prerequisite for comparing these theories is operationalizing the notion of continuous distance in state space. Such an operationalization must be both continuous and context sensitive, so that one context of use (e.g., the word "lamb" in "marinated lamb") can be compared to another (e.g., in "friendly lamb"), for example, by calculating the distance between these contexts.

To operationalize this notion of continuity, we used BERT (Devlin et al., 2018), a state-of-the-art neural language model (NLM). There is a growing body of literature using BERT and other NLMs as operationalizations of human lexical-semantic knowledge in general (Haber & Poesio, 2020a, 2020b; Li & Joanisse, 2021; Nair et al., 2020; Trott & Bergen, 2021), and to test Elman's (2004; 2009) cues to meaning framework in particular (Li & Joanisse, 2021; Trott & Bergen, 2021). It is important to note that BERT (like most

Table 1
Predicted Effects of Each Theory

	Mental Dictionary Framework		Continuity of Meaning Framework	Hybrid Framework	
Predicted effects	Sense enumeration	Core representation	Pure exemplar theory	Hybrid	Hybrid+
Graded effects of context	—	—	Yes	Yes	Yes
Effect of sense boundaries	Yes	Yes	—	Yes	Yes
Effect of sense boundary larger for homonyms than polysemes	—	Yes	—	—	Yes

NLMs) is trained on linguistic input alone (Bender & Koller, 2020), and lacks access to any extralinguistic sources of information that humans might use to represent the meanings of a word, such as sensorimotor associations. Thus, BERT reflects a particular operationalization of the Continuity of Meaning Framework: Its representational space is continuous, and the topology of this continuous space is determined by statistical regularities in which words co-occur with which other words. While this operationalization has clear limitations (Bender & Koller, 2020), it is compatible with views of linguistic meaning that emphasize the role of usage (Wittgenstein, 1953), such as the *distributional semantic hypothesis* (Firth, 1957; Harris, 1954; Lenci, 2008). The distributional semantic hypothesis states that words with more similar meanings should appear in more similar contexts—and consequentially, that meaning similarity should be derivable from contextual similarity. Importantly, we do not necessarily view BERT as a psychological model per se: Rather, the information contained in BERT's representational space (i.e., statistical relationships between words and word contexts) could be viewed as a *component* of (or input to) a psychological model, which may be particularly well-suited to representing contextual similarity (Lake & Murphy, 2021). The open question raised by the current work is just how much of human behavior can be explained by this component.

BERT (base) was trained on a large text corpus (>3 billion word tokens) using two objectives: (a) a masked language modeling task, in which the model learns to predict a "masked" word in some sentential context (e.g., "I went to the [MASK] bank"); and (b) next-sentence prediction, in which the model must learn to predict whether two sentences occurred next to each other. After training, BERT can be used to produce *contextualized embeddings* of a given wordform, a vector representation reflecting both that wordform's statistical distribution in the training corpus, as well as the immediate context in which that word appears. That is, rather than producing a context-insensitive embedding for a given string, as earlier distributional semantic measures like Latent Semantic Analysis and Hyperspace Analogue to Language do, BERT's embeddings are sensitive to the linguistic context in which a word token is observed. BERT's architecture appears to naturally encode a number of linguistic features, such as part of speech, semantic roles, and others (Tenney et al., 2019). These contextualized embeddings have been shown to improve performance on a number of downstream Natural Language Processing tasks involving lexical ambiguity, such as word sense disambiguation (Aina et al., 2019; Loureiro et al., 2020). Past work also suggests that BERT can be used to distinguish monosemous and polysemous words, or even polysemy and homonymy (Haber & Poesio, 2020a, 2020b; Nair et al., 2020; Soler & Apidianaki, 2021), and that BERT's representations encode sense-like information (Karidi et al., 2021). Most relevantly for our purposes, BERT's contextualized embeddings are well-suited for measuring contextual distance in a graded manner—given two contextualized embeddings of an ambiguous target word (e.g., for "marinated *lamb*" and "friendly *lamb*"), we can compute the cosine distance between those vectors, a metric often used to assess proximity in vector space.[4] Smaller cosine distances indicate that the embeddings are closer, while larger values indicate they are further apart.

Accounting for contextual distance in a primed sensibility judgment task allows us to adjudicate among the theories outlined above. Pure exemplar theory predicts that the difficulty in transitioning between two contexts of use should be affected solely by their proximity in usage space—thus, the existence of a sense boundary (or the distinction between homonymy and polysemy) should not predict variance in RT or accuracy above and beyond cosine distance. Both varieties of the Mental Dictionary Framework predict the opposite, that is, cosine distance should *not* explain variance in RT or Accuracy above and beyond the existence of a sense boundary. And both hybrid theories predict a systematic distortion of this continuous usage space, such that the existence of a sense boundary (or the distinction between homonymy and polysemy) should increase measures relating to processing cost (e.g., RT or Accuracy), above and beyond the cosine distance as measured by BERT.

Experiment 1

The primary goal of Experiment 1 was adjudicating between the competing theoretical accounts outlined above (see Table 1). This work is also (to our knowledge) the first attempt to directly test the Continuity of Meaning Framework using a measure of online processing ease. Past work (Li & Joanisse, 2021; Nair et al., 2020; Trott & Bergen, 2021) has used NLM-derived measures (e.g., cosine distance) to predict relatedness judgments, but has not directly pitted those continuous measures against categorical factors (such as the existence of a sense boundary) to ask whether both explain independent sources of variance in processing difficulty.

The experimental design, hypotheses, and analyses were preregistered on Open Science Framework (OSF) in advance of data collection (https://osf.io/gj48a). Additionally, data and code to reproduce the preregistered analyses are available on OSF (https://osf.io/2s7mg/); additional data and code to reproduce the supplementary analyses are also available on GitHub (https://github.com/seantrott/trott_ph_amb).

Method

Participants

We recruited 216 participants from the UC San Diego Psychology Department Subject Pool. After following the exclusion criteria listed in our preregistration (https://osf.io/gj48a), we had a total of 180 participants (our target sample size). The exclusion criteria included: participants who self-reported as nonnative speakers of English, participants who failed at least one of the two "bot check" questions at the beginning of the experiment, participants who self-reported as having completed the experiment on a mobile device, and participants for whom more than half of critical trials were excluded because of overly slow (RT > 3 *SD* above the subject-level mean) or overly fast (<500 ms) responses. Of the final set of participants, 144 self-identified as female (33 male, 2 nonbinary, and 1 preferred not to answer). The average age was 20.5 ($SD = 1.67$) and ranged from 18 to 29.

The target sample size of 180 was based on a pilot study with 74 participants. In the pilot study, we detected significant ($p < .001$) effects of both cosine distance and sense boundary, but only a marginally significant interaction between sense boundary and ambiguity type in predicting accuracy. Thus, it was inconclusive from the pilot whether hybrid meaning theory or hybrid+ theory was

[4] Note that we also replicated the primary analyses using ELMo, another well-known contextualized language model. Our preregistered analyses used BERT because it tends to outperform ELMo on word sense disambiguation tasks (Wiedemann et al., 2019) and predicting relatedness judgments (Trott & Bergen, 2021), and because it was more predictive of response time in a pilot study.

a better explanation of the data. We conducted a simulation-based power analysis using the simR package (Green & MacLeod, 2016) to determine the number of participants we would need to detect the interaction between Ambiguity Type and Sense Boundary with 95% power at an α of .025 (to correct for the two dependent variables). The power analysis indicated that 95% power could be achieved with 180 participants; we then estimated the number of participants we would need based on applying the exclusion criteria to the pilot data. (More details are included in the preregistration.)

The study was carried out with the approval of the UC San Diego Institutional Review Board.

Materials

We adapted materials from several previous studies (Brown, 2008; Klepousniotou, 2002; Klepousniotou & Baum, 2007; Klepousniotou et al., 2008). These studies either used sentence fragments containing an ambiguous word (e.g., "marinated *lamb*" or "*fixed* the radio") or used homonymous and polysemous words in isolation (e.g., "bat"). For each ambiguous word, we created four sentences (two for each of the primary senses). Thus, there were six possible sentence pairs for each word: two same-sense pairs and four different-sense pairs. Each sentence for each word contained the same sentence frame (e.g., "They liked the ___ lamb"), but differed in the disambiguating word (e.g., "marinated" vs. "friendly"); a minority of words (13) had at least one sentence which used a different article before the disambiguating word than the other sentences (e.g., "a" vs. "an"). We began with 115 items total (460 sentences).

We used two dictionaries (Merriam Webster and the Oxford English Dictionary) to determine whether the two meanings expressed by a word were categorized by lexicographic experts as different senses. There were three words for which neither dictionary listed the meanings as separate senses at all (e.g., "glossy magazine" vs. "sports magazine"), suggesting that lexicographers viewed these meanings as the same. These items were included in the norming study, but not in the final stimulus set (leaving us with 112 words). We also used both dictionaries to annotate whether different-sense items were classified by lexicographers as related via homonymy or polysemy; meanings listed as separate entries were annotated as homonymy, and those listed in the same entry were annotated as polysemy. There was one word ("drill") for which the two dictionaries did not agree; in this case, we labeled the two meanings as homonymy, following the Oxford English Dictionary.

We also created a number of filler items (112 unique wordforms). Each filler word was matched for the concreteness, frequency, part of speech, and length (number of syllables) of one of the critical wordforms. Then, for each filler, we constructed two sentences containing that word, that is, a minimal sentence pair. For 38 of these filler items (approximately one-third), both sentences were nonsensical; for the remaining 74 (approximately two-thirds), only one of the two sentences was nonsensical (counterbalanced for whether the first or second was nonsensical). This was to prevent participants from learning any contingencies between the prime and target item.

Finally, we ran a norming study to obtain relatedness judgments for all of the critical sentence pairs (Trott & Bergen, 2021). Eight of the words had very low relatedness judgments for their same-sense pairs, so we excluded these from the final stimulus set, leaving us with 104 wordforms (and 624 unique sentence pairs, not accounting for order). In this final set, 30 wordforms were labeled as homonymous, and 74 were polysemous. Seventy-six of the target wordforms were used as nouns, and 28 were used as verbs.

Among this final set of 104 words, mean relatedness from the norming study was (as expected) higher among same-sense (M = 3.53, SD = 0.451) than different-sense (M = 1.38, SD = 1.13) pairs. Further, different-sense homonyms were less related on average and also exhibited less variability (M = 0.44, SD = 0.37), than different-sense polysemes (M = 1.76, SD = 1.11). This was also expected: The polysemous meanings ranged considerably in their relatedness, from highly related meanings (e.g., "marinated *lamb*" vs. "friendly *lamb*") to less related meanings (e.g., "brain *cell*" vs. "prison *cell*"). Additional details about the norming procedure can be found in Trott and Bergen (2021); note that some of the descriptive statistics will differ from those presented here, given that Trott and Bergen (2021) report analyses on the original set of 112 words.

Procedure

Participants completed the study online. They were told that they would read a series of sentences, and that some of these sentences would make sense, while others would not. Their task was to determine which sentences made sense and which did not; they were told to indicate this via button press (*m* for "makes sense," and *x* for "does not make sense"). The instructions encouraged participants to complete each trial as accurately and quickly as possible. Before beginning the primary experiment, participants completed ten practice trials (five sentence pairs). After each trial, they were given feedback indicating whether their response was correct. We used the default intertrial interval in JsPsych, which is 0 ms.

The primary experiment contained 56 critical sentence pairs, randomly sampled from the list of possible trials. Each sentence pair contained an overlapping word (e.g., "lamb") and sentence frame ("They liked the ___"), with one disambiguating word (e.g., "marinated/friendly"). Fifty-six words were randomly sampled from the set of possible words; then, for each word, two of the corresponding sentences were sampled. A similar process was implemented for sampling 56 filler sentence pairs as well.

On any given trial (i.e., a target sentence), a participant saw a sentence appear in the center of the browser page (e.g., "They liked the marinated lamb"). A reminder of their task instructions appeared below the target sentence ("Does this sentence make sense? X = No; M = Yes"). Prime and target sentences appeared on different pages, with an intertrial interval of 0 ms.

After completing the primary experiment, participants answered several demographic questions, regarding their self-identified gender, age, whether or not they were a native speaker of English, and whether or not they completed the experiment on a mobile device.

The experiment was implemented in JsPsych, Version 6.0.5 (de Leeuw, 2015).

Results

All analyses described below were conducted in R Version 3.6.3 (R Core Team, 2020). Mixed effects models were constructed using *lmer* (for Reaction Time data) and *glmer* (for Correct Response data) commands from the *lme4* package (Bates et al., 2015). Random effects structure was determined by beginning with the maximal model, then reducing random effects as needed for model

convergence (Barr et al., 2013); in this case, all models contained by-subject random slopes for the effects of cosine distance, sense boundary, and ambiguity type, as well as random intercepts for subjects and items.

All models also contained the following covariates relating to the target word: concreteness, log frequency, part-of-speech, and length (number of characters). Nested models were compared using log-likelihood ratio tests.

Each explanatory variable of interest (e.g., cosine distance) was used in two separate analyses, to predict either reaction time (RT) or correct response (correct vs. incorrect); thus, we corrected for multiple comparisons using the Holm–Bonferroni method (Holm, 1979). Only adjusted p values are reported below.

All analyses were of the target trial (i.e., the second sentence in each sentence pair). Analyses of RT included only correct responses.

Planned analyses were preregistered on OSF (https://osf.io/gj48a); all exploratory analyses are marked as such in a separate section.

Planned Analyses

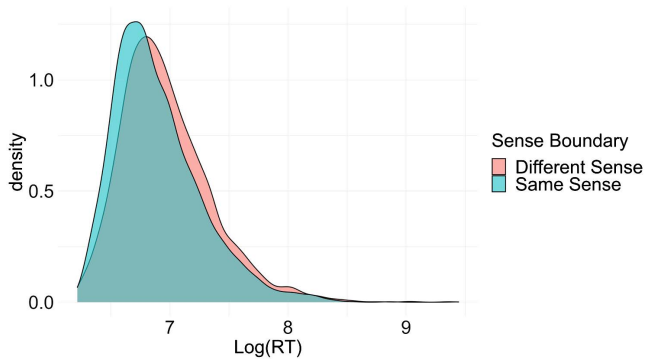
First, we compared a model with fixed effects for sense boundary and cosine distance to a model omitting only the fixed effect of sense boundary. The full model had significantly better fit than the reduced model for both accuracy ($\chi^2[1] = 77.17$, $p < .001$) and RT ($\chi^2[1] = 34.43$, $p < .001$). This disconfirms the prediction of pure exemplar theory; even after adjusting for continuous differences in a word's context of use, the existence of a sense boundary explained additional variance in how accurately and quickly participants responded to the target item. Subjects were more likely to respond correctly to same-sense items (89.3%) than different-sense items (79.9%). Response times were also faster for same-sense ($M = 1{,}068$, $SD = 542$) than different-sense ($M = 1{,}159$, $SD = 598$) items (see also Figure 3).

Second, we constructed a full model including fixed effects of sense boundary, cosine distance, and ambiguity type, as well as an interaction between ambiguity type and sense boundary. The full model explained significantly more variance in RT than the same model without cosine distance ($\chi^2[1] = 15.42$, $p < .001$); larger cosine distances were associated with longer response times ($\beta = 0.14$, $SE = 0.04$). This disconfirms predictions of both theories falling under the Mental Dictionary Framework, that is, both the core representation and sense enumeration accounts; continuous gradations in a word's context of use predicted behavior above and beyond sense boundary and ambiguity type. There was no significant effect of cosine distance on accuracy ($p > .2$).

Finally, we compared the full model to a model omitting only the interaction between ambiguity type and sense boundary. The full model did not explain significantly more variance for either accuracy ($p = .16$) or RT ($p > .2$). This is inconsistent with hybrid+ meaning theory—at least on this task, there is no evidence that homonymy and polysemy elicit fundamentally different behavior, all other things being equal.

Figure 3

Log Reaction Time for Correct Trials Only, Displayed as a Function of Same-Sense Versus Different-Sense



Note. Different-sense trials resulted in longer response times on average than same-sense trials. RT = response time. See the online article for the color version of this figure.

Exploratory Analyses

We also conducted several exploratory analyses, motivated in part by manual inspection of coefficients on the full models for each dependent variable. One finding revealed by this inspection was an apparent main effect of ambiguity type (homonymy vs. polysemy) on accuracy: Participants were considerably less accurate when responding to homonymous than polysemous items, as evidenced by a significant coefficient for ambiguity type ($\beta = -1.14$, $SE = 0.24$, $p < .001$). Because this effect occurred in both the same-sense and different-sense conditions, it is unlikely to be driven by relative differences in the degree of cross-sense inhibition (or facilitation) across polysemy and homonymy. That is, the main effect cannot be due to priming. Further support for this interpretation comes from inspection of accuracy on the first, unprimed half of each trial, which reveals a similar main effect of ambiguity type (but not, crucially, of sense boundary). Together, this suggests that the main effect of ambiguity type arises because of properties of the sentences themselves—either because participants are less accurate when responding to sentences with homonyms in general, or because these *particular* sentences were not sufficiently controlled for plausibility across homonymy and polysemy.

To account for the latter possibility, we replicated the planned analyses from above, but substituted mean first-trial accuracy (or RT) for each version of each item in place of variables relating to the target word like concreteness or frequency. The goal of including first-trial accuracy (or RT) as a covariate was to account for uncontrolled properties of the stimuli (e.g., perhaps some of the sentences are intrinsically more plausible than others, despite all being sensible sentences of English). In principle, accounting for this variance should allow us to better estimate parameters of interest (e.g., coefficients for cosine distance, sense boundary, and the Ambiguity Type × Sense Boundary interaction) as they relate to the task itself, as opposed to uncontrolled properties of the stimuli. Thus, for each unique sentence, we calculated the mean accuracy (or RT) of participants responding to that sentence when it occurred in the *prime* position (i.e., first-trial). We then added first-trial accuracy (or RT) to the models. The main effect of ambiguity type in the full model was now not significantly different from zero ($p > .5$ for both accuracy and RT). Importantly, however, the main effect of sense boundary was preserved for both accuracy ($\chi^2[1] = 84.8$, $p < .001$), and RT ($\chi^2[1] = 36.57$, $p < .001$), as was the main effect of cosine distance on RT ($\chi^2[1] = 17.73$, $p < .001$). Again, there was no significant interaction between ambiguity type and sense boundary for either RT or accuracy ($ps > .2$).

While this does not answer directly the question of why homonymous sentences had lower accuracy rates than polysemous sentences overall, it does suggest a method for directly accounting for any uncontrolled differences in the stimuli.[5] This approach has the advantage of more directly adjusting for any variance due to features intrinsic of the individual sentence in question; differences in first-trial accuracy or first-trial RT are not plausibly attributed to the structure of the task—given that the target ambiguous word has not been directly primed or inhibited by a previous use—and instead, reflect processing difficulties relating to the sentence itself. Thus, in Experiment 2, we sought to replicate the findings reported in Experiment 1 using this refined analysis.

Experiment 2

In Experiment 1, we found that behavior was predicted both by cosine distance and sense boundary, but not by the interaction between sense boundary and ambiguity type. However, there was a main effect of ambiguity type: accuracy was lower for homonymous than polysemous sentences. This main effect could have arisen from uncontrolled properties of the stimuli—indeed, when we controlled for the first-trial accuracy of each item, the main effect of ambiguity type disappeared, but the main effects of sense boundary and cosine distance were preserved. However, this analysis was exploratory. Thus, the primary goal of Experiment 2 was replicating the main findings of Experiment 1 while preregistering this new analysis (https://osf.io/4ej6t).

Method

Participants

As in Experiment 1, we aimed to collect data from 180 participants. Rather than try to estimate the rate of exclusion ahead of time, we iteratively collected data in batches and applied the exclusion criteria to each batch until *at least 180* included participants were reached.

Subjects were recruited through the UC San Diego Psychology Department Subject Pool. When we finished collecting data, there were 239 subjects in the final pool, with 187 remaining after applying the exclusion criteria. Of the final 187 participants, 129 self-reported as female (53 male, 2 nonbinary, and 3 preferred not to answer). The average age was 20.4 ($SD = 2.04$), and ranged from 18 to 32.

Materials and Procedure

The materials used and experimental design were identical to Experiment 1.

Results

Planned Analyses

The analyses were identical to those carried out in Experiment 1, except that the lexical statistics of the target word (e.g., concreteness or log frequency) were replaced by the average first-trial accuracy (or RT) for the target sentence.

As demonstrated by model comparisons, responses were both slower ($\chi^2[1] = 45.57$, $p < .001$) and less accurate ($\chi^2[1] = 96.96$, $p < .001$) for different sense uses. This disconfirms the predictions of pure exemplar theory. Further, as in Experiment 1, we found a significant effect of cosine distance above and beyond sense boundary and ambiguity type (and their interaction), when predicting RT ($\chi^2[1] = 39.64$, $p < .001$) but not Accuracy ($p > .2$). Finally, we detected no significant interaction between ambiguity type and sense boundary for either accuracy or RT ($ps > .2$). Altogether, these results are most consistent with hybrid meaning theory: Behavior was correlated with both the existence of a sense boundary and by the distance between the prime and target context, but this effect did not extend to the difference between polysemes and homonyms.

Exploratory Analyses

The analyses above, precisely like the results from Experiment 1, are most consistent with hybrid meaning theory. However, as noted in the Introduction, there are multiple mechanisms by which sense categories could be implemented in a continuous space. In the sense attraction account, distances in usage space are *reduced* for within-sense tokens, and *exaggerated* for tokens that span a sense boundary. Crucially, within-cluster variance is not eliminated entirely—it is merely reduced. In contrast, the sense distillation account claims that within-cluster variance is entirely distilled into a single point, that is, the centroid of that cluster. The metric properties of the underlying continuous space are preserved across-sense clusters, but within-cluster variance is removed.

These accounts make testable predictions about whether, and how, cosine distance is related to processing ease for same-sense items. Specifically, the sense attraction account predicts that even for same-sense items, reaction time should increase as a function of cosine distance (as it does when all items are considered). However, because the sense distillation account claims that within-cluster variance is removed entirely, it predicts that cosine distance should not be systematically related to reaction time.

We tested these accounts by building a linear mixed-effects model with Log RT as a dependent variable, fixed effects of both cosine distance and ambiguity type, by-subject random slopes for both cosine distance and ambiguity type, and random intercepts for subjects and items. This model was constructed for same-sense pairs only. This model explained significantly more variance than a model omitting cosine distance alone ($\chi^2[1] = 11.02$, $p = .001$), indicating that even within same-sense pairs, cosine distance was positively correlated with RT ($\beta = .17$, $SE = 0.05$). This is inconsistent with the predictions of sense distillation account, which predicts no difference in RT within same-sense pairs.

Discussion

Combined with Experiment 1, these results are inconsistent with three of the five accounts under investigation. As noted earlier, each account made specific predictions about which variables should or should not influence behavior in a primed sensibility judgment task (see Table 1 for a summary). Only the hybrid+ theory could not be strictly disconfirmed, given that it predicts significant effects of all the relevant experimental variables; failing to find a significant effect is not necessarily grounds for rejecting a theory. Nevertheless, a simpler theory that explains the data equally well is still preferable from the standpoint of theoretical parsimony. In this case, that reasoning tips the scales toward the hybrid meaning theory.

[5] See the General Discussion for possible explanations for this result.

To summarize the results, first, we found that the existence of a sense boundary between two contexts of use (e.g., "marinated *lamb*" vs. "friendly *lamb*") resulted in slower response times and less accurate responses overall, as compared to two contexts of use that fall under the same-sense category (e.g., "marinated *lamb*" vs. "roast *lamb*"). This replicates the sense consistency effect obtained in past work, using both identical task paradigms (Klein & Murphy, 2001; Yurchenko et al., 2020) and alternative approaches (Klein & Murphy, 2002). Importantly, this effect held even after controlling for contextual distance, which is inconsistent with the prediction of pure exemplar theory. That is, behavior on this task can be better explained by positing some form of categorical sense representation above and beyond the distance between two contexts of use.

Second, we found that response times were systematically longer for larger contextual distances, as measured by the cosine distance between BERT's contextualized representations of the ambiguous target word, when controlling for sense boundaries. This disconfirms predictions of both accounts falling under the Mental Dictionary Framework (i.e., the core representation and sense enumeration accounts), neither of which allow for graded effects of context: Behavior on this task varied not only as a function of discrete sense representations, but rather was related to a measure that captures the context-dependent nature of word meaning. To our knowledge, this is the first empirical demonstration that online processing difficulty of ambiguous words can be explained by a continuous measure of contextual distance, above and beyond discrete variables like sense boundary.

This leaves the two hybrid theories: hybrid versus hybrid+. The former predicts no difference in behavior across polysemous and homonymous words, while the latter does. Crucially, we failed to detect a difference in how people processed different-sense polysemous meanings and different-sense homonymous meanings, after controlling for differences in first-trial accuracy or response time. Although we cannot strictly reject the hybrid+ theory—absence of evidence does not entail evidence of absence—this does suggest that the hybrid meaning theory is a more parsimonious explanation of the data from both experiments.[6] According to this theory, the clusters in context space that arise as a function of purely distributional properties of language use are systematically "warped" in psychological space, as in the categorical perception of speech (Goldstone & Hendrickson, 2010), according to sense boundaries. Further, a post hoc analysis of the data from both experiments found that variance in contextual distance within same-sense words (i.e., within a sense category) was also predictive of reaction time. This suggests that of the two compression mechanisms explored in the Introduction (sense attraction vs. sense distillation), sense attraction is a better explanation of the data.

This result raises a number of questions about the nature of these sense representations. How exactly does *contextual distance* map onto *conceptual distance*? Which functional transformation best accounts for the behavioral data, and what are the parameters underlying this transformation? These questions are explored in the section below.

Hybrid Meaning Theory: A Further Test and Computational Model

Above, we concluded that hybrid meaning theory—and the sense attraction mechanism in particular—was the best explanation of the behavioral data. This theory claims that distance in *context space* is systematically warped by the existence of sense boundaries, such that within-sense distances are reduced, and across-sense distances are amplified.

One potential objection to this conclusion is that BERT's representation of the context space is not analogous to that of human participants. In principle, it is possible that human meaning representations are completely continuous (as predicted by pure exemplar theory)—and even derived from distributional statistics alone—but that the topology of this representational space is distinct from BERT's, for reasons other than the existence of putative sense boundaries. If this interpretation is correct, BERT's representational space already contains sufficient information to account for human behavior on the priming task described above, provided it is transformed in the appropriate way.

Critically, to be consistent with pure exemplar theory, such a transformation must be *bottom-up*: that is, it must not depend on information extrinsic to what is observable via a word's pattern of use. On the other hand, if hybrid meaning theory is correct, no bottom-up transformation to the underlying BERT-space will be sufficient to account for human sense knowledge. Instead, contextual distance must be transformed using auxiliary information about the existence of a sense boundary, for example, as defined by a lexicographer. For example, contextual distance could be systematically transformed according to whether the two contexts of use straddle a sense boundary or not. For contrast, we call this latter type of transformation "top-down" and include it in the modeling experiments below primarily as a baseline for the bottom-up transformations.

In the current section, we asked whether a top-down or bottom-up transformation to cosine distance improved the fit of a model predicting human behavior on the primed sensibility judgment task. Specifically, we compared the success of several bottom-up transformations to top-down transformations relying on the value of the sense boundary parameter (i.e., same-sense vs. different-sense). As a second-order question, we also considered two distinct functions to apply to cosine distance (for both bottom-up and top-down transformations): (a) an additive function, which increased or decreased cosine distance as a function of sense boundary (or the induced cutoff parameter); and (b) a multiplicative function, which scaled with the original value of cosine distance. The point of identifying the cutoff parameter (in the bottom-up versions) was to compare a sense boundary obtained from a dictionary to a parameter derived from the actual value of cosine distance. Both functions, as well as the procedure for identifying the optimal parameters for each transformation, are described in more detail in the Methods section below.

Once the parameters for each transformation were identified, we asked which transformation best predicted human behavior on Experiments 1–2. The best transformation was then selected using Akaike information criterion (AIC), a measure of model fit (Akaike, 1974; Burnham & Anderson, 2002). That is, we compared the predictive power of a series of statistical models, each containing a specific implementation of transformed distance.

Functional Transformations

The first functional transformation was additive and top-down. That is, it assumed a fixed mapping between contextual distance and

[6] The question of whether homonymy is truly just a form of "distant" polysemy is further explored in the General Discussion.

conceptual distance, according to whether or not two contexts of use were separated by a lexicographer-classified sense boundary. This mapping can be described as follows:

$$\text{Semantic Distance} = \left\{ \begin{array}{l} \text{same} = 1 : x - \beta_1 \\ \text{same} = 0 : x + \beta_1 \end{array} \right\} \quad (1)$$

If two contexts of use correspond to the same-sense, this function decreases conceptual distance by a fixed amount[7] (β_1); if two contexts of use correspond to distinct senses, this function increases conceptual distance by a fixed amount (β_1). The "bottom-up" version of this transformation is identical, but uses an optimized cutoff parameter instead of sense boundary:

$$\text{Semantic Distance} = \left\{ \begin{array}{l} x \leq c : x - \beta_1 \\ x > c : x + \beta_1 \end{array} \right\} \quad (2)$$

The second functional transformation was still linear, but no longer applied a fixed transformation to a given value of cosine distance. Rather, transformed distance was scaled proportionally to the original value of cosine distance: for same-sense pairs, more distant pairs were "attracted" more relative to closer pairs; for different-sense pairs, closer pairs were "repelled" more relative to already distant pairs. This was based on research suggesting that certain category effects are particularly large near category boundaries (Kuhl, 1991), and that cocategorized exemplars undergo a larger perceptual transformation when they are further apart (Kuhl, 1991). This mapping can be described as follows:

$$\text{Semantic Distance} = \left\{ \begin{array}{l} \text{same} = 1 : \frac{x}{\beta_1} \\ \text{same} = 0 : x + \frac{1-x}{\beta_2} \end{array} \right\} \quad (3)$$

That is, the contextual distance of same-sense pairs is divided by a fixed amount (β_1), which results in a proportionately larger transformation to distant pairs than close pairs. Conversely, the contextual distance of different-sense pairs is increased by an amount that decreases as cosine distance increases—different-sense pairs that are already very distant (i.e., close to 1) will be adjusted less than different-sense pairs that are very close (i.e., close to 0). As with (1), the bottom-up version of this transformation uses an optimized cutoff parameter instead of the sense boundary variable:

$$\text{Semantic Distance} = \left\{ \begin{array}{l} x \leq c : \frac{x}{\beta_1} \\ x > c : x + \frac{1-x}{\beta_2} \end{array} \right\} \quad (4)$$

For each transformation, we performed a grid search over a constrained parameter space to identify the optimal set of parameters that would best approximate relatedness. For the additive transformation, we considered values of β_1 ranging from a lower bound of 0 (i.e., no transformation) to an upper bound of 1. For the multiplicative transformation, we considered parameter values ranging from [.1, 15] for both β_1 and β_2. For the bottom-up versions of each transformation, we considered cutoff parameters between [0, 1].

Parameter Optimization

To determine the optimal values of each parameter for each functional transformation, we sought to optimize the strength of the relationship between transformed distance and human relatedness judgments. Past work (Trott & Bergen, 2021) has found that although cosine distance is strongly correlated with relatedness ($\rho = -.58$), it underperforms human interannotator agreement by a considerable margin ($\rho = -0.79$); further, cosine distance systematically *underestimates* human relatedness judgments of same-sense pairs, and *overestimates* the relatedness of different-sense pairs. Thus, we used a grid search to identify the parameters for each transformation that optimized the correlation strength between transformed distance and mean relatedness.

The optimal parameters and resulting correlations between mean relatedness and transformed distance are included in Table 2, and the transformations themselves are depicted in the figures below (See Figures 4 and 5).

Model Specification and Evaluation

Our primary goal was to identify the transformation that best predicted human behavior on Experiments 1–2. To this end, we compared a series of models with distinct parameterizations, predicting both correct response and RT. Recall that in both Experiments 1–2, cosine distance did not improve model fit when predicting correct response, but it did explain independent variance in RT; sense boundary explained variance in both dependent variables. There were four transformed models total, accounting for the transformation itself (additive vs. multiplicative) and the implementation (bottom-up vs. top-down). All models included the same random effects structure and differed only in which fixed effects were added.

1. Transformed Distance (Additive):
   a. D-Add-BU
   b. D-Add-TD
2. Transformed Distance (Multiplicative):
   a. D-Mul-BU
   b. D-Mul-TD
3. Original Cosine Distance: D
4. Sense Boundary: SB
5. A model containing both (3) and (4): D + SB
6. A model containing an interaction between D and SB, along with their main effects.

The top-down additive and multiplicative models (D-Add-TD and D-Mul-TD) represent hypothesized implementations of the sense attraction account, that is, distinct mechanisms by which within-sense distance is reduced and across-sense distance is increased. In this sense, they are each examples of a "hybrid" model. Thus, to the extent that cosine distance and sense boundary each explain unique variance in behavior, as they do for reaction time, these hybrid models should improve upon models with only cosine distance (D) or sense boundary (SB). Their bottom-up counterparts are included to test whether equivalent transformations to cosine distance *without* the use of extrinsic information (i.e., a sense boundary) would suffice.

[7] Using a single same term (β_1) produces the same optimal solution as using distinct terms (β_1, β_2) for same and different-sense pairs.

Table 2
Final Parameters for Each Transformation, Including Both Bottom-Up (BU) and Top-Down (TD) Implementations

Transformation	Parameters	Pearson's *r*
Additive (BU)	$\beta_1 = 0.2$, $C = 0.2$	−0.59
Multiplicative (BU)	$\beta_1 = 0.6$, $\beta_2 = 14.6$, $C = 0.5$	−0.62
Additive (TD)	$\beta_1 = 0.4$	−0.76
Multiplicative (TD)	$\beta_1 = 10.6$, $\beta_1 = 3.6$	−0.77

Note. The table also includes the measure of correlation (Pearson's *r*) between the transformed distance measure and relatedness (Trott & Bergen, 2021).

Model D + SB is another example of a "hybrid" model, which is agnostic to the particular transformation applied to cosine distance, but which simply accounts for both sense boundary and contextual distance using distinct parameters. If D + SB is superior to both of the models with transformed distance, it suggests that neither functional transformation is sufficient to capture the underlying psychological transformation. But if either D-Add or D-Mul improves upon D + SB, it suggests that the corresponding functional transformation is, in fact, a good approximation of the true mapping between contextual distance and conceptual distance. Finally, we considered a model with an interaction between cosine distance and sense boundary (D × SB). This model can be seen as a superset of the multiplicative transformations, since it allows for a different slope of the effect of cosine distance for same-sense versus different-sense pairs.

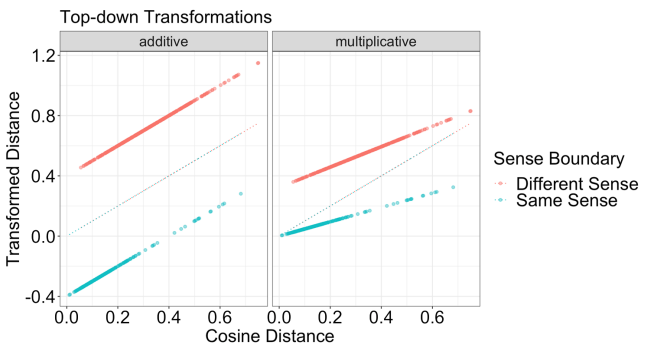
Further, because there are both top-down and bottom-up implementations of D-Add and D-Mul, we can ask whether—and to what degree—an explicit, supervised transformation improves upon one that simply warps cosine distance according to some cutoff parameter. If the top-down transformations do not represent an improvement, it suggests that the relevant information to form human-like sense boundaries is already captured by the distributional regularities of language use—that is, the transformation does not require information *external* to contextual distance (as measured by BERT). Importantly, this outcome would be consistent with pure exemplar theory: Human lexical knowledge can be explained using information present in the distributional statistics of linguistic input alone. But if the top-down transformations do improve upon the bottom-up ones, it suggests that other sources of information, or other manners of representation, are necessary to account for human behavior.

We then calculated the AIC for each model. AIC is a measure of model fit and is defined as:

$$\mathrm{AIC} = 2k - 2\ln(L) \tag{5}$$

where k is the number of parameters in the model, and L is the likelihood of the model. Models with better fit will have higher values of L, and thus lower AIC values overall. As is standard practice (Burnham & Anderson, 2002; Burnham et al., 2011), we rescaled each value of AIC by subtracting the AIC of the best model (i.e., the one with the lowest AIC) of that model set.

Figure 4
Final Result of Top-Down Transformations to Cosine Distance



Note. Different functional transformations are applied to cosine distance as a function of sense boundary. The dotted line reflects the line of identity, that is, if the transformed distance was identical to the original cosine distance measure. Colors reflect the original categorization of a given context pair as belonging to the same or different-sense. See the online article for the color version of this figure.

Results

Predicting Response Time

First, we considered the distribution of AIC values across models for predicting RT on the target sentence, aggregated across Experiments 1–2.

The best model (i.e., the one with the lowest AIC) is the model containing both of the original predictors (cosine distance and sense boundary), followed by the model containing their interaction; presumably, the interaction does not substantially improve model fit, and is penalized for adding an extra parameter.

None of the transformations considered were sufficient to account for the information provided by cosine distance and sense boundary. On the other hand, the top-down implementations of the additive and multiplicative transformations represented a substantive improvement over cosine distance alone, as well as sense boundary. This is not entirely surprising, given that the transformed variables explicitly incorporate both cosine distance and sense boundary, systematically adjusting the former as a function of the latter. Of the two transformations, the simpler additive transformation resulted in a lower AIC than the multiplicative transformation (See Figure 6).

Finally, the bottom-up implementations of both transformations actually performed *worse* than cosine distance alone. This is more surprising, given that they were optimized to improve the correlation between cosine distance and mean relatedness.

Predicting Accuracy

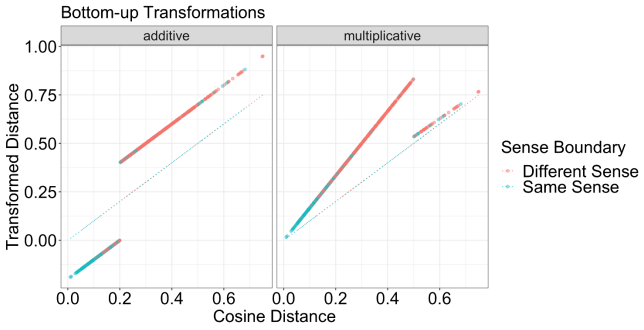
Second, we asked how well the transformed versions of cosine distance predicted correct response. Recall that in Experiments 1–2, a fixed effect of cosine distance did not improve model fit above a model containing only sense boundary.

In this case, the original measure of cosine distance performed the worst, followed by the bottom-up (BU) transformations; unlike with RT, the bottom-up transformations did represent an improvement upon the original cosine distance measure (See Figure 7).

The best model was the one containing only sense boundary. Again, this is not surprising, given that the addition of cosine distance did not improve model fit for Experiments 1–2. Of the top-down (TD) transformations, the additive transformation resulted

Figure 5

Final Result of Bottom-Up Transformations to Cosine Distance



Note. Here, distinct transformations are applied according to some cutoff parameter, as opposed to the value of sense boundary variable; the cutoff refers to the value of cosine distance at which to apply one transformation or the other, and stands in for a "bottom-up" or induced value of sense boundary. Note that in the case of the multiplicative transformation, this results in a surprising transformation: because some different-sense pairs are grouped under the cutoff value, there is a pressure to increase the distance of those pairs even more, to differentiate them from the same-sense pairs also grouped under the cutoff value. The dotted line reflects the line of identity, that is, if the transformed distance was identical to the original cosine distance measure. Colors reflect the original categorization of a given context pair as belonging to the same or different-sense (note that because this was a bottom-up transformation, the same/different-sense information was not used to transform cosine distance). See the online article for the color version of this figure.

in a slightly lower AIC, but this difference was quite small (~0.53), considering the differences between other models.

Discussion

In this section, we attempted to formalize and compare different implementations of hybrid meaning theory. This theory claims that distance in context space is systematically warped in conceptual space by the existence of sense boundaries, such that within-sense distance is reduced and across-sense distance is increased.

We considered two high-level questions. First, what information is required to account for the effect of sense boundaries? Can these effects be simulated by applying a bottom-up transformation to cosine distance, or does a successful approximation require some top-down, external source of information? And second, which functional transformation (i.e., additive vs. multiplicative) results in a parameter that best predicts human behavior?

We addressed the first question by comparing a top-down and bottom-up version of each transformation. The key difference was that the top-down transformations explicitly relied on the value of sense boundary (i.e., same vs. different-sense), while the bottom-up transformations induced an optimal "cutoff" parameter to apply to cosine distance. Models equipped with the top-down transformations consistently outperformed those using the bottom-up transformations, as measured by a lower AIC value. This pattern held across both dependent variables (correct response and response time) and both types of transformation (additive vs. multiplicative). This suggests that distributional statistics alone, at least as operationalized by certain state-of the-art NLMs, are insufficient to account for the effect of sense categories. Rather, an explanatory theory must posit that sense category structure is derived from some source of information or representation that goes beyond linguistic co-occurrence statistics; plausible candidates are explored in the General Discussion.

We addressed the second question by comparing two functional transformations. The additive transformation was intended to model a main effect of sense boundary: Within-sense distance was reduced by some fixed amount, and across-sense distance was increased by that same amount. In contrast, the multiplicative transformation allowed the magnitude of a given transformation to vary with the original distance in context space: Distant same-sense pairs were attracted more than pairs that were already close together; and nearby different-sense pairs were repelled more than pairs that were already distant. When predicting RT, the top-down additive transformation was better than the top-down multiplicative transformation; this was also true when predicting correct response, but the difference in predictive power was comparatively very small.

General Discussion

We began with the question of how humans store and represent the meanings of ambiguous words. Traditional theories fall under the Mental Dictionary Framework, with discrete entries corresponding to each meaning of a wordforms. In contrast, the Continuity of Meaning Framework views word meaning as trajectories in a continuous, context-dependent state space (Elman, 2004; Li & Joanisse, 2021). Some theories falling under this framework (e.g., pure exemplar theory) eschew the notion of discrete meaning representations altogether. In this article, we also introduced two "hybrid" theories, which allow for the possibility of graded, context-sensitive meaning representations, but also posit the existence of mediating categorical representations (see Table 1 for a summary).

Two behavioral experiments provided support for the simpler of these two hybrid theories, which we uncreatively call hybrid meaning theory. Using a primed sensibility judgment paradigm, we found that response time on the target trial was systematically related to the continuous distance between the prime and target contexts, above and beyond the existence of a sense boundary between those contexts. This result is inconsistent with the predictions of either of the theories falling under the Mental Dictionary Framework, at least with respect to this task. Both response time and accuracy were further modulated by the existence of a sense boundary, which is inconsistent with the predictions of pure exemplar theory for this task. We also found no evidence that the size of this effect depended on the kind of ambiguity (i.e., homonymy vs. polysemy), suggesting that these phenomena do not elicit categorically distinct behavior on the task. Altogether, this suggests that hybrid meaning theory accounts best for the behavioral signatures we measured of how humans represent the meaning of ambiguous words.

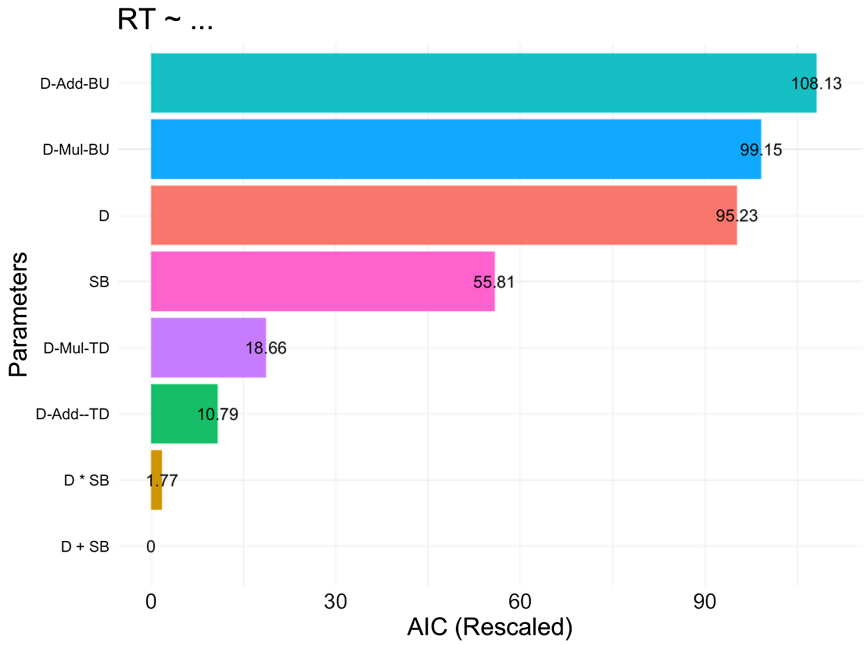
Below, we discuss limitations of the current work, and explore implications for future research.

Accuracy Versus Reaction Time (RT)

As described above, both dependent measures (Accuracy and RT) were predicted by the existence of a sense boundary, but only RT was significantly correlated with cosine distance. Although we

Figure 6

Rescaled AIC for Each of the Models Predicting RT



Note. A lower AIC indicates better fit. The models containing top-down transformations (D-Add-TD and D-Mul-TD) exhibited better fit than those containing only sense boundary (SB) or the original cosine distance variable (D). The bottom-up transformations (D-Add-BU and D-Mul-BU) exhibited the worst fit. RT = response time. See the online article for the color version of this figure.

treated these measures as testing the same hypothesis (hence correcting for multiple comparisons), it is worth exploring potential post hoc explanations for why they would diverge with respect to cosine distance.

Accuracy is a discrete measure (correct or incorrect), reflecting discrete responses on the task (sensible vs. nonsensical). Consequently, it reflects the outcome of imposing a decision threshold on a process that may, at root, be continuous. In cases where the effect of graded context distances (here, cosine distance) is relatively small—they may be detectable only on the process of *arriving* at a decision, but not necessarily the outcome of a decision itself.

In contrast, RT reflects the amount of time required to correctly identify a sentence as plausible. As a more fine-grained measure of the process by which a participant *arrived* at their decision, RT may thus be more suitable for identifying small, continuous effects like that of cosine distance. Indeed, other researchers (Spivey & Dale, 2004, 2006) have pointed out that fine-grained, continuous measures are important for investigating putatively continuous processes. If this is true, it suggests a potential avenue for future work: Researchers might deploy more fine-grained measures (e.g., mouse-tracking, eye-tracking, or electroencephalography) to identify whether and to what extent the mental lexicon exhibits continuity.

Limitations of the Language Model

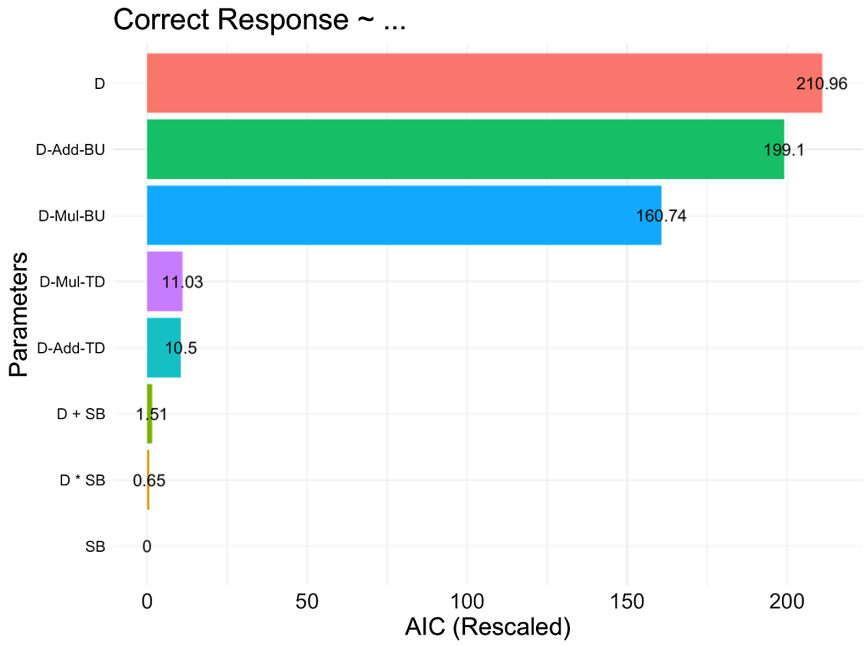
One possible objection to the current work is that BERT represents a poor operationalization of pure exemplar theory and that other language models would be a better choice. This objection might manifest in two different ways: First, that BERT already has representational abstractions and is thus ill-suited to operationalizing an account that eschews sense representations (i.e., it is already too human-like); and second, that BERT is too limited, either in its training data or its architecture (i.e., it is not powerful enough).

First, as others have pointed out (Mahowald et al., 2020), signals elicited from NLMs—for example, surprisal, hidden unit activation, etc.—often covary with psychological or linguistic categories, such as parts of speech, animacy, semantic roles, and more (Tenney et al., 2019); this is sometimes interpreted as reflecting the formation of representational abstractions. If models like BERT are capable of forming abstractions, it is conceivable that sense representations might already be encoded by BERT, in which case BERT would indeed be a poor implementation of a theory that posits no sense representations. However, this objection can be rejected on empirical grounds: In our studies, BERT demonstrably failed to capture variance in human behavior that *was* explained by the existence of a sense boundary. Similarly, bottom-up transformations to cosine distance alone failed to improve model fit above and beyond the effect of sense boundary—the best transformations were "top-down," in that they relied on an external source of information (in this case, human-annotated sense knowledge). Together, these findings empirically demonstrate that even if BERT is capable of forming representational abstractions, these abstractions cannot account for the effect of human sense knowledge.

A second, alternative objection is that BERT is not sufficiently powerful. NLMs are evolving rapidly—models like Generative Pretrained Transformer 3 already surpass BERT on a number of

Figure 7

Rescaled AIC of the Models Predicting Correct Response



Note. A lower AIC indicates better fit. As with RT, the models with the top-down transformations (D-Mul-TD and D-Add-TD) exhibited better fit than those with the bottom-up transformation (D-Mul-BU and D-Add-BU), though in this case, these were not as successful as a model with sense boundary (SB) alone. RT = response time. See the online article for the color version of this figure.

metrics (Brown et al., 2020), and increases in computing resources will likely yield even better models in years to come (Kaplan et al., 2020). Thus, our "best guess" for how much information can be extracted from linguistic context alone may change as well; it is possible that a future generation of NLMs *will* display something equivalent to human sense knowledge. Importantly, however, improvements in performance along some dimensions (e.g., perplexity) do not always entail better predictions of human behavior on other tasks (e.g., reading time or eye-tracking; Kuribayashi et al., 2021). This suggests that even as models improve at the tasks they are designed to do (e.g., masked word prediction), they may continue to diverge from humans in important, cognitively relevant ways. As an analogy, the best computer-chess programs now outperform virtually all humans, but that does not entail that these programs play chess in a human-like way.

Additionally, as noted in Supplemental Material Analysis 1, we did replicate our analyses with ELMo, another language model (Peters et al., 2018) that typically underperforms BERT on word sense disambiguation tasks (Wiedemann et al., 2019), and with BERT-large, which contains twice as many layers and many more parameters. Interestingly, although BERT-large produced better predictions of human relatedness judgments, it was a worse predictor (as measured by AIC) of reaction time than BERT-base. This reinforces the point that improvement on one task does not entail improvement across the board—so it is no guarantee that future language models will acquire human-like sense knowledge in the absence of other methodological interventions intended to render them more human-like (i.e., fine-tuning them to a word sense disambiguation data set, or incorporating grounding into their training regime). Further, in Supplemental Material Analysis 3, we asked whether a different metric (surprisal) explained more variance than the preregistered measure (cosine distance); while surprisal was indeed predictive of behavior, it did not eliminate the explanatory power of sense boundaries, consistent with the predictions of hybrid meaning theory. Finally, in the Computational Modeling section, we explored several bottom-up transformations to cosine distance, all of which suggest that distributional statistics alone are insufficient to account for the category effects of sense boundaries.

On this note, it is worth reiterating that BERT represents a particular implementation of pure exemplar theory—that is, one in which continuous meaning representations are derived from distributional regularities in linguistic input alone. BERT (and most other NLMs) lack extralinguistic grounding (Bender & Koller, 2020; Lake & Murphy, 2021). Thus, any semantic knowledge that relies on extralinguistic information (e.g., perceptual experience) will be inaccessible to BERT. While this limits BERT's predictive power, it also offers a useful inferential tool: models like BERT help establish empirical limits on how much human linguistic knowledge can be captured from distributional regularities alone (see Supplemental Material Analysis 4, for an additional analysis along these lines). As Elman (2011) notes, a continuous meaning space could be constituted by many different dimensions of experience, including the sensorimotor or even social associations with individual words and constructions. Accordingly, future work could make use of recent developments in grounded language models

(Johns, 2021; Su et al., 2019; Zellers et al., 2021) to ask whether access to particular dimensions of sensorimotor information elicits more human-like behavior. One possible outcome is that the variance in human behavior currently explained by sense boundaries can actually be attributed to aspects of sensorimotor or social experience uncaptured by BERT. Under one interpretation, this would salvage a version of pure exemplar theory, which simply admits more dimensions of human experience into the continuous state space. Note, however, that this possibility is tested in Supplemental Material Analysis 4; the results suggest that even the inclusion of a new continuous measure, derived from similarity in sensorimotor space, cannot explain away the categorical effect of sense boundary. A related objection is that our operationalization measures only the difference between two specific contexts, as opposed to considering the entire distribution of usages of a particular wordform; future work could explore whether additional explanatory power is obtained by comparing clusters of usages, as opposed to specific contexts.

Is Homonymy Just "Distant" Polysemy?

We found no evidence that homonymous meanings exhibited different priming effects than polysemous meanings. This is somewhat surprising, given the extensive evidence that the two phenomena elicit systematically different behavior on a number of tasks (Floyd & Goldberg, 2021; Klepousniotou, 2002; Klepousniotou & Baum, 2007; Klepousniotou et al., 2012; Rodd et al., 2002, 2012), and the fact that they are typically treated as distinct phenomena in theoretical linguistics and lexicography (Valera, 2020). Some past work has nevertheless acknowledged the possibility of homonymy and polysemy lying along a continuum, both in theoretical cognitive linguistics (Tuggy, 1993) and experimental psycholinguistics (Brown, 2008; Klein & Murphy, 2002; Klepousniotou et al., 2008; Rodd et al., 2002). As noted in the Introduction, however, the majority of work in this area has not incorporated this notion of a continuum between homonymy and polysemy into theoretical or formal models of lexical ambiguity (with some exceptions, e.g., Rodd, 2020).

Does this mean that homonymy is simply "distant" polysemy—at least when it comes to their respective impacts on cognitive processing? There are several possible reasons why it does not. First, we might simply have failed to detect a real, nonzero difference between polysemy and homonymy (e.g., because the study was underpowered). On the other hand, while there is always a real possibility of a false negative result, we obtained null results across two large N studies ($N \geq 180$); further, a power analysis suggested that we should have had 95% power to detect an effect of the size we detected in a pilot study. Combined, this suggests that the behavioral differences in this paradigm are either nonexistent, or small enough to be of negligible theoretical interest.

Second, it is possible that our operationalization of homonymy and polysemy—that is, determining whether two meanings were listed as separate entries in the dictionary—was somehow deficient. However, it is unclear how better to operationalize these variables. Binning according to some behavioral variable (e.g., relatedness) would impose semiarbitrary structure on a continuous space, which is precisely the question we are attempting to address. The expertise of lexicographers for Merriam Webster and the Oxford English Dictionary may be the closest approximation to the received expert view that can be found. Nevertheless, it is possible that another operationalization, perhaps relying on finer grained distinctions between semantic relations (e.g., metaphor vs. metonymy), could result in the detection of behavioral differences across categories of ambiguity.

Third, our original, preregistered analyses did not account for sense dominance (i.e., when one meaning of an ambiguous word is more frequent than another), which is known to influence ease of processing (Blott et al., 2021; Duffy et al., 1988; Klepousniotou et al., 2008). However, we counterbalanced the order of the prime and target sentences across participants. Thus, if many of the sentence pairs contained unbalanced meanings, our results would essentially be averaging across a null or small effect (i.e., moving from a subordinate to a dominant sense) and a strong effect (i.e., moving from a dominant to subordinate sense); we believe this is unlikely to account for the failure to find a significant difference in the priming effect across polysemous and homonymous pairs. Additionally, we did run a post hoc analysis using normed dominance judgments for different-sense items only (see Supplemental Material Analysis 2). This analysis replicated the effect of dominance found in past experiments (Klepousniotou et al., 2008), as well as the main effect of cosine distance reported in Experiments 1–2. There was also a possible main effect of ambiguity type for different-sense pairs only—but as noted below, this main effect could be driven by uncontrolled differences among the stimuli themselves and is not necessarily attributable to differences in the strength of priming across homonymous and polysemous stimuli (i.e., we failed to detect a Sense Boundary × Ambiguity Type interaction in both experiments).

Finally, and perhaps most importantly, it is possible that the primed sensibility judgment task is simply not well-suited for detecting a difference between homonymy and polysemy. Similarly, it may be that BERT is unusually good at predicting behavior on this kind of task, but would not do a good job of predicting human behavior on other tasks that require deeper comprehension (Lake & Murphy, 2021). Our failure to detect a difference on one task does not entail that the two phenomena are not psychologically distinct in general. To make this more general claim—that is, that homonymy is "distant" polysemy—one would need to demonstrate a null effect across a number of tasks that have provided evidence for a categorical difference between polysemy and homonymy. If, by process of elimination, each task fails to elicit behavioral differences above and beyond the continuous distance between two contexts of use, one might at last conclude that homonymy and polysemy truly do lie along a continuum; if, on the other hand, some tasks *do* continue to elicit different behavior, that would provide deeper insight into exactly when and under what conditions this categorical distinction is cognitively and behaviorally relevant.

Here, it is worth revisiting the finding that accuracy did significantly differ across sentences containing homonymous and polysemous sentences, on *both* prime and target trials. Since the size of this effect did not differ across prime and target trials, this indicates that there was no difference in the priming effect itself. There are several possible explanations for this main effect. First, it could be due to uncontrolled differences in the stimuli: Perhaps, the sentences containing homonyms, or the homonymous items themselves, happened to be less natural than those containing polysemes. Although we adjusted for a number of features in our analyses (e.g., frequency, length, concreteness), it is possible that we failed to account for a crucial determinant of lexical processing. Second, the

effect could be driven by a theoretically meaningful difference in how homonymous and polysemous words are processed. Past work (Klepousniotou, 2002; Rodd et al., 2002) has found differences in reaction time and accuracy on isolated lexical decision tasks. If accessing the meaning of a homonym involves competition from its other, unrelated meanings (Rodd et al., 2002), then sentences containing homonyms might also be genuinely harder to process than those with polysemes, even independent of priming.

Sense Representations in a Continuous State Space

The experimental results reported above support hybrid meaning theory, which claims: (a) word meanings are context-dependent trajectories through a continuous state space; and (b) these trajectories are mediated by sense representations, such that *contextual distance* is transformed into a sense-mediated *conceptual distance*. The second claim raises a number of questions about the nature of these sense representations.

First, there are a number of distinct computational mechanisms by the use of which sense representations might mediate contextual distance. In the Introduction, we distinguished between sense attraction, in which tokens within a sense cluster shrink toward their centroid, and sense distillation, in which within-cluster variance is removed altogether, preserving only the centroid or prototypical member. An exploratory analysis provided evidence in favor of the sense attraction mechanism; even considering only same-sense uses, we found that response time was positively correlated with contextual distance, suggesting that some within-cluster variance is preserved. Further, we applied several functional transformations to cosine distance and asked which transformation yielded improvements in predicting human behavior. We found that an additive transformation to cosine distance best improved a statistical model's fit; crucially, the best transformation was "top-down" and explicitly used the sense boundary variable, that is, information *external* to the underlying BERT-space. This suggests that distributional regularities alone—even after applying a bottom-up transformation—are insufficient to account for the emergence of sense-like representations.

This leads to a second, related question: How and when do these sense representations emerge? Given that every context of use constitutes a slight variation in meaning, what degree—or what dimensions—of variation results in the creation of a sense boundary? Klein and Murphy (2001, p. 279) summarize the question as follows (emphasis ours):

> If two senses are only very subtly different, it seems unlikely that speakers will develop separate entries for them, since a single entry will suffice to specify most of the meaning for both. If two senses are strikingly different, then a single entry will probably be unsuccessful at representing both meanings, which will presumably lead to the formation of separate entries … *What is needed is a more specific model of what causes a sense to be separately represented, from which one could derive predictions about which uses would involve the same senses and which would involve different senses.*

One promising avenue would be to look to related research on how children acquire ambiguous words (Rabagliati et al., 2010). There is some evidence that children are better able to acquire new meanings for a known wordform when those meanings are related, rather than unrelated, to its existing meanings (Floyd & Goldberg, 2021). This echoes previous findings that homonyms are challenging to learn (Casenhiser, 2005), possibly because children have a bias against assuming homophony—though more recent work (Dautriche et al., 2016) suggests that children reliably postulate homophony if the exemplars presented from each meaning are sufficiently distinct. Finally, work by Srinivasan and Snedeker (2011) suggests that children rely on a common representation for polysemous words with highly regular meaning relationships (e.g., "heavy *book*" and "popular *book*"). As the authors note, this common representation could be lexical, with rules for deriving each meaning stored with the word itself (Pustejovsky, 1995); alternatively, they might rely on more general conceptual knowledge, likely reflecting systematic conceptual relations within and across languages (Srinivasan & Rabagliati, 2015). Yet to our knowledge, it remains unknown whether and when these related meanings drift apart into distinct sense clusters. In our task, English-speaking adults demonstrated an effect of sense boundary above and beyond the distance between two contexts of use, and there was no significant difference in the size of this effect between polysemes and homonyms—suggesting that at least in adults, polysemous meanings manifest in distinct sense clusters. Future work could use a similar paradigm with children, and ask at what age children begin to differentiate highly related polysemous meanings.

A second avenue would be to develop hypotheses about which dimensions of contextual variability are most likely to predict the emergence of a new sense. If word meaning is at least partially grounded in sensorimotor experience (Barsalou, 1999; Bergen, 2015; Pulvermüller, 2013), one possibility is that a new sense cluster is generated when the associated sensorimotor profile is sufficiently distinct. For example, one meaning might be more concrete than the other, as is the case with much of conceptual metaphor (e.g., "a wooden *table*" vs. "a data *table*"). Alternatively, different contexts of use might be similarly concrete, but involve different bodily effectors, different perceptual modalities, or even different instruments. For example, "*cut* the paper" and "*cut* the hair" both typically involve scissors, whereas "*cut* the grass" often involves a lawn mower. If psychological senses are motivated by sensorimotor distinctions, then one would predict that "*cut* the paper" and "*cut* the hair" are more likely to behave as same-sense items, while "*cut* the paper" and "*cut* the grass" should be more likely to behave like different-sense items, all other things being equal. Similarly, the difficulty of transitioning across a sense boundary might be highest when those senses have very different sensorimotor profiles.

Conclusion

Word meaning is highly context-sensitive and often outright ambiguous. Accordingly, mental representations of word meaning must be flexible enough to accommodate this context-sensitivity. However, traditional theoretical frameworks analogize mental representations to entries in a physical dictionary, which are static and discrete; this conceptualization is challenging to reconcile with the flexible, context-dependent nature of word meaning. We reviewed evidence supporting the Continuity of Meaning Framework, in which word meanings are conceptualized as context-sensitive trajectories in a continuous state space; we also introduced two "hybrid" theories, which posit discrete, psychologically real categories atop this continuous space. In two behavioral experiments

using a primed sensibility paradigm, we found that human behavior was best predicted by a theory that posits both continuous, flexible meaning representations as well as discrete senses.

References

Aina, L., Gulordava, K., & Boleda, G. (2019). Putting words in context: LSTM language models and lexical ambiguity. *PsyArXiv*. https://doi.org/10.18653/v1/P19-1324

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Ambridge, B. (2020). Against stored abstractions: A radical exemplar model of language acquisition. *First Language*, *40*(5–6), 509–559. https://doi.org/10.1177/0142723719869731

Armstrong, B. C., & Plaut, D. C. (2008). *Settling dynamics in distributed networks explain task differences in semantic ambiguity effects: Computational and behavioral evidence* [Conference session]. Proceedings of the Annual Meeting of the Cognitive Science Society.

Bambini, V., Bertini, C., Schaeken, W., Stella, A., & Di Russo, F. (2016). Disentangling metaphor from context: An ERP study. *Frontiers in Psychology*, *7*, Article 559. https://doi.org/10.3389/fpsyg.2016.00559

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*(4), 577–660. https://doi.org/10.1017/S0140525X99002149

Bates, B., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Beer, R. D. (2003). The dynamics of active categorical perception in an evolved model agent. *Adaptive Behavior*, *11*(4), 209–243. https://doi.org/10.1177/1059712303114001

Bender, E. M., & Koller, A. (2020, July). *Climbing towards NLU: On meaning, form, and understanding in the age of data* [Conference session]. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.

Bergen, B. (2015). Embodiment, simulation and meaning. In N. Riemer (Ed.), *The Routledge handbook of semantics* (pp. 158–173). Routledge.

Blott, L. M., Rodd, J. M., Ferreira, F., & Warren, J. E. (2021). Recovery from misinterpretations during online sentence processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *47*(6), 968–997.

Boutonnet, B., Dering, B., Viñas-Guasch, N., & Thierry, G. (2013). Seeing objects through the language glass. *Journal of Cognitive Neuroscience*, *25*(10), 1702–1710. https://doi.org/10.1162/jocn_a_00415

Brown, S. (2008). Polysemy in the Mental Lexicon. *Colorado Research in Linguistics*, *21*(1), Article 2. https://doi.org/10.25810/s1d0-gj21

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. *PsyArXiv*.

Burnham, K. P., & Anderson, D. R. (2002). Avoiding Pitfalls when using information-theoretic methods. *Journal of Wildlife Management*, *66*(3), 912–918. https://doi.org/10.2307/3803155

Burnham, K. P., Anderson, D. R., & Huyvaert, K. P. (2011). AIC model selection and multimodel inference in behavioral ecology: Some background, observations, and comparisons. *Behavioral Ecology and Sociobiology*, *65*(1), 23–35. https://doi.org/10.1007/s00265-010-1029-6

Casenhiser, D. M. (2005). Children's resistance to homonymy: An experimental study of pseudohomonyms. *Journal of Child Language*, *32*(2), 319–343. https://doi.org/10.1017/S0305000904006749

Chemero, A. (2011). *Radical embodied cognitive science*. MIT Press.

Cruse, D. A. (1986). *Lexical semantics*. Cambridge University Press.

Dautriche, I. (2015). *Weaving an ambiguous lexicon* (Doctoral dissertation, Sorbonne Paris Cité).

Dautriche, I., Chemla, E., & Christophe, A. (2016). Word learning: Homophony and the distribution of learning exemplars. *Language Learning and Development*, *12*(3), 231–251. https://doi.org/10.1080/15475441.2015.1127163

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*(1), 1–12. https://doi.org/10.3758/s13428-014-0458-y

Deane, P. D. (1988). Polysemy and cognition. *Lingua*, *75*(4), 325–361. https://doi.org/10.1016/0024-3841(88)90009-5

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *PsyArXiv*. https://doi.org/10.48550/arXiv.1810.04805

Duffy, S. A., Morris, R. K., & Rayner, K. (1988). Lexical ambiguity and fixation times in reading. *Journal of Memory and Language*, *27*(4), 429–446. https://doi.org/10.1016/0749-596X(88)90066-6

Elman, J. L. (2004). An alternative view of the mental lexicon. *Trends in Cognitive Sciences*, *8*(7), 301–306. https://doi.org/10.1016/j.tics.2004.05.003

Elman, J. L. (2009). On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive Science*, *33*(4), 547–582. https://doi.org/10.1111/j.1551-6709.2009.01023.x

Elman, J. L. (2011). Lexical knowledge without a lexicon? *The Mental Lexicon*, *6*(1), 1–33. https://doi.org/10.1075/ml.6.1.01elm

Firth, J. R. (1957). A synopsis of linguistic theory, 1930–1955. *Studies in linguistic analysis*. Oxford.

Floyd, S., & Goldberg, A. E. (2021). Children make use of relationships across meanings in word learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *47*(1), 29–44. https://doi.org/10.1037/xlm0000821

Ganong, W. F., III. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, *6*(1), 110–125. https://doi.org/10.1037/0096-1523.6.1.110

Geeraerts, D. (1993). Vagueness's puzzles, polysemy's vagaries. [includes Cognitive Linguistic Bibliography]. *Cognitive Linguistics*, *4*(3), 223–272. https://doi.org/10.1515/cogl.1993.4.3.223

Goldstone, R. L., & Hendrickson, A. T. (2010). Categorical perception. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(1), 69–78. https://doi.org/10.1002/wcs.26

Green, P., & MacLeod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*(4), 493–498. https://doi.org/10.1111/2041-210X.12504

Haber, J., & Poesio, M. (2020a, December). *Assessing polyseme sense similarity through co-predication acceptability and contextualised embedding distance* [Conference session]. Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics, Barcelona, Spain.

Haber, J., & Poesio, M. (2020b, June). *Word sense distance in human similarity judgements and contextualised word embeddings* [Conference session]. Proceedings of the Probability and Meaning Conference.

Hanks, P. (2000). Do word meanings exist?. *Computers and the Humanities*, *34*(1/2), 205–215. https://www.jstor.org/stable/30204810

Harris, Z. S. (1954). Distributional structure. *Word*, *10*(2–3), 146–162. https://doi.org/10.1080/00437956.1954.11659520

Hoffman, P., Lambon Ralph, M. A., & Rogers, T. T. (2013). Semantic diversity: A measure of semantic ambiguity based on variability in the contextual usage of words. *Behavior Research Methods*, *45*(3), 718–730. https://doi.org/10.3758/s13428-012-0278-x

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*(2), 65–70. https://www.jstor.org/stable/4615733

Johns, B. T. (2021). Distributional social semantics: Inferring word meanings from communication patterns. *Cognitive Psychology*, *131*, Article 101441. https://doi.org/10.1016/j.cogpsych.2021.101441

Jurafsky, D. (2014). *The language of food: A linguist reads the menu*. W.W. Norton.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling laws for neural language models. *PsyArXiv*. https://doi.org/10.48550/arXiv.2001.08361

Karidi, T., Zhou, Y., Schneider, N., Abend, O., & Srikumar, V. (2021). Putting words in BERT's mouth: Navigating contextualized vector spaces with pseudowords. *PsyArXiv*. https://doi.org/10.18653/v1/2021.emnlp-main.806

Kawamoto, A. H. (1993). Nonlinear dynamics in the resolution of lexical ambiguity: A parallel distributed processing account. *Journal of Memory and Language*, *32*(4), 474–516. https://doi.org/10.1006/jmla.1993.1026

Kearns, K. (2006). Lexical Semantics. In B. Aarts (Ed.), *The handbook of English linguistics* (p. 557–580). Blackwell Publishing Ltd.

Kempson, R. M. (1977). *Semantic theory*. Cambridge University Press.

Kilgarriff, A. (2007). Word senses. In E. Agirre, & P. Edmonds (Eds.), *Word sense disambiguation* (pp. 29–46). Springer. https://doi.org/10.1007/978-1-4020-4809-8_2

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

Klein, D. E., & Murphy, G. L. (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, E. (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

Klepousniotou, E., & Baum, S. R. (2007). Disambiguating the ambiguity advantage effect in word recognition: An advantage for polysemous but not homonymous words. *Journal of Neurolinguistics*, *20*(1), 1–24. https://doi.org/10.1016/j.jneuroling.2006.02.001

Klepousniotou, E., Pike, G. B., Steinhauer, K., & Gracco, V. (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

Klepousniotou, E., Titone, D., & Romero, C. (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

Krishnamurthy, R., & Nicholls, D. (2000). Peeling an onion: The Lexicographer's experience of manual sense-tagging. *Computers and the Humanities*, *34*(1), 85–97. https://doi.org/10.1023/A:1002407003264

Kuhl, P. K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics*, *50*(2), 93–107. https://doi.org/10.3758/BF03212211

Kuribayashi, T., Oseki, Y., Ito, T., Yoshida, R., Asahara, M., & Inui, K. (2021). Lower perplexity is not always human-like. *PsyArXiv*. https://doi.org/10.18653/v1/2021.acl-long.405

Lake, B. M., & Murphy, G. L. (2021). Word meaning in minds and machines. *Psychological Review*. Advance online publication. https://doi.org/10.1037/rev0000297

Lehrer, A. (1990). Polysemy, conventionality, and the structure of the lexicon. *Cognitive Linguistics*, *1*(2), 207–246. https://doi.org/10.1515/cogl.1990.1.2.207

Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. *Italian Journal of Linguistics*, *20*(1), 1–31.

Li, J., & Joanisse, M. F. (2021). Word senses as clusters of meaning modulations: A computational model of polysemy. *Cognitive Science*, *45*(4), Article e12955. https://doi.org/10.1111/cogs.12955

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, *54*(5), 358–368. https://doi.org/10.1037/h0044417

Lopukhina, A., Laurinavichyute, A., Lopukhin, K., & Dragoy, O. (2018). The mental representation of polysemy across word classes. *Frontiers in Psychology*, *9*, Article 192. https://doi.org/10.3389/fpsyg.2018.00192

Loureiro, D., Rezaee, K., Pilehvar, M. T., & Camacho-Collados, J. (2020). Language models and word sense disambiguation: An overview and analysis. *PsyArXiv*.

Mahowald, K., Kachergis, G., & Frank, M. C. (2020). What counts as an exemplar model, anyway? A commentary on Ambridge (2020). *First Language*, *40*(5–6), 608–611. https://doi.org/10.1177/0142723720905920

Malt, B. C. (1989). An on-line investigation of prototype and exemplar strategies in classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(4), 539–555. https://doi.org/10.1037/0278-7393.15.4.539

Miikkulainen, R., & Elman, J. (1993). *Sub-symbolic natural language processing: An integrated model of scripts, lexicon, and memory*. MIT Press.

Mo, L., Xu, G., Kay, P., & Tan, L. H. (2011). Electrophysiological evidence for the left-lateralized effect of language on preattentive categorical perception of color. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(34), 14026–14030. https://doi.org/10.1073/pnas.1111860108

Nair, S., Srinivasan, M., & Meylan, S. (2020). Contextualized word embeddings encode aspects of human-like word sense knowledge. *PsyArXiv*. https://doi.org/10.48550/arXiv.2010.13057

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

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). *Deep contextualized word representations* [Conference session]. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, Louisiana. https://doi.org/10.18653/v1/N18-1202

Pinker, S. (1997). Words and rules in the human brain. *Nature*, *387*(6633), 547–548. https://doi.org/10.1038/42347

Pulvermüller, F. (2013). Semantic embodiment, disembodiment or misembodiment? In search of meaning in modules and neuron circuits. *Brain and Language*, *127*(1), 86–103. https://doi.org/10.1016/j.bandl.2013.05.015

Pustejovsky, J. (1995). *The generative lexicon*. MIT Press.

Pustejovsky, J. (2002). The generative lexicon. *Language*, *73*(3), 597–600. https://doi.org/10.2307/415891

Pustejovsky, J., & Bouillon, P. (1995). Aspectual coercion and logical polysemy. *Journal of Semantics*, *12*(2), 133–162. https://doi.org/10.1093/jos/12.2.133

Rabagliati, H., Marcus, G. F., & Pylkkänen, L. (2010). Shifting senses in lexical semantic development. *Cognition*, *117*(1), 17–37. https://doi.org/10.1016/j.cognition.2010.06.007

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rice, S. A. (1992). *Polysemy and lexical representation: The case of three English prepositions* [Conference session]. Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society.

Rodd, J., Gaskell, G., & Marslen-wilson, W. (2002). Making sense of semantic ambiguity: Semantic competition in lexical access. *Journal of Memory and Language*, *266*(2), 245–266. https://doi.org/10.1006/jmla.2001.2810

Rodd, J. M. (2020). Settling into semantic space: An ambiguity-focused account of word-meaning access. *Perspectives on Psychological Science*, *15*(2), 411–427. https://doi.org/10.1177/1745691619885860

Rodd, J. M., Berriman, R., Landau, M., Lee, T., Ho, C., Gaskell, M. G., & Davis, M. H. (2012). Learning new meanings for old words: Effects of semantic relatedness. *Memory & Cognition*, *40*(7), 1095–1108. https://doi.org/10.3758/s13421-012-0209-1

Rodd, J. M., Gaskell, M. G., & Marslen-Wilson, W. D. (2004). Modelling the effects of semantic ambiguity in word recognition. *Cognitive Science*, *28*(1), 89–104. https://doi.org/10.1207/s15516709cog2801_4

Schane, S. (2002). Ambiguity and misunderstanding in the law. *Thomas Jefferson Law Review*, *25*(1), 167–193.

Schneider, N., Srikumar, V., Hwang, J. D., & Palmer, M. (2015, June). *A hierarchy with, of, and for preposition supersenses* [Conference session]. Proceedings of the 9th Linguistic Annotation Workshop, Denver, Colorado, United States.

Soler, A. G., & Apidianaki, M. (2021). Let's play mono-poly: BERT can reveal words' polysemy level and partitionability into senses. *Transactions of the Association for Computational Linguistics*, *9*, 825–844. https://doi.org/10.1162/tacl_a_00400

Spivey, M. (2008). *The continuity of mind*. Oxford University Press.

Spivey, M. J., & Dale, R. (2004). On the continuity of mind: Toward a dynamical account of cognition. In B. H. Ross (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 45, pp. 87–142). Elsevier Academic Press.

Spivey, M. J., & Dale, R. (2006). Continuous dynamics in real-time cognition. *Current Directions in Psychological Science*, *15*(5), 207–211. https://doi.org/10.1111/j.1467-8721.2006.00437.x

Srinivasan, M., & Rabagliati, H. (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

Srinivasan, M., & Snedeker, J. (2011). Judging a book by its cover and its contents: The representation of polysemous and homophonous meanings in four-year-old children. *Cognitive Psychology*, *62*(4), 245–272. https://doi.org/10.1016/j.cogpsych.2011.03.002

Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., & Dai, J. (2019). VL-BERT: Pre-training of generic visual-linguistic representations. *PsyArXiv*. https://doi.org/10.48550/arXiv.1908.08530

Tenney, I., Das, D., & Pavlick, E. (2019). BERT rediscovers the classical NLP pipeline. *PsyArXiv*. https://doi.org/10.18653/v1/P19-1452

Thierry, G., Athanasopoulos, P., Wiggett, A., Dering, B., & Kuipers, J. R. (2009). Unconscious effects of language-specific terminology on preattentive color perception. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(11), 4567–4570. https://doi.org/10.1073/pnas.0811155106

Trott, S., & Bergen, B. (2021). *RAW-C: Relatedness of ambiguous words—In context (A new lexical resource for English)*. PsyArXiv. https://doi.org/10.48550/arXiv.2105.13266

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

Valera, S. (2020). Polysemy versus homonymy. In M. Aronoff (Ed.), *Oxford research encyclopedia of linguistics*. Oxford University Press. https://doi.org/10.1093/acrefore/9780199384655.013.617

Wiedemann, G., Remus, S., Chawla, A., & Biemann, C. (2019). Does BERT make any sense? Interpretable word sense disambiguation with contextualized embeddings. *PsyArXiv*. https://doi.org/10.48550/arXiv.1909.10430

Wittgenstein, L. (1953). *Philosophical Investigations*. Blackwell.

Yee, E., & Thompson-Schill, S. L. (2016). Putting concepts into context. *Psychonomic Bulletin & Review*, *23*(4), 1015–1027. https://doi.org/10.3758/s13423-015-0948-7

Yurchenko, A., Lopukhina, A., & Dragoy, O. (2020). Metaphor is between metonymy and homonymy: Evidence from event-related potentials. *Frontiers in Psychology*, *11*, Article 2113. https://doi.org/10.3389/fpsyg.2020.02113

Zellers, R., Holtzman, A., Peters, M., Mottaghi, R., Kembhavi, A., Farhadi, A., & Choi, Y. (2021). PIGLeT: Language grounding through neuro-symbolic interaction in a 3D world. *PsyArXiv*. https://doi.org/10.18653/v1/2021.acl-long.159

Received November 30, 2021
Revision received October 21, 2022
Accepted December 31, 2022 ■