This is the accepted version of the following article: Barend Beekhuizen, Blair Armstrong and Suzanne Stevenson (2021). 'Probing Lexical Ambiguity: Word Vectors Encode Number and Relatedness of Senses'. *Cognitive Science* 45: e12943, which has been published in final form at https://onlinelibrary.wiley.com/doi/full/10.1111/cogs.12943. This article may be used for non-commercial purposes in accordance with the Wiley Self-Archiving Policy http://olabout.wiley.com/WileyCDA/Section/id-828039.html.

ORIGINAL ARTICLE

Journal Section

Probing Lexical Ambiguity: Word Vectors Encode Number and Relatedness of Senses

Barend Beekhuizen[1] | **Blair C. Armstrong**[2] | **Suzanne Stevenson**[3]

Lexical ambiguity – the phenomenon of a single word having multiple, distinguishable, senses – is pervasive in language. Both the degree of ambiguity of a word (roughly, its number of senses), and the relatedness of those senses, have been found to have widespread effects on language acquisition and processing. Recently, distributional approaches to semantics, in which a word's meaning is determined by its contexts, have led to successful research quantifying the degree of ambiguity, but these measures have not distinguished between the ambiguity of words with multiple related senses versus multiple unrelated meanings. In this work, we present the first assessment of whether distributional meaning representations can capture the ambiguity structure of a word, including both the number and relatedness of senses. On a

[1]Department of Language Studies, University of Toronto, Mississauga, 3359 Mississauga Road North, Mississauga, ON, L5L 1C6, Canada

[2]Department of Psychology and Department of Language Studies, University of Toronto Scarborough, Basque Center on Cognition, Brain, & Language

[3]Department of Computer Science, University of Toronto

Correspondence

Barend Beekhuizen, Department of Language Studies, 3359 Mississauga Road North, Mississauga, ON, L5L 1C6, Canada

Email: barend.beekhuizen@utoronto.ca

Funding information

This research was supported by NSERC grant RGPIN-2019-06917 to Barend Beekhuizen, NSERC grant RGPIN-2017-06310 to Blair Armstrong, and by NSERC grant RGPIN-2017-06506 to Suzanne Stevenson.

very large sample of English words, we find that some, but not all, distributional semantic representations that we test exhibit detectable differences between sets of monosemes (unambiguous words) [$N = 964$], polysemes (with multiple related senses) [$N = 4096$], and homonyms (with multiple unrelated senses [$N = 355$]. Our findings begin to answer open questions from earlier work regarding whether distributional semantic representations of words, which successfully capture various semantic relationships, also reflect fine-grained aspects of meaning structure that influence human behavior. Our findings emphasize the importance of measuring whether proposed lexical representations capture such distinctions: in addition to standard benchmarks that test the similarity structure of distributional semantic models, we need to also consider whether they have cognitively plausible ambiguity structure.

KEYWORDS

lexical ambiguity; semantic ambiguity; homonymy; polysemy; distributional semantic models; vector space models

1 | INTRODUCTION

Lexical ambiguity – the phenomenon of a single word having multiple, distinguishable, senses – is pervasive in language: no language has been found to lack ambiguity at the word level (e.g., Youn et al., 2016), and within a language, large numbers of words are found to be ambiguous (e.g., Klein and Murphy, 2001). Indeed, lexical ambiguity is suggested to be a necessary property of language, as a way to efficiently express a large number of concepts with a small, finite lexicon (e.g., Bartsch, 1984; Piantadosi et al., 2012; Ramiro et al., 2018; Schaff, 1964). As such, lexical ambiguity is a central concern for the cognitive science of language, and the nature of the representations that support the encoding and processing of multiple senses of a word is key to understanding this phenomenon.[1]

Ambiguity is not a single monolithic property. Since Bréal (1897), linguistic research has by and large adopted a representational taxonomy of words as **monosemes** – those with a single distinguishable sense, such as *tango* referring to a type of dance; **polysemes** – those with multiple related senses, such as *chicken* referring to both the animal and the meat of that animal; and **homonyms** – those with multiple unrelated meanings, such as *bat* referring to a flying mammal and a type of sporting equipment. (Note that in the psycholinguistics literature, the term "senses" is often used more restrictively to refer to senses that are related, whereas "meanings" is taken to refer to senses that are not related to each other. We will use the terms "sense" and "meaning" here loosely along those lines, but will specify "related" or "unrelated" when the distinction is important.) We recognize that the categories of monosemes, polysemes, and homonyms may be notional "endpoints" of underlyingly-continuous properties, as sometimes understood in linguistics (e.g., Bartsch, 1984; Geeraerts, 1993; Tuggy, 1993) and psycholinguistics (e.g., Brisard et al., 2001; Hoffman et al., 2013). Nevertheless, the explanatory power of these coarse categories for elucidating the nature of lexical representations has been supported by experi-

[1] In the remainder of this paper, we use the term *ambiguity* to refer to "lexical ambiguity" (sometimes also referred to as "semantic ambiguity"), rather than the many other types of ambiguity that arise at other linguistic levels of representation.

mental work on various aspects of lexical processing (*lexical decision*: Klepousniotou and Baum 2007; Rodd et al. 2002, *semantic categorization*: Hino et al. 2006; Pexman et al. 2017, *semantic priming*: Klepousniotou et al. 2008; Williams 1992, *picture naming*: Rabagliati and Snedeker 2013, *sentence processing*: Brocher et al. 2016; Frazier and Rayner 1990; Frisson and Pickering 1999), as well as by computational cognitive modeling (e.g., Rodd et al., 2004; Armstrong and Plaut, 2016). We thus adopt this tripartite distinction between monosemes, polysemes, and homonyms as a useful construct in our work here.

An issue that arises when studying the impact of ambiguity on processing is how to measure the relevant aspects of this phenomenon. A number of measures have relied on processes involving human judgments, such as number of dictionary senses (e.g. Rodd et al., 2002), human ratings of ambiguity (e.g. Hino et al., 2006), or number of semantic features listed (e.g., McRae et al. 2005's feature naming norms as used by Pexman et al. 2008). However, some successful approaches have drawn on corpus-based measures that exploit the distributional hypothesis regarding word meaning: that is, the hypothesis that (some significant part of) a word's meaning may be derived from its usage contexts (Firth, 1957; Jones et al., 2017; Sahlgren, 2008). This distributional perspective has led to much research quantifying the degree of ambiguity not by directly assessing the representations of words, but *indirectly* by looking at the diversity of their (local linguistic) contexts: for example, the difference between the distribution of a word over its contexts and their prior distribution (McDonald and Shillcock, 2001); the number of different documents a word occurs in (Adelman et al., 2006); or the semantic dissimilarity between the contexts of a word (Hoffman et al., 2013; Jones et al., 2012). The general reasoning is as follows: ambiguous words occur in more diverse contexts (intuitively, more senses mean more contexts in which the word is applicable), and thus measuring diversity of a word's contexts provides a window into how ambiguous it is.

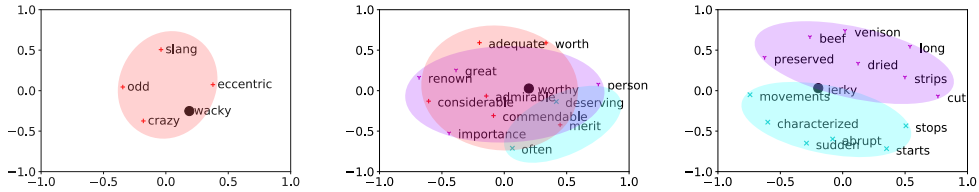
These approaches have contributed to our understanding of the relationship between contextual usages of a word and ambiguity, but have two important limitations. First, thus far,

research in the distributional tradition has focused on the *degree* of ambiguity, but not the *structure* of the ambiguity. In particular, these measures have not distinguished between the ambiguity of words with multiple related senses or unrelated senses. For example, both McDonald and Shillcock (2001) and Hoffman et al. (2013) point out that their measures do not capture the differences between related senses (polysemes) and unrelated meanings (homonyms) found by Rodd and colleagues (Rodd et al., 2000, 2002), and leave the assessment of the relatedness structure of ambiguity to future work.

A second limitation of this prior work on measures of ambiguity has (to our knowledge) not been previously identified: While the proposed measures have been motivated by the distributional hypothesis to consider distinctions among contexts, they have not actually considered whether and how a distributional representation derived from contexts captures those distinctions. That is, looking at only the contexts of a word assumes that relevant differences in those contexts are *necessarily* captured in its meaning representation. This is not the case: It is possible to measure something about the contexts themselves that any particular learning algorithm fails to capture in forming a representation of the contexts as the meaning of the word. In short, if we want to know what lexical properties actually influence acquisition and processing, we need to consider what is encoded in the lexical item – that is, we need to directly assess its representation. In particular, because previous work measuring lexical ambiguity only looked at the contexts, they do not take into account the relation of the word's representation to the semantic space in which it is embedded.[2] A fuller understanding of the lexical properties of ambiguity can benefit from considering how the word is structured within that semantic space, and variations in those relations that result from different aspects of ambiguity.

In this work, we directly address both of these limitations of prior work by presenting the first assessment of whether distributional meaning representations reflect the ambiguity structure

[2] Others have looked at the semantic neighborhood of words in a distributional framework (Buchanan et al., 2001; Burgess, 1998; Landauer and Dumais, 1997), but have not used this to measure ambiguity. For example, Buchanan et al. (2001) focus on the idea of semantic neighborhood density, and only speculate on how this might relate to different aspects of ambiguity. We return to this point later, since we make use of a similar measure to theirs in our work.



(a) Monoseme (*wacky*) (b) Polyseme (*worthy*) (c) Homonym (*jerky*)

FIGURE 1 Multidimensional scaling plots for the word2vec vectors of three target words (large black dot) and of words in their dictionary definitions (see text Section 2 and Experiment 1 in Sections 3 and 4) for the three ambiguity types. Words from the same dictionary definition are indicated by the color and shape of the symbols and the ellipses around them (definition 1 in red '+'s, definition 2 in purple 'y's, definition 3 in cyan 'x's). Distances in the MDS plot are comparable across the three subfigures.

of a word. Rather than testing whether the *contexts of a word* are informative about its number and relatedness of senses, we explore whether *its semantic representation* (derived from context) is sensitive to those aspects of meaning structure. Specifically, we consider the relation within the distributional semantic space of a target word's representation to those of **probe words** that are related to various senses of the target. Figure 1, which is described in detail in the next section, provides some intuition regarding our hypothesis: Monosemes should show the tightest relation in semantic space between the target word and the probe words of its single sense; polysemes, with related senses, should show the next tightest relation to probes of those senses; and homonyms, which encompass unrelated meanings, should show the loosest relation to the identified probes.

To preview our results, on a very large sample of English words, we find that some, but not all, distributional semantic representations that we test exhibit a detectable difference between sets of monosemes ($N = 964$), polysemes ($N = 4096$), and homonyms ($N = 355$), showing the predicted distinction between all three levels in the design. Our work thus extends understanding of the encoding of ambiguity within distributional representations of meaning in several ways. First, our findings begin to answer open questions from earlier work regarding whether distributional semantic representations of words, which have been found to successfully capture various semantic relationships (e.g., Baroni et al., 2014; Pereira et al., 2016), also reflect

fine-grained aspects of meaning structure found to influence human behavior (as cited above). Moreover, while there has been some debate as to whether a single distributional vector-based representation can capture multiple senses of a word (e.g., Jamieson et al., 2018; Li and Jurafsky, 2015; Reisinger and Mooney, 2010), our findings suggest that such a single representation of a word may in fact be sensitive to its multiple senses (complementing work from various angles, such as Arora et al., 2018; Beekhuizen et al., 2019; Burgess, 2001; Kintsch, 2001; Mu et al., 2017). Second, our approach shows that the postulated differences in meaning structure manually built into some previous computational models (Armstrong and Plaut, 2016; Rodd et al., 2004) may arise within a large scale distributional semantic space derived from natural language corpora. While these earlier modelers designed their meaning representations to make the relevant distinctions, we can see that at least some natural representations reflect them. Finally, our results show that not all distributional meaning representations exhibit the predicted pattern of differences between all three levels of ambiguity in our design. This suggests that in addition to comparisons of meaning representations on how well they predict human judgments (e.g., Baroni et al., 2014; Pereira et al., 2016), an assessment of adequate cognitive plausibility must also consider whether the representational structure reflects lexical properties of ambiguity that have been demonstrated to influence lexical processing in people.

2 | OUR APPROACH TO DELINEATING MONOSEMES, POLYSEMES, AND HOMONYM

Our goal is to see whether distributional semantic representations capture the ambiguity structure of words. We consider representations resulting from a distributional approach – that is, learned from the linguistic contexts of word usages – because they have been shown to successfully extract word meaning from samples of natural language (corpora) (e.g., Baroni et al., 2014; Pereira et al., 2016), and have been studied extensively in psycholinguistics as

cognitively plausible lexical representations (e.g., Burgess, 1998; Hollis and Westbury, 2016; Jones et al., 2017; Landauer and Dumais, 1997). Distributional approaches vary widely in their precise method of forming meaning based on the usage contexts of words, but typically create a geometric semantic space in which word meanings are encoded in distributional semantic vectors (DSVs), whose relative locations in space capture meaning relations among words.

Within this framework, we hypothesize that there are detectable differences between DSVs of words with single versus multiple senses, as well as between DSVs of ambiguous words with related senses versus unrelated meanings. To explore this hypothesis, we investigate how DSVs relate to relevant portions of the high-dimensional semantic space they occur in. In particular, we consider the similarity between target DSVs from each of the three ambiguity types – monosemes, polysemes, and homonyms – and various regions of the space, using relevant "probes" in the DSV space. Following much previous work (e.g., Burgess, 1998; Erk, 2012; Jones and Mewhort, 2007; Landauer and Dumais, 1997; Mikolov et al., 2013a; Pennington et al., 2014), we assume that the semantic similarity between two DSVs is indicated by their relative positioning in the semantic space: DSVs that occur close together in a region of the space are more semantically similar than those that are more spread out in the space. We thus tap into the ambiguity structure of a target DSV using a straightforward measure of its similarity in semantic space to the DSVs of probes related to its sense(s) in various ways.

With this in mind, Figure 1 (shown earlier) illustrates the components of our main hypothesis of a tripartite distinction between monosemes, polysemes, and homonyms, wherein each of these ambiguity types is different from the other two: (1) We assume that, because the DSV for a monoseme encodes a single sense, the expected similarity between its DSV and (vectors representing) probes related to that sense should be relatively high (represented visually as a low distance). (2) A DSV for a polyseme will be relatively less similar to (the vectors for) its related probes, since its multiple senses pull its word vector representation somewhat away from any one particular sense. (3) A DSV for a homonym will be the least similar to its related

probes, since its encoding reflects various meanings with no overlap in semantics; the resulting DSV must encode and thus "sit between" these more dissimilar, non-overlapping semantic regions.[3]

To test the hypothesis above, we require a distributional semantic space created from a large-scale sample of language use, so that we can measure the similarity between the DSVs of a set of target words from the three ambiguity types (monosemes, polysemes, and homonyms), and identified probes that tap into their senses. We use standard, off-the-shelf DSVs whose usage is widespread in psycholinguistics and computational linguistics. Our selection was guided by findings in the literature regarding the ability of various distributional models to capture human judgments in semantic tasks, such as similarity and analogy, in order to ensure the DSVs are capturing word meaning effectively. The goal is then to see if these DSVs also capture the ambiguity structure of words.

There are various ways to identify relevant probes within the semantic space to compare the target representations to. Because our goal is to probe the ambiguity structure of a target word – that is, to see whether its encoding is sensitive to the number and relatedness of the word's senses – we select probes that are expected to evoke the range of senses and meanings of the word. A natural choice is linguistic usage contexts, which (following the distributional hypothesis) are informative about the word's various senses – and indeed contribute to the target representation in the DSVs we use to test our hypothesis. We also select additional probes that tap into the word's range of senses in different ways, in order to robustly test whether other relational aspects of the target representation within the semantic space can (instead of or in addition to usage contexts) highlight the ambiguity structure of the word. This leads us to a set of experiments using three different probe types that relate to the target representations in various ways.

First we use as probes the **dictionary definition(s)** of a target word. Such definitions have

[3] Similar intuitions underlie the word sense disambiguation models of Schütze (1998) and Burgess (2001).

been carefully constructed to elaborate the distinctive semantics of the target, and as such, they serve as a set of clearly biased probes that reflect all of its senses, and (potentially) their relative relatedness. In this way, we expect the definition words of a target to constitute highly effective probes of how its ambiguity structure is captured in its DSV. Going back to the lay-out of monosemes, polysemes, and homonyms in semantic space, we expect these definitional probes to accurately pinpoint salient spatial regions whose similarities to the target are highly informative (with monosemes most similar to their probes, followed by polysemes, then homonyms; cf. Figure 1).

Second, we use the actual **linguistic usage contexts** of a word as probes. Specifically, we use a sample of corpus usages of the target word as examples of its natural contexts. Since these contexts are similar to the contexts used to create the target vectors, they are a natural probe to measure the extent to which the resulting representation of a word is more or less similar to DSVs reflecting its range of senses. These results should help reveal how the DSV is related to actual contextual aspects of its meaning, in contrast to the definitional aspects.[4]

Finally, we consider the local context of the target DSV within the semantic space; that is, we use as probes the target's most semantically-similar **neighbors** in the distributional model (cf. Buchanan et al., 2001; Burgess, 1998). Here, we are probing whether hypothesized differences in the make-up of the DSVs across the three types of words lead to different degrees of similarity to their nearest semantic neighbors.[5] Again, going back to the lay-out of monosemes, polysemes, and homonyms in semantic space, we hypothesize that the various neighbors of a target word may be more or less similar to the target depending on the variety of their shared semantic dimensions.

These three probe types – dictionary definitions, usage contexts, and neighbors in semantic

[4] Again, as noted in the introduction, measuring the similarity of the target representation to the word's usage contexts is not the same as measuring the relation of usages contexts to each other (e.g., Hoffman et al., 2013; Jones et al., 2012). While the latter may be informative about the contexts that contribute to a word's meaning, our approach is focused on determining how the resulting semantic representation reflects the ambiguity properties of interest.

[5] Note Buchanan et al. (2001) use dissimilarity to neighboring words as a measure of density of semantic neighborhood, rather than as a probe of ambiguity structure.

space – will be used to test our hypothesis in several ways, in order to shed light on whether and how the representation of a target word encodes differences among the ambiguity types that reflect the properties of interest – that is, whether the word is ambiguous or not, and whether an ambiguous word has related senses or distinct, unrelated meanings. Next, we describe the detailed experimental setup we use to test our hypothesis that we would observe differences between monosemes, polysemes, and homonyms.

3 | EXPERIMENTAL SET-UP

Here we describe the selection of the target words and distributional semantic representations for our study, and the details of our experimental approach. Note that we carry out our investigation on English words and semantic spaces due to the wealth of resources available for guiding selection of our experimental items, and the availability of frequently-used off-the-shelf DSVs for that language.[6]

3.1 | Target words

To identify appropriate monosemes, polysemes, and homonyms, we drew on the Wordsmyth dictionary (Parks et al., 1998). Wordsmyth structures definitions such that unrelated meanings of a word have separate entries, while related senses are grouped under a single entry.[7] This property enabled us to categorize extracted words as monosemes (a single entry with one sense), polysemes (a single entry with multiple related senses), and homonyms (multiple unrelated

[6] All data generated by us, as well as code necessary to replicate our experiments, can be found at https://osf.io/9q8ce/?view_only=3039b3b37a6b45cebda6919f7d24b83a.

[7] Manual inspection of Wordsmyth reveals that in a small number of cases the definitions may not cover some senses of a word, and some choices of senses as related (or not) may not be clear-cut. However, overall the unrelated meaning and related sense counts have been found to correlate significantly with ambiguity effects in several prior behavioural experiments (e.g., Armstrong and Plaut, 2016; Rodd et al., 2002).

entries, with possibly related senses within those).[8]

We selected target words from the three ambiguity types with the aim of minimizing the influence of other, potentially confounding, properties, by range-matching the targets in each category on a set of psycholinguistic covariates.[9] These properties include measures of word frequency, word length, phonological neighborhood, and number of senses. For the full list of psycholinguistic covariates we considered, see the first column of Table 1. To be able to take these properties into account, we limited our target words to those in found in Wordsmyth as well as the following sources: the SUBTL word frequency database (derived from movie/television subtitles; Brysbaert and New, 2009), the CMU pronouncing dictionary (Weide, 1998), and the measures of orthographic neighborhood reported in Yarkoni et al. (2008)

We used the eDom norms (Armstrong et al., 2012a) to further narrow down the retrieved set of homonyms from Wordsmyth, because eDom includes a large set of pre-screened homonyms suitable for psycholinguistic experimentation, as well as norms on additional psycholinguistic properties that may be of interest for later studies. Because homonyms are the least numerous of our three word types, we then further excluded any monosemes and polysemes whose values on our psycholinguistic covariates (see Table 1) fell outside of the ranges of values of the homonyms, in order to minimize the influence of possible confounds.[10]

Following all filtering steps, our target words included 335 homonyms, 4096 polysemes, and 964 monosemes, for 5395 items in total. The ranges, means, and variances of each of the psycholinguistic covariates for the three groups of items is detailed in Table 1.

[8] We filtered out all words that contained meanings in Wordsmyth that were morphologically derived from another word. That is, we excluded words like *stole*, 'past tense of *steal*' versus 'a woman's long, scarf-like garment', as such cases would have required cross-referencing meanings of the derived word's base form (in this case, *steal*), which turned out to be a complex task beyond the scope of this work.

[9] We further regress out (a subset of) these covariates in our experiments (see Section 3.5), and in addition analyze an item-matched dataset that further controls for their impact (see Section 4.2 and Appendix B).

[10] We later removed several words that upon manual inspection were found to be (1) dominantly used as a proper name, or (2) morphologically complex in at least one of their meanings even after our earlier filtering step ($n = 230$). Because of this, the range values per covariate do not exactly line up.

TABLE 1 Descriptive statistics for features used in range matching, and as covariates in our experiments (boldfaced).

Property	monosemes				polysemes				homonyms			
	min	max	mean	var	min	max	mean	var	min	max	mean	var
Number of phonemes	3	10	6.2	2.5	3	10	5.6	3.0	3	8	3.8	0.8
Number of letters	3	10	7.3	2.9	3	10	6.6	3.2	3	8	4.6	1.2
Number of syllables	1	4	2.4	0.6	1	4	2.1	0.7	1	4	1.2	0.2
Phonological Levenshtein Dist.	1	5.6	2.79	0.81	1	5.7	2.25	0.67	1	3.65	1.37	0.14
Coltheart's *N* (phonology)	0	34	1.5	16.1	0	41	3.4	38.9	0	39	12.2	87.9
Orthographic Levenshtein Dist.	1.05	5	2.84	0.60	0.60	4.95	2.36	0.51	1	3.4	1.54	0.14
Coltheart's *N* (orthography)	0	23	0.7	4.1	0	28	1.6	10.9	0	27	6.9	35.5
Positional unigram frequency	93	2733	1271	2036028	56	3187	1249	236571	84	2686	872	274952
Positional bigram frequency	5	753	151	11058	2	903	160	13738	5	554	106	10587
Log_{10} word frequency	0.69	4.59	1.34	0.47	0.69	4.61	1.88	1.00	0.69	4.59	2.19	0.96
Number of meanings	1	1	1	0	1	1	1	0	2	6	2.2	0.26
Number of senses	1	1	1	0	2	29	4.4	8.7	2	25	8.0	19.9
Number of noun senses	0	1	0.7	0.2	0	14	2.1	3.2	0	11	4.0	5.3
Number of verb senses	0	0	0	0	0	0	0	0	9	1	0.0	0.0
Number of adjective senses	0	1	0.2	0.2	0	12	0.8	2.2	0	8	0.5	1.4

3.2 | Distributional semantic spaces

Our goal is to see whether distributional semantic representations – which have served as proxies for cognitive representations of meaning – can capture the distinctions between monosemes, polysemes, and homonyms. In order to test a range of approaches, we worked with pre-trained vectors of four distributional semantic models that draw on different learning algorithms, and that have been of interest in both psycholinguistics and computational linguistics. We selected word2vec (using skipgram with negative sampling, SGNS; Mikolov et al., 2013a) because it is the top-performing model in various extensive tests on semantic tasks (Baroni et al., 2014; Pereira et al., 2016). In addition, we chose GloVe (Pennington et al., 2014) as another high-performing method from computational linguistics, and Latent Semantic Analysis (LSA; Landauer and Dumais, 1997) as a founding method in psycholinguistics. Early results on these three were reported in Beekhuizen et al. (2018); we also did further follow-up with Non-Negative Sparse Embeddings (NNSE; Murphy et al., 2012), as these representations are argued to be highly interpretable.

We found the most consistent patterns across these pilot experiments with word2vec vectors trained on English Wikipedia and Gigaword (Fares et al., 2017).[11] Such a result is consistent with the studies noted above showing the superior ability of word2vec to capture word meaning. Because the goal of our work is to test whether a distributional semantic representation of word meaning can reveal aspects of its ambiguity structure, in this paper we focus on our full set of new results using word2vec, which was most successful in this regard. For completeness, we report the additional results on the other three vector spaces in Appendix A.1. In addition, in Appendix A.2 we present some follow-up comparisons with GloVe (Pennington et al., 2014) to try to uncover what might give rise to some of the differences we found.

[11] These vectors were gathered from `http://vectors.nlpl.eu/repository/`.

3.3 | Our Experimental Measure

As noted above, our aim is to use a simple measure that can reveal basic properties of the meaning structure of a word, specifically focused on the relation of the word to its various types of probes – its definitions, its usages, and its semantic neighbors. We adopt as our experimental measure the mean cosine similarity between the DSV of the target word, and the DSVs of each instance of a probe of a certain type. For example, the relevant value for dictionary definition probes for a target word with 5 senses will be the mean cosine similarity between the target word and 5 definition DSVs. For each instance of a probe that consists of multiple words (i.e., most dictionary definitions and usage probes), we follow a common approach in computational linguistics (e.g., Schütze, 1998) of aggregating the DSVs of the content words in the probe text to form a single DSV, and compare that to the target DSV.

This similarity measure is both simple and consistently applicable. As noted earlier, pairwise context similarity does not take the target DSV itself into account, and our hypothesis concerns the *relation of a word's representation to that of its probes.* That is, we consider that the meaning structure of a word involves the relations of the meaning within the semantic space, and is not simply a property of its contexts themselves (although they contribute to the learned meaning). Other measures may tap into this construct (such as considering how the DSVs of a word and its contexts cluster in semantic space), but we chose to start with the average cosine similarity between the target and probes as a simple measure with minimal assumptions.

3.4 | Probe Types and Experiments

We operationalized the proposed experiments from Section 2 as follows.

Experiment 1: Dictionary definition probes are each of the definitions for the target given in Wordsmyth. The total number of definition probe DSVs for a word depends on the number of its

definitions. Each probe DSV is formed from a single definition by averaging the DSVs of all gloss words in that definition (omitting stopwords using *NLTK*; Bird et al., 2009). We take the mean cosine similarity of a target to its definition DSVs to get its similarity value for definition probes.

Experiment 2: Linguistic usages are defined as follows, in two sub-experiments. In Experiment 2a, we extract usages from (a part of) the corpus the vector space was trained on, in order to maximize the similarity of the usages we test here with the linguistic data the semantic space was learned from. Specifically, usages are fragments containing the target word plus five context words on either side, extracted from a dump of Wikipedia (dated July 1st, 2017). in Experiment 2b, we instead draw on a corpus frequently used in psycholinguistic experiments – the SUBTLEXus corpus of movie and television subtitles (Brysbaert and New, 2009) – as the genre represented in this corpus reflects distributions of colloquial language use. Here, again the usages are fragments containing the target words with (up to) five content words of context on either side. In both cases, we use a maximum of 100 usage contexts for a word (due to computational time): either a random sample of 100 usages, or all of its usages, if there were fewer than 100 tokens of the word in the corpus. (We achieve similar patterns using a maximum of 200 usage contexts per word as well.) Each usage context DSV is formed by averaging the DSVs of all words in the corpus fragment (excluding stopwords and the target word itself). For each corpus, we take the mean cosine similarity of a target to its usage context DSVs (in that corpus) to get its similarity value for usage contexts.

Experiment 3: Neighbor probes are the DSVs with the highest cosine similarity to the target vector (cf. Buchanan et al., 2001, who use this approach to investigate semantic neighborhood density).[12] Here we use the 100 nearest neighbors of the target word in the semantic space. (We achieve similar patterns using 20 and 200 neighbors as well.) We take the mean cosine similarity

[12] It is interesting to note here that while Buchanan et al. (2001) did not test for differences in the ambiguity structure of words (their focus was on effects of semantic neighborhood size in lexical decision and naming), they do speculate, contrary to our hypothesis and our findings below, that homonyms would show a *larger* average similarity than other words (a denser neighborhood) because of the presence of words related to more than one meaning. This illustrates the importance we stress here of considering the actual distributional representation of a word and what it encodes: The hypothesis of Buchanan et al. follows the intuition that each meaning of a homonym has a set of highly similar nearest neighbors, but does not consider that an aggregate semantic representation, which encodes all those meanings simultaneously, necessarily must "push away" one set of neighbors when bringing closer the other set.

of a target to its neighbor DSVs to get its similarity value for neighbor probes.

3.5 | Statistical methods

In each experiment, we used a hierarchical multiple linear regression procedure to test for differences between the similarity measures across the three ambiguity types of monoseme, polyseme, and homonym.

First, we regressed out the effects of the previously identified covariates, which are known to covary with measures of lexical representation and processing. This step was taken in order to avoid potential confound effects from other lexical factors (aside from ambiguity type) and to establish a conservative estimate of the unique effect of ambiguity type (Baayen et al., 2006).[13] Specifically, we used the features in bold in Table 1, omitting *Coltheart's N (orthography)*, *Coltheart's N (phonology)*, and *Number of letters*, to avoid collinearity with *Orthographic Levenshtein Distance*, *Phonological Levenshtein Distance*, and *Number of phonemes*, respectively. (Note that our particular choice of covariates to omit was found not to affect the results; see Appendix E for a follow-up experiment that establishes this.) We further left out the variables for *Number of meanings*, *Number of senses*, and the three variables for *Number of senses* per syntactic category, as these variables vary by definition across the three ambiguity types. For each of our three experiments (the three types of probes), we ran a multiple regression with the 7 identified covariates as independent variables and the similarity measure for that experiment as the dependent variable. Taking the residuals from this regression gives us the portion of the similarity measure unaccounted for by these standard psycholinguistic covariates.

Next, we tested for significant differences between the ambiguity types in predicting the these residuals; these are the results reported in Section 4 (and various Appendices). That is, here the predictor is the three-level variable of ambiguity type (monoseme, polyseme, and

[13] This approach follows previous similar literature (e.g. Boukadi et al., 2016; Cortese and Khanna, 2007; Cortese and Schock, 2013; Sánchez-Gutiérrez et al., 2018) in which the variable of interest was added separately in the last step of the regression models.

homonym), and the dependent variable is the residual similarity score, for each experiment. In these analyses, the baseline level of ambiguity type was rotated to run all pairwise comparisons between types: that is, we compared monosemes to polysemes, monosemes to homonyms, and polysemes to homonyms, and examined whether the regression coefficients in each case were statistically significant. The Type-I error rates in each experiment were held constant at $p < .05$ (two-tailed) and were corrected using the Bonferroni-Holm procedure. Each experiment formed one family of 12 comparisons: a residual similarity measure as the dependent variable and three pairs of (binary) independent variable values (3 ambiguity types compared to each other), tested for 4 vector spaces (word2vec and the other three from our experiments reported in Appendix A).

4 | EXPERIMENTAL RESULTS

We hypothesized that the properties of learned distributional semantic representations reflect both whether or not a word has multiple senses and, if so, whether those senses are all related or not. We further proposed that we can probe this meaning structure by considering the similarity of a target word's DSV to various probe DSVs, comparing these similarities across sets of monosemes, polysemes, and homonyms. Here we present the results of our experiments that consider the following two questions, corresponding to the two parts of the hypothesis:

1. Are ambiguous words (both polysemes and homonyms) less similar in semantic space to their probes than are unambiguous words (monosemes)?
2. Are ambiguous words with only related senses (polysemes) more similar to their probes than ambiguous words that have distinct (unrelated) meanings (homonyms)?

Our hypothesis will be supported to the extent that the experimental evidence suggests an answer of 'yes' to each of these questions.

We first present the experimental results in Section 4.1, discuss their implications in Section 4.2, and pose follow-up questions and further analyses in Section 4.3.

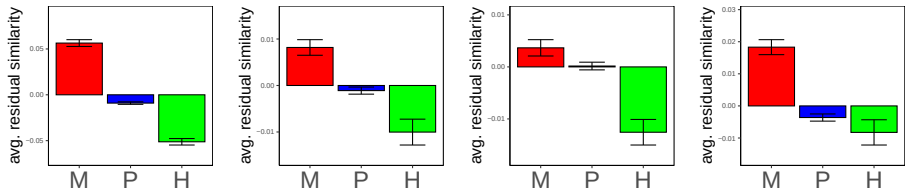
4.1 | Experimental Results

Given our hypothesis above, operationalized as questions (1) and (2), we predict the following results for similarities of probes to monosemes (M), polysemes (P), and homonyms (H): With regard to question (1), we predict both $M > P$ and $M > H$, and with regard to question (2), we predict $P > H$. Thus, if the answer to both questions is 'yes', we expect to observe a difference in the similarities for all three item types, $M > P > H$, indicating support for both parts of our hypothesis. A positive answer to the first question with a negative answer to the second question would yield a pattern of significant differences such that $M > P, H$; that is, unambiguous words are distinct from ambiguous ones, but the latter are not distinguishable. Less expected but also possible is a positive answer to the second question ($P > H$) with a "split" answer to the first ($M > H$, but **not** $M > P$), implying that only the homonyms are distinct due to their unrelated meanings: $M, P > H$.

Figure 2 shows the results in each of our experiments—with the statistically significant orderings of M, P, and H indicated—using word2vec as the distributional semantic representation. We discuss the results of each experiment in turn below.

4.1.1 | Experiment 1: Similarity to dictionary definitions

The mean residual similarities between the target DSVs and their definition DSVs for each of our ambiguity types are presented in Figure 2a. In line with our predictions, both parts of our hypothesis are supported here: the similarity of the target DSV to the definition DSVs was significantly greater for monosemes compared to each of the two ambiguous types, as well



(a) Exp. 1: Definition words. **M > P > H**

(b) Exp. 2a: Usage contexts (Wiki). **M > P > H**

(c) Exp. 2b: Usage contexts (SUBTL). **M, P > H**

(d) Exp. 3: Semantic neighbors. **M > P, H**

FIGURE 2 Average residual similarity of word2vec target vectors in the full dataset to each type of probe (as indicated in each subcaption), by ambiguity type: monosemes (*M*), polysemes (*P*), and homonyms (*H*). Significant differences (if any) between *M*, *P*, and *H*, and the direction, are indicated in each subcaption.

as significantly greater for polysemes compared to homonyms. The sample words from this experiment that were used to generate Figure 1 underscore this point: the definition words of the monoseme are most tightly clustered around the target word, followed by the polysemes, and with the homonym displaying the set of definition words that are most dissimilar from it in vector space. This indicates that a DSV trained on samples of natural language text, and probed with dictionary definitions, can reveal the predicted aspects of the meaning structure of a word.

4.1.2 | Experiment 2: Similarity to linguistic usage contexts

Figure 2b presents the mean residual similarities between the target DSVs and their linguistic usage DSVs using linguistic usages from the Wikipedia corpus, part of the training corpus for the DSVs we use. Here we find the same significant distinction between all three ambiguity types as with dictionary definitions, with monosemes more similar to their context probes than polysemes, which are in turn more similar than homonyms.

Figure 2c presents the mean residual similarities between the target DSVs and their linguistic usage DSVs derived from the SUBTLEXus corpus, a different genre from the training corpus for the DSVs. We find that these usage contexts for homonyms are significantly less similar to the target word than those of polysemes and of monosemes. There is no significant difference

between the latter two categories, however.

These results show that, when probed with usage contexts compatible with the kind of linguistic input the semantic model was trained on, the meaning structure of a DSV is sensitive to both the ambiguity of a word and the relatedness of its senses, as we predicted. However, the relationship of the target DSV to usage contexts from a different linguistic genre shows a different pattern, with only homonyms—words with multiple unrelated meanings—significantly different from the other two.

4.1.3 | Experiment 3: Similarity to nearest semantic neighbors

Figure 2d presents the mean residual similarities between the target DSVs and their nearest neighbor DSVs. Here, we observe that homonyms and polysemes both show greater dissimilarities than the monosemes, but are not statistically different from one another. In other words, these results indicate that ambiguous words have nearest neighbors that are less similar than those of unambiguous words, but the degree of similarity is not impacted by the relatedness of an ambiguous word's senses.

4.2 | Discussion of Results for Experiments 1, 2, 3

In these experiments, we looked at the similarities between word vectors and various probes, operationalized as the aggregate vectors of the dictionary definition words (Experiment 1), the aggregate vectors of the usage contexts of the words, either drawn from Wikipedia (Experiment 2a) or from SUBTLEXus (Experiment 2b), or the word vectors of the nearest neighbors in the vector space (Experiment 3).

Experiments 1 and 2a—using dictionary definitions and using usage contexts from the training corpus—show the significant pattern of differences that we predicted between all three

ambiguity types : homonyms being most dissimilar from their probes, followed by polysemes, then by monosemes. These results support our hypothesis that distributed semantic representations can reflect important properties of the meaning structure of a word—that is, both whether it is ambiguous and, if so, the relatedness of its multiple senses.

On the other hand, Experiment 3—using nearest neighbors—shows only a distinction between unambiguous and ambiguous words: monosemes are more similar to their probes than either of polysemes or homonyms, which are not significantly different. It appears that (at least, using our similarity measure) the structure of the semantic neighborhood of a word is sensitive only to a word having multiple senses, and not to their relatedness.

Finally, Experiment 2b shows a less expected pattern: although the numeric trend is in line with our prior findings of differences between all three ambiguity types , only the homonyms are significantly different from the other two. In isolation, this might indicate that only (un)relatedness of meanings is captured by the learned lexical representations we examine, rather than whether a word is ambiguous or not. However, given the significant distinctions between all three ambiguity types on Wikipedia usages (Experiment 2a), we suggest the lack of a significant difference between monosemes and polysemes on SUBTLEXus usages is likely caused by the mismatch between these usage context probes and the type of usage data the DSVs were trained on (Wikipedia and Gigaword). In particular, detecting the finer-grained difference between having a single sense (monosemes) and having multiple related senses (polysemes) may be sensitive to the probes capturing the same distribution of senses as the training corpus.

Overall, then, we see strong support for a positive answer to both our research questions: The representations of ambiguous words are generally less similar to their variously-related probes than unambiguous words, and (given suitably strong probes) we find that homonyms (encoding unrelated meanings) are less similar to their probes than are polysemes (encoding related senses). In short, distributional semantic representations of word meaning can capture key properties of the meaning structure of a word—whether it is ambiguous, and whether its

senses are related.

4.3 | Follow-up Questions and Further Analyses

While the results are promising regarding the ability of a learned distributional representation to detectably encode aspects of the ambiguity structure of a word, they raise further questions as well. For example, a major issue for consideration is whether the observed pattern of results might be due to factors other than the ambiguity type of the words. The three ambiguity types —monosemes, polysemes, and homonyms—were intended to capture the differences between words with a single sense and words with multiple senses ($M > P, H$), as well as between ambiguous words with related senses and unrelated meanings ($P > H$). By range-matching the items in the three ambiguity types and regressing out multiple covariates, we intended to minimize the influence on our results of these other lexical properties (such as word frequency or length; see Section 3.5). But, as Table 1 shows, there remain substantial differences between the distributions of these covariates across the ambiguity types.

To address that potential source of confound (as well as to support future research designing behavioural experiments), we created an item-matched dataset ($N = 335$ for each of monosemes, polysemes, and homonyms) to minimize the difference on the covariates, and repeated our experiments on that dataset; details are reported in Appendix B. For dictionary definition probes, we again found significant distinctions in the predicted order for the three ambiguity types ($M > P > H$). The experiments with other probes all showed a significant distinction between ambiguous and unambiguous words, but not between the two types of ambiguous words ($M > P, H$). Thus we get further strong support for our hypothesis that the meaning structure of DSVs can detectably encode the ambiguity of a word (having one or multiple senses), and, given the sufficiently strong probes of definitional words, the relatedness of senses.

This weaker effect of sense relatedness in the item-matched experiments is worth further consideration. The two ambiguity types of polysemes and homonyms were intended to capture the distinction of an ambiguous word having only related senses vs. having unrelated meanings, respectively. However, another key property that differs substantially between these items in the full dataset is their total number of senses (see Table 1). Thus, we must consider whether the more robust findings of a difference between polysemes and homonyms in the full dataset (compared to the item-matched set) are driven largely by the difference in the **number** of senses, rather than their **degree of relatedness**.

To study the effects of number of senses and their relatedness more directly, we ran follow-up analyses in which we directly assessed these two factors on all the polysemes and homonyms from our full dataset. We ran the same analyses as before, except that we used two independent variables, including the integral 'number of senses' in addition to the categorical 'ambiguity type' variable; see Appendix C for details. Interestingly, we do find that number of senses is significant in all experiments, suggesting that the magnitude of sense variation, and not just our initial categorical distinction of unambiguous (monoseme) vs. ambiguous (polyseme and homonym), is detectable in the distributional semantic representations of words. Crucially, in addition we observe that in three of the four experiments, 'ambiguity type' remains a significant predictor, over and above 'number of senses'. (These three analyses are the same ones in which, in our original experiments in Section 4.1, polysemes and homonyms showed a significant difference.) Thus when we analyze the polysemes and homonyms from our full dataset, we find that the difference between words with related senses and unrelated meanings cannot be reduced to the two groups having different numbers of senses. Disentangling of these nuanced effects in more detail will require future research—in particular, large-scale sense relatedness ratings that would enable creation of a dataset designed to tease apart the effects of number of senses versus their degree of relatedness. Until then, our results here establish that there is indeed an effect of ambiguity type pertaining to the relatedness—in addition to the number—of senses.

Another issue raised by our results in Section 4.1 is why, as alluded to earlier, the patterning of results seen with word2vec in Figure 2 is not as apparent when testing on other distributional semantic spaces. In particular, experiments on several other distributional models had yielded inconsistent results regarding the ability to capture distinctions in number and relatedness of senses; see Appendix A.1 for details of these results. We performed follow-up analyses to explore the possible source of the differences in behaviour, particularly focusing on GloVe vectors, due to claims of their improvement in some cases over word2vec (Pennington et al., 2014). In these analyses, we found that GloVe vectors are much more sensitive to word frequency than word2vec vectors; see Appendix A.2 for details. We tentatively conclude that word2vec may better generalize over infrequent data, and thus better capture semantic aspects of distributional behaviour.

In any case, our experiments make clear that not all distributional semantic representations *necessarily* capture important lexical distinctions that may be evident from a word's context. A natural question arises as to whether measures found to be successful at capturing degree of ambiguity solely from a word's contexts can show the differences between monosemes, polysemes, and homonyms that we found here in testing lexical representations. In further analyses, we examined three such measures: Contextual Distinctiveness, as proposed in McDonald and Shillcock (2001), Contextual Diversity, as presented by Adelman et al. (2006), and Semantic Diversity, as formulated in Hoffman et al. (2013) (cf. Jones et al., 2012); details are in Appendix D. We find that, as expected for measures of degree of ambiguity, these methods do indeed distinguish between unambiguous and ambiguous words. However, none of these measures separates polysemes from homonyms, failing to reliably detect the differences attributable to relatedness of senses (although Contextual Distinctiveness and Contextual Diversity do show the appropriate numeric trend). These findings further support our contention that lexical representations must be tested directly for whether they encode important aspects of ambiguity structure.

5 | GENERAL DISCUSSION

Our goal here was to investigate the nature of distributional lexical representations in which (part of) the meaning of words is captured by learning over their usage contexts. We proposed that key aspects of the *ambiguity structure* of a word could be revealed by the relationship of the distributional semantic vector (DSV) for the word to those of variously-related probes. In particular, we sought to determine whether the ambiguity of a word and the relatedness of its senses influence its semantic representation in a detectable way. We hypothesized that monosemes, encoding a single sense, should show the highest degree of semantic similarity in semantic space between a target word and its probes; polysemes, with multiple related senses, should exhibit a lower similarity between a target word and probes of those different senses; and homonyms, which encode unrelated meanings, should yield the lowest similarity between the target representation and the variously-related probes. We illustrate the intuition behind this hypothesis with visualizations of examples from our data in Figure 1.

We tested this hypothesis in several experiments that considered the similarities of DSVs to several types of probes: dictionary definitions that highlight the defining or prototypical semantic aspects of a word (Experiment 1), linguistic usage contexts that emphasize the co-occurrence relations of a word (Experiment 2), and neighbors in the vector space that indicate the relation of a word to semantically similar words (Experiment 3). Our expectation was that the calculated similarities in each case would show the distinctions as outlined above across the three ambiguity types—monosemes, polysemes, and homonyms.

Our key findings, using word2vec (Mikolov et al., 2013a) as the distributional semantic model, were as follows. All of the experiments showed a numerical ranking in the direction predicted by our hypothesis—that is, similarities for monosemes are higher than those of polysemes, which are in turn higher than those of homonyms—although the patterns of significance varied according to the probe type. In Experiment 1 (using dictionary definition probes) and Experiment

2a (with usage contexts from the corpus the DSVs were trained on), the predicted distinction between all three ambiguity types was found to be significant. This finding suggests that with the strongest probes—those drawing on definitional aspects of the word or on words highly associated to the learned representation—our hypothesis is born out that both the ambiguity of a word and the relatedness of its senses can influence the meaning structure of its DSV.

However, in Experiment 2b, using probes from a different genre, only the similarity of homonyms compared to the other word types was significantly different. This suggests that the ability of distributional lexical representations to reveal a fuller picture of meaning structure is dependent on the suitability of the probes, and in particular their congruence with the original training data. This may be especially the case with polysemes, whose finer-grained sense distinctions may be less apparent when probes have a different distribution of such senses from training data; this is in contrast to homonyms, whose distinctive meanings appear to be encoded robustly enough in the lexical representation to be detectable by probes across genres.

Moreover, in Experiment 3, we found only a significant distinction between monosemes and the two ambiguous word types (but no differences between homonyms and polysemes), when probed with neighboring words in the semantic space. This finding suggests that the actual meaning structure within the semantic space itself is less sensitive to the relatedness of the word's senses, such that homonyms and polysemes may not differ in the coarse layout of their local semantic neighborhood. Our simple measure of average similarity may just not be sensitive enough to capture the finer-grained structure of the semantic space, and considering a method such as relative clusterability of the neighbors may be required (cf. related discussion in Hoffman et al., 2013).

We aimed to reduce potential confound effects by matching the words across the three groups, that is, by making sure that for every homonym (the least numerous of our target items), there are exactly one polyseme and one monoseme that are maximally similar on a number of psycholinguistically-relevant covariates. With this matched dataset (in Appendix B), we found

significant distinctions between all three ambiguity types with the dictionary definition probes (Experiment 1), and a significant distinction between ambiguous and unambiguous words (but not between homonyms and polysemes) in all remaining experiments. Thus, here we see strong support for the hypothesis that the ambiguity of a word influences its semantic structure, while the relatedness of senses is only detectable using probes strongly biased to the various senses. Potential sources of the differences between the full and matched dataset are the reduced size of the data (and concomitant reduction in statistical power), along with the change in the make-up of the set of target items by matching them—especially for polysemes, for which the matched items are a relatively small and arguably non-representative sample of the population.

In particular, the polysemes in the full dataset have a notably larger mean number of senses than the homonyms. This raises the possibility that the detectable difference between polysemes and homonyms in the full dataset (in contrast to the matched dataset) is due to the difference in number of senses, rather than the difference in relatedness intended by our ambiguity type variable (i.e., of 'polyseme' vs. 'homonym'). In further analyses to test this (see Appendix C), we found that the effect of relatedness explains a significant amount of variance over and above the variance explained by number of senses. Thus, the learned representations of polysemes and homonyms are sensitive to both constructs of number and relatedness of senses.

We also experimented with other vector space models in addition to word2vec, as reported in Appendix A.1. Some results display the same distinctions found with word2vec between monosemes, polysemes, and homonyms—that is, either all three ambiguity types were significantly different from one another, or the unambiguous words were different from the ambiguous words but there were no differences between polysemes and homonyms. However, the patterns overall were not as consistent. In further analyses (in Appendix A.2), we investigated why the stable pattern for word2vec was not observed in particular for GloVe, another popular vector space model trained on the same corpus. Our findings (compatible with those of Schnabel et al.,

2015) showed that, compared to word2vec, GloVe representations are more sensitive to word frequency. Whether due to word2vec's predictive objective function, its local (usage-by-usage) training paradigm, or its well-tuned hyperparameters (for discussion of the latter, see Levy et al., 2015), word2vec's superior performance in various semantic tasks (e.g., Baroni et al., 2014; Pereira et al., 2016) is complemented by our findings here that its distributional representations reflect key properties of the meaning structure of a word – both the ambiguity of a word and the relatedness of its senses. Future research will need to aim at better understanding why some distributional representations have detectable ambiguity structure and others do not, helping to elucidate what properties of algorithms for learning word meanings are relevant to the cognitively plausibility of the resulting representations.

In any case, our results emphasize the importance of directly testing properties of actual lexical representations—as opposed to assessing properties of a word's contexts, as done in most previous measures of ambiguity (e.g., Adelman et al., 2006; Hoffman et al., 2013; Jones et al., 2012; McDonald and Shillcock, 2001). We find that such measures can distinguish the contexts of unambiguous versus ambiguous words, but in contrast to our approach, cannot reliably separate polysemes from homonyms; see Appendix D. Our results suggest that, while the distributional semantic hypothesis focuses on context as the locus of word meaning, research must also consider whether a learning algorithm over such contexts can robustly capture the aspects of ambiguity structure that are known to be cognitively relevant.

When we probed with a target word's nearest neighbors in the semantic space, we did not find a distinction between polysemes and homonyms. This was surprising because we thought that the semantic neighborhood of a word might be sensitive to whether the word's senses are related or not. That is, we expected that some of the nearest neighbors of a homonym would relate to very distinct meanings (and they do!), and that these would show a different degree of similarity with the target word compared to the neighbors of a polyseme with related senses. The lack of support for a distinction of this kind between homonyms and polysemes may be

indicative of the fact that senses can be *related* to each other without being semantically *similar* (which is the relation we measure among neighbors in the semantic space). For example, a film or sports *star* is not very similar to a celestial *star*, despite the former sense having a metaphorical relation to the latter. It is an open question whether or not relatedness of senses is encoded analogously to similarity, for example, as captured by semantic feature overlap (Armstrong and Plaut, 2016; Hino et al., 2006; Rodd et al., 2004). In either case, an understanding of how the representations of polysemes and homonyms relate to the semantic space in ways beyond simple semantic similarity may be needed to more fully elucidate their meaning structure.

This leads to the general question of whether the representational variation *within* specific ambiguity types is meaningful. In particular, it has been argued in both linguistics (e.g., Bartsch, 1984; Geeraerts, 1993; Tuggy, 1993) and psycholinguistics (e.g., Klepousniotou et al., 2008) that polysemes display varying degrees of semantic relatedness between their senses. Indeed, such differences are one possible explanation for the disparate findings regarding whether there are behavioural distinctions between polysemes and homonyms (Armstrong and Plaut, 2016; Hino et al., 2006)—that is, these varying effects may arise from differing item selection in such experiments. If polysemes do show substantial differences in sense relatedness, we might expect high variation in our similarity measures across the set of polysemes here. However, the similarities we find for the polysemes have no higher variance than the similarities for monosemes and homonyms; this suggests that the polysemes do not show a wide variation in the degree of relatedness of their senses, at least as tapped into by our measures. Future work will need to explore directly whether subgroups of polysemes have differing behaviour. For instance, do we find lower similarities in our experiments for metaphorical polysemes, such as FILM versus CELESTIAL *star*, compared to metonymic polysemes, such as *chicken* (ANIMAL or MEAT OF THAT ANIMAL)? It is possible that we may need to develop a more sensitive measure to detect such distinctions, and thereby make progress on the broader consideration of relatedness of senses as a continuum, rather than falling neatly into the discrete categories of

polysemes and homonyms.

Our choice of using the average similarity between each of the probe vectors and the target word vector as an indicator of semantic structure was motivated by its simplicity in directly capturing the relation of the target word representation to the variously-related probes. However, this measure does not capture the actual layout in semantic space of the probes with respect to the target. We have been experimenting with other, potentially richer, measures involving clusterability of context vectors, but found similar results to the ones using the simple similarity measure. In future work, we plan to experiment with measures that could tap into further properties of the meaning structure of a DSV, such as whether these representations capture the type of relation that exists among related senses of a polyseme. Given the success of word2vec in semantic analogy tasks (such as "*man* is to *king* as *woman* is to _"; Mikolov et al., 2013b), it is possible that the analogical relation among *chicken*, *fish*, and *lamb* (all referring to both ANIMALs and MEATs) may be detectable given a suitable measure.

Our results here have shown that a simple measure of similarity between distributional semantic vectors—from the vector of a target word to those of various semantic probes—enables us to detect differences between unambiguous and ambiguous target words, and (among the latter) between those with related and unrelated senses. Interestingly, these representational differences correspond to distinctions that prior computational cognitive modelers have assumed in order to simulate human behavioral data (Armstrong and Plaut, 2016; Rodd et al., 2004). While we have shown that distributional semantic representations created from natural corpora exhibit this ambiguity structure, it remains as future work to see whether the representations would show the behavioral correlates found in these models. In any case, our findings emphasize the importance of measuring whether proposed lexical representations capture important aspects of ambiguity type: in addition to standard benchmarks that test the similarity structure of distributional semantic models, we need to also consider whether they have cognitively plausible ambiguity structure.

Acknowledgements

We are very grateful to Saša Milić for her contributions to our preliminary work on this topic, and to Allan Jepson for extensive discussions on vector spaces. We also thank the attendees who commented on the work at CogSci 2018, and the anonymous reviewers of our papers both there and here, whose constructive input helped to improve the research.

References

Adelman, J. S., Brown, G. D. and Quesada, J. F. (2006) Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science*, **17**, 814–823.

Armstrong, B. C. and Plaut, D. C. (2016) Disparate semantic ambiguity effects from semantic processing dynamics rather than qualitative task differences. *Language, Cognition, and Neuroscience*, **31**, 940–966.

Armstrong, B. C., Tokowicz, N. and Plaut, D. C. (2012a) eDom: Norming software and relative meaning frequencies for 544 English homonyms. *Behavior Research Methods*, **44**, 1015–1027.

Armstrong, B. C., Watson, C. E. and Plaut, D. C. (2012b) SOS: An algorithm and software for the stochastic optimization of stimuli. *Behavior Research Methods*, **44**, 675–705.

Arora, S., Li, Y., Liang, Y., Ma, T. and Risteski, A. (2018) Linear algebraic structure of word senses, with applications to polysemy. *Transactions of the Association for Computational Linguistics*, **6**, 483–495.

Baayen, R. H., Feldman, L. B. and Schreuder, R. (2006) Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language*, **55**, 290–313.

Baroni, M., Dinu, G. and Kruszewski, G. (2014) Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the Association for Computational Linguistics*.

Bartsch, R. (1984) Norms, tolerance, lexical change, and context-dependence of meaning. *Journal of Pragmatics*, **8**, 367–393.

Beekhuizen, B., Cui, C. X. and Stevenson, S. (2019) Representing lexical ambiguity in prototype models of lexical semantics. In *Proceedings of the 41st Annual Meeting of the Cognitive Science Society*, 1376–1382.

Beekhuizen, B., Milić, S., Armstrong, B. and Stevenson, S. (2018) What company do semantically ambiguous words keep? Insights from distributional word vectors. In *Proceedings of the 40th Annual Meeting of the Cognitive Science Society*.

Bird, S., Klein, E. and Loper, E. (2009) *Natural Language Processing with Python: Analyzing text with the Natural Language Toolkit*. O'Reilly.

Boukadi, M., Zouaidi, C. and Wilson, M. A. (2016) Norms for name agreement, familiarity, subjective frequency, and imageability for 348 object names in tunisian arabic. *Behavior Research Methods*, **48**, 585–599.

Bréal, M. (1897) *Essai de sémantique: science des significations*. Hachette.

Brisard, F., Van Rillaer, G. and Sandra, D. (2001) Processing polysemous, homonymous, and vague adjectives. In *Polysemy in Cognitive Linguistics* (eds. H. Cuyckens and B. E. Zawada), 261–284. John Benjamins.

Brocher, A., Foraker, S. and Koenig, J.-P. (2016) Processing of irregular polysemes in sentence reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **42**, 1798.

Brysbaert, M. and New, B. (2009) Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, **41**, 977–990.

Buchanan, L., Westbury, C. and Burgess, C. (2001) Characterizing semantic space: Neighborhood effects in word recognition. *Psychonomic Bulletin & Review*, **8**, 531–544.

Burgess, C. (1998) From simple associations to the building blocks of language: Modeling meaning in memory with the HAL model. *Behavior Research Methods, Instruments, & Computers*, **30**, 188–198.

— (2001) Representing and resolving semantic ambiguity: A contribution from high-dimensional memory modeling. In *On the Consequences of Meaning Selection: Perspectives on resolving lexical ambiguity* (ed. D. S. Gorfein), 233–261. American Psychological Association.

Cortese, M. J. and Khanna, M. M. (2007) Age of acquisition predicts naming and lexical-decision performance above and beyond 22 other predictor variables: An analysis of 2,342 words. *Quarterly Journal of Experimental Psychology*, **60**, 1072–1082.

Cortese, M. J. and Schock, J. (2013) Imageability and age of acquisition effects in disyllabic word recognition. *The Quarterly Journal of Experimental Psychology*, **66**, 946–972.

Erk, K. (2012) Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compass*, **6**, 635–653.

Fares, M., Kutuzov, A., Oepen, S. and Velldal, E. (2017) Word vectors, reuse, and replicability: Towards a community repository of large-text resources. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*. Linköping University Electronic Press.

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

Frazier, L. and Rayner, K. (1990) Taking on semantic commitments: Processing multiple meanings vs. multiple senses. *Journal of Memory and Language*, **29**, 181–200.

Frisson, S. and Pickering, M. J. (1999) The processing of metonymy: Evidence from eye movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **25**, 1366–1383.

Geeraerts, D. (1993) Vagueness's puzzles, polysemy's vagaries. *Cognitive Linguistics*, **4**, 223–272.

Günther, F., Dudschig, C. and Kaup, B. (2015) LSAfun - an R package for computations based on Latent Semantic Analysis. *Behavior Research Methods*, **47**, 930–944.

Hino, Y., Pexman, P. M. and Lupker, S. J. (2006) Ambiguity and relatedness effects in semantic tasks: Are they due to semantic coding? *Journal of Memory and Language*, **55**, 247–273.

Hoffman, P., Ralph, M. A. L. and Rogers, T. T. (2013) Semantic diversity: a measure of semantic ambiguity based on variability in the contextual usage of words. *Behavior research methods*, **45**, 718–730.

Hollis, G. and Westbury, C. (2016) The principals of meaning: Extracting semantic dimensions from co-occurrence models of semantics. *Psychonomic bulletin & review*, **23**, 1744–1756.

Jamieson, R. K., Avery, J. E., Johns, B. T. and Jones, M. N. (2018) An instance theory of semantic memory. *Computational Brain & Behavior*, **1**, 119–136.

Jones, M. N., Dye, M. and Johns, B. T. (2017) Context as an organizing principle of the lexicon. In *Psychology of Learning and Motivation*, vol. 67, 239–283. Elsevier.

Jones, M. N., Johns, B. T. and Recchia, G. (2012) The role of semantic diversity in lexical organization. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, **66**, 115.

Jones, M. N. and Mewhort, D. J. (2007) Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, **114**, 1.

Kintsch, W. (2001) Predication. *Cognitive science*, **25**, 173–202.

Klein, D. E. and Murphy, G. L. (2001) The representation of polysemous words. *Journal of Memory and Language*, **45**, 259–282.

Klepousniotou, E. and Baum, S. R. (2007) Disambiguating the ambiguity advantage effect in word recognition: An advantage for polysemous but not homonymous words. *Journal of Neurolinguistics*, **20**, 1–24.

Klepousniotou, E., Titone, D. and 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**, 1534–1543.

Landauer, T. K. and Dumais, S. T. (1997) A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, **104**, 211.

Levy, O., Goldberg, Y. and Dagan, I. (2015) Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, **3**, 211–225.

Li, J. and Jurafsky, D. (2015) Do multi-sense embeddings improve natural language understanding? In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (eds. L. Màrquez, C. Callison-Burch, J. Su, D. Pighin and Y. Marton), 1722–1732. The Association for Computational Linguistics.

McDonald, S. A. and Shillcock, R. C. (2001) Rethinking the word frequency effect: The neglected role of distributional information in lexical processing. *Language and Speech*, **44**, 295–322.

McRae, K., Cree, G. S., Seidenberg, M. S. and McNorgan, C. (2005) Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, **37**, 547–559.

Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013a) Efficient estimation of word representations in vector space. *CoRR*, **abs/1301.3781**.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. and Dean, J. (2013b) Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26* (eds. C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani and K. Q. Weinberger), 3111–3119.

Mu, J., Bhat, S. and Viswanath, P. (2017) Geometry of polysemy. In *Proceedings of the International Conference on Learning Representations*.

Murphy, B., Talukdar, P. and Mitchell, T. (2012) Learning effective and interpretable semantic models using non-negative sparse embedding. *Proceedings of the International Conference on Computational Linguistics*, 1933–1950.

Parks, R., Ray, J. and Bland, S. (1998) Wordsmyth English Dictionary-thesaurus [Retrieved September 2008 from wordsmyth.net].

Pennington, J., Socher, R. and Manning, C. D. (2014) Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*.

Pereira, F., Gershman, S., Ritter, S. and Botvinick, M. (2016) A comparative evaluation of off-the-shelf distributed semantic representations for modelling behavioural data. *Cognitive Neuropsychology*, **33**, 175–190.

Pexman, P. M., Hargreaves, I. S., Siakaluk, P. D., Bodner, G. E. and Pope, J. (2008) There are many ways to be rich: Effects of three measures of semantic richness on visual word recognition. *Psychonomic Bulletin & Review*, **15**, 161–167.

Pexman, P. M., Heard, A., Lloyd, E. and Yap, M. J. (2017) The calgary semantic decision project: concrete/abstract decision data for 10,000 english words. *Behavior Research Methods*, **49**, 407–417.

Piantadosi, S. T., Tily, H. and Gibson, E. (2012) The communicative function of ambiguity in language. *Cognition*, **122**, 280–291.

Rabagliati, H. and Snedeker, J. (2013) The truth about chickens and bats: Ambiguity avoidance distinguishes types of polysemy. *Psychological Science*, **24**, 1354–1360.

Ramiro, C., Srinivasan, M., Malt, B. C. and Xu, Y. (2018) Algorithms in the historical emergence of word senses. *Proceedings of the National Academy of Sciences*, **115**, 2323–2328.

Reisinger, J. and Mooney, R. J. (2010) Multi-prototype vector-space models of word meaning. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 109–117. Association for Computational Linguistics.

Rodd, J., Gaskell, G. and Marslen-Wilson, W. (2000) The advantages and disadvantages of semantic ambiguity. In *Proceedings of the 22nd Annual Conference of the Cognitive Science Society, Mahwah, New Jersey*, 405–410.

Rodd, J. M., Gaskell, G. and Marslen-Wilson, W. (2002) Making sense of semantic ambiguity: Semantic competition in lexical access. *Journal of Memory and Language*, **46**, 245–266.

Rodd, J. M., Gaskell, M. G. and Marslen-Wilson, W. D. (2004) Modelling the effects of semantic ambiguity in word recognition. *Cognitive Science*, **28**, 89–104.

Sahlgren, M. (2008) The distributional hypothesis. *Rivista di Linguistica (Italian Journal of Linguistics)*, **20**, 33–53.

Sánchez-Gutiérrez, C. H., Mailhot, H., Deacon, S. H. and Wilson, M. A. (2018) Morpholex: A derivational morphological database for 70,000 english words. *Behavior research methods*, **50**, 1568–1580.

Schaff, A. (1964) Unscharfe ausdrücke und die grenzen ihrer präzisierung. In *Sprache und Erkenntnis: Essays über die Philosophie der Sprache.*, 220–243. Europa Verlag.

Schnabel, T., Labutov, I., Mimno, D. and Joachims, T. (2015) Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 298–307.

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

Tuggy, D. (1993) Ambiguity, polysemy, and vagueness. *Cognitive Linguistics*, **4**, 273–290.

Weide, R. L. (1998) The CMU pronouncing dictionary. *http://www.speech.cs.cmu.edu/cgi-bin/cmudict*.

Williams, J. N. (1992) Processing polysemous words in context: Evidence for interrelated meanings. *Journal of Psycholinguistic Research*, **21**, 193–218.

Yarkoni, T., Balota, D. A. and Yap, M. (2008) Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, **15**, 971–979.

Youn, H., Sutton, L., Smith, E., Moore, C., Wilkins, J. F., Maddieson, I., Croft, W. and Bhattacharya, T. (2016) On the universal structure of human lexical semantics. *Proceedings of the National Academy of Sciences*, **113**, 1766–1771.

A | ALTERNATIVE VECTOR SPACE MODELS

A.1 | Experiments on Other Vector Space Models

We tested our hypothesis using pre-trained vectors of three vector space models in addition to word2vec. We chose these particular sets of word vectors because they have all been made available for public use and have been shown to match human semantic judgments in various aspects. Early results on GloVe[14] (Pennington et al., 2014) and Latent Semantic Analysis[15] (LSA; Landauer and Dumais, 1997) (along with word2vec) were reported in Beekhuizen et al. (2018). We also did further follow-up with Non-Negative Sparse Embeddings[16] (NNSE; Murphy et al., 2012). Here we present the results on GloVe, LSA, and NNSE, using the identical experimental set-up to that described in Section 3. The results for word2vec are repeated here in Figure 2 for ease of comparison.

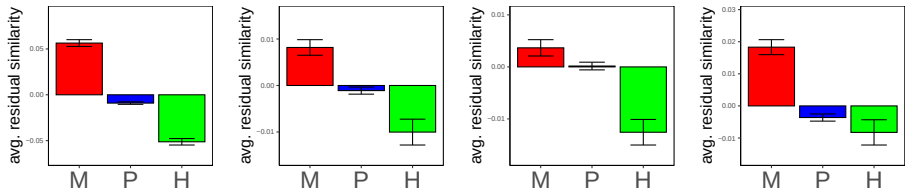
Figure 3 presents the results on the experiments for GloVe. The GloVe similarities using dictionary definition probes fit the prediction of observing differences between all three ambiguity types , and on SUBTLEXus usages and neighbors, they show the predicted difference of polyseme similarities being higher than homonym similarities. However, for both usage contexts, monosemes are *less similar* than polysemes, and for neighbors, no different from them.

LSA (Figure 4) and NNSE (Figure 5) also significantly display the hypothesized distinction between all three ambiguity types for the dictionary definitions. However, LSA has no significant differences between the ambiguity types for either of the usage contexts, and for neighbors, only distinguishes polysemes from homonyms, and in the opposite direction from expected. NNSE produces a split between monosemes and polysemes in the predicted direction for both usages contexts and for neighbors, but the split between polysemes and homonyms is either in

[14] These vectors were retrieved from `http://vectors.nlpl.eu/repository/`, trained on English Wikipedia and Gigaword (Fares et al., 2017), the same training corpus as the word2vec vectors we used.

[15] The LSA vectors used here, trained on the TASA corpus (Günther et al., 2015), are a standard set that has been the subject of extensive research over 20 years. These vectors were retrieved from `https://sites.google.com/site/fritzgntr/software-resources/semantic_spaces`

[16] These vectors were retrieved from `http://www.cs.cmu.edu/~bmurphy/NNSE/`, trained on approximately 15 billion words of web text.



(a) Exp. 1: Definition words. **M > P > H**

(b) Exp. 2a: Usage contexts (Wiki). **M > P > H**

(c) Exp. 2b: Usage contexts (SUBTL). **M, P > H**

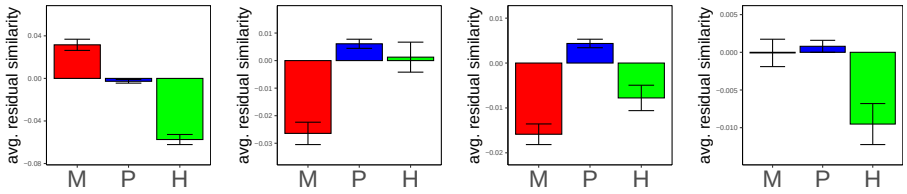
(d) Exp. 3: Semantic neighbors. **M > P, H**

FIGURE 2 Average residual similarity of word2vec target vectors in the full dataset to each type of probe (as indicated in each subcaption), by ambiguity type: monosemes (*M*), polysemes (*P*), and homonyms (*H*). Significant differences (if any) between *M*, *P*, and *H*, and the direction, are indicated in each subcaption.

the wrong direction (both usage contexts) or is not significant (neighbors).

Taken together, we see that all three vector spaces show the predicted ordering of $M > P > H$ when probed with dictionary definition words. As with word2vec, using these very biased probe words reveals a distinction between monosemes, polysemes, and homonyms in each vector space. However, these three pre-trained vector spaces mostly give inconsistent results across the other experiments and from each other. LSA, due to its resource intensivity, was trained on a much smaller corpus (the TASA corpus; Günther et al., 2015), so it is not clear whether the smaller training set or the algorithm itself leads to its differing results. NNSE was trained on approximately 15 billion words of text, and GloVe was trained on approximately 6 billion words of text (the same corpus as word2vec), so that differences between these vector spaces cannot be due to the size of the training corpus. Future work would be required to determine whether differences related to NNSE may be related to the nature of the training corpora.

We conclude that, of the pre-trained distributional semantic vectors we tested (all chosen for their use in computational and psycholinguistic work), word2vec best exhibits the distinctions in number and variety of senses that people appear to be sensitive to in their own lexical representations. Whether this is due to the particular learning algorithm or to judicious setting of hyperparameters (as suggested for word2vec's performance in various tasks; see Levy et al.,



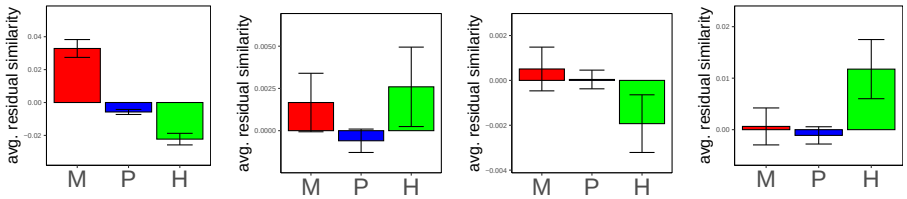
(a) Exp. 1: Definition words. **M > P > H**

(b) Exp. 2a: Usage contexts (Wiki). **M < P, H**

(c) Exp. 2b: Usage contexts (SUBTL). **M < P > H**

(d) Exp. 3: Semantic neighbors. **M, P > H**

FIGURE 3 Average residual similarity of GloVe target vectors in the full dataset to each type of probe (as indicated in each subcaption), by ambiguity type: monosemes (M), polysemes (P), and homonyms (H). Significant differences (if any) between M, P, and H, and the direction, are indicated in each subcaption.



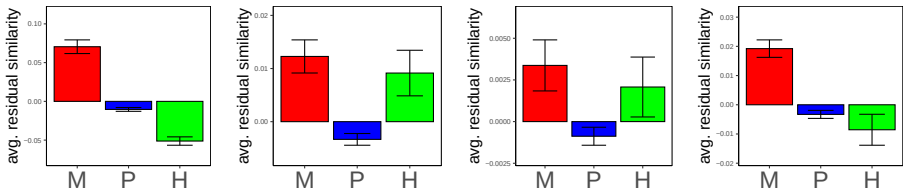
(a) Exp. 1: Definition words. **M > P > H**

(b) Exp. 2a: Usage contexts (Wiki). **No sig. differences**

(c) Exp. 2b: Usage contexts (SUBTL). **No sig. differences**

(d) Exp. 3: Semantic neighbors. **P < H**

FIGURE 4 Average residual similarity of LSA target vectors in the full dataset to each type of probe (as indicated in each subcaption), by ambiguity type: monosemes (M), polysemes (P), and homonyms (H). Significant differences (if any) between M, P, and H, and the direction, are indicated in each subcaption.



(a) Exp. 1: Definition words. **M > P > H**

(b) Exp. 2a: Usage contexts (Wiki). **M > P < H**

(c) Exp. 2b: Usage contexts (SUBTL). **M > P < H**

(d) Exp. 3: Semantic neighbors. **M > P, H**

FIGURE 5 Average residual similarity of NNSE target vectors in the full dataset to each type of probe (as indicated in each subcaption), by ambiguity type: monosemes (M), polysemes (P), and homonyms (H). Significant differences (if any) between M, P, and H, and the direction, are indicated in each subcaption.

2015) is a matter for future research. While it is an open research problem in computational linguistics to determine the precise reasons behind the differing performance of various distributional semantic methods (and beyond the scope of this paper), in the next section we further analyze GloVe to consider one possible factor at play in the results here.

A.2 | Further analysis of GloVe

Given that GloVe is a recent and popular vector space model that has been shown to accurately simulate several properties of semantic behaviour, it is surprising that it displayed inconsistent results across our experiments despite being trained on the same corpus as word2vec. Here we consider one potential cause of the differences found between the word2vec results in the main body of the paper, and those just above for GloVe. Following observations that GloVe encodes low-frequency words less accurately than word2vec (as evaluated on various semantic tasks; e.g., Schnabel et al., 2015), we propose that this sensitivity to frequency entails that GloVe does not reflect the semantic properties of interest to us as well as word2vec does.

To substantiate this hypothesis, we look at the correlations between a word's frequency and its similarity value in each of our experiments, for both GloVe and word2vec. We adopted those similarity measures on the assumption that they are primarily determined by semantic factors. If a model shows a very strong correlation of the similarities with frequencies, it indicates that, for that model, the similarity measures may simply not be sensitive enough to the semantic factors they are intended to probe.

In addition to using log-transformed word frequency from the SUBTLEXus corpus (which is one of the psycholinguistic covariates we considered in range-matching items), we also consider log-transformed word frequency in Wikipedia.[17] Since both GloVe and word2vec were trained on the Wikipedia + Gigaword corpus, these frequencies are more representative of the source

[17] The word frequency in Wikipedia was based on the frequency list retrieved from `https://github.com/IlyaSemenov/wikipedia-word-frequency/blob/master/results/enwiki-20150602-words-frequency.txt`, based on a Wikipedia dump from June 2nd, 2015.

	word2vec similarities				GloVe similarities			
variable	E1	E2a	E2b	E3	E1	E2a	E2b	E3
log word freq (SUBTLEXus)	-0.20	-0.12	-0.08	-0.09	0.32	0.57	0.58	0.41
log word freq (Wikipedia)	-0.30	-0.17	-0.38	-0.15	0.49	0.85	0.70	0.56

TABLE 2 Correlations between the similarities in the experiments (Exp. 1–Exp. 3) and two word frequency measures.

of the vector representations. We report the correlation coefficients per vector space in Table 2. All reported correlations are significant at the $p < .05$ level or lower.

The GloVe similarities display strikingly strong positive correlations with both kinds of word frequencies, ranging from $r = 0.32$ to $r = 0.85$, with five out of the eight comparisons having an r value over 0.5. These correlations mean that, in line with the observations of Schnabel et al. (2015), GloVe representations of frequent words have closer semantic links to their related words (our probes here) than do those of infrequent words. It is also worth noting that the weakest correlations using GloVe occur with similarities to definition words and semantic neighbors, which show the closest match to the predicted behaviour (of $M > P > H$) in our experiments (cf. Figure 3). Thus, the experiments where the similarities are least sensitive to word frequency do appear to tap into the semantic factors we are probing for.

The word2vec similarities, on the other hand, show much less prominent (but nonetheless significant) negative correlations, ranging from $r = -0.08$ to $r = -0.38$, with all but three of the eight comparisons having an absolute value of r under 0.2. Clearly, the similarities in word2vec space are less sensitive to the frequencies of the represented words, perhaps underlying the ability of the similarities to tap into ambiguity factors more effectively. (We have no explanation for why the correlations for word2vec are negative, while those of GloVe are positive, but we assume this arises from the differences in objective functions of the algorithms.)

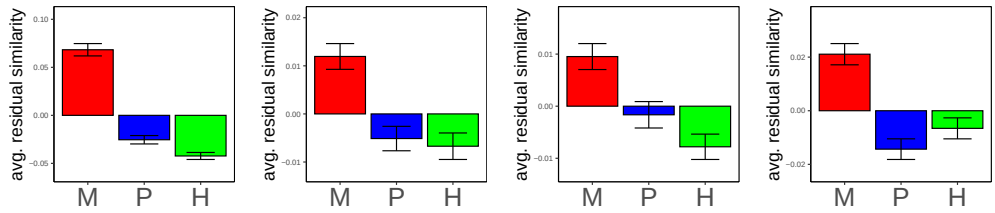
B | EXPERIMENTS ON A MATCHED DATASET

We noted that there are differences in the means and variances (and, in some cases, ranges) of the range-matched co-variates of our full dataset; see Table 1 in Section 3. In order to address that potential source of confound, we also created an item-matched dataset that minimizes these differences. The matched dataset is also intended to be of use in future behavioural experiments. Because the homonyms were the least numerous of our ambiguity types ($N = 335$), we selected 335 monosemes and 335 polysemes that were matched to the homonyms on the covariates to the greatest extent possible at the item level, using the SOS stimulus optimization software (Armstrong et al., 2012b). Table 3 presents descriptive statistics for the matched dataset.

We ran the experiments described in Section 3 on this matched dataset, obtaining the results shown in Figure 6. Here we observe the same distinction in Experiment 1 (dictionary definitions) as with the full dataset (cf. Figure 6a to Figure 2a in Section 4), such that all three ambiguity types were significantly different from one another. For Experiment 3 (nearest neighbors), we also see the same result as with the full dataset, of a distinction between unambiguous and ambiguous words but not between homonyms and polysemes (cf. Figure 6d to Figure 2d). For the two experiments on usage contexts, we also find a significant difference only between unambiguous and ambiguous words. (On the full dataset, these showed either a significant distinction between all three ambiguity types, or a difference between homonyms and the other ambiguity types; cf. Figure 6b and Figure 6c to Figure 2b and Figure 2c, respectively.) Thus, while two of the experiments differ from the full dataset, all experiments on the matched data support our hypothesis that the meaning structure of DSVs reflects the ambiguity of words, and one of the experiments further shows that DSVs may be sensitive to the relatedness of senses as well. These results gives us additional reassurance that the attested pattern of differences between ambiguity types is not due to confounding variables.

TABLE 3 Descriptive statistics for features used in item-matching, and as covariates in our experiments reported in Appendix B (boldfaced).

Property	monosemes				polysemes				homonyms			
	min	max	mean	var	min	max	mean	var	min	max	mean	var
Number of phonemes	3	10	5.0	1.9	3	10	4.4	1.5	3	8	3.8	0.8
Number of letters	3	10	6.1	2.5	3	10	5.3	1.7	3	8	4.6	1.2
Number of syllables	1	4	1.9	0.4	1	4	1.5	0.4	1	4	1.2	0.2
Phonological Levenshtein Dist.	1	5.30	2.12	0.56	1	3.85	1.67	0.27	1	3.65	1.37	0.14
Coltheart's *N* (phonology)	0	34	3.9	3.7	0	35	6.9	6.0	0	29	12.2	8.8
Orthographical Levenshtein Dist.	1	4.75	2.26	0.46	1	3.8	1.85	0.21	1	3.4	1.54	0.14
Coltheart's *N* (orthography)	0	23	1.9	9.6	0	20	3.4	1.7	0	27	6.9	3.6
Positional unigram frequency	93	2733	1235	298883	63	2433	1083	289853	84	2686	871	274952
Positional bigram frequency	5	658	146	143397	3	903	127	13063	5	554	106	10687
Log_{10} word frequency	0.69	4.59	1.89	0.72	0.69	4.57	2.12	0.90	0.69	4.59	2.19	0.96
Number of meanings	1	1	1	0	1	1	1	0	2	6	2.2	0.3
Number of senses	1	1	1	0	2	25	7.7	17.6	2	25	8.0	19.9
Number of noun senses	0	1	0.8	0.2	0	11	3.9	4.6	0	11	4.0	5.3
Number of verb senses	0	1	0.0	0.0	0	15	3.3	8.9	0	16	3.3	9.3
Number of adjective senses	0	1	0.2	0.2	0	5	0.5	1.2	0	8	0.5	1.4



(a) Exp. 1: Definition words. **M > P > H**

(b) Exp. 2a: Usage contexts (Wiki). **M > P, H**

(c) Exp. 2b: Usage contexts (SUBTLEX). **M > P, H**

(d) Exp. 3: Semantic neighbors. **M > P, H**

FIGURE 6 Average residual similarity of word2vec target vectors in the **matched** dataset to each type of probe (as indicated in each subcaption), by ambiguity type: monosemes (M), polysemes (P), and homonyms (H). Significant differences (if any) between M, P, and H, and the direction, are indicated in each subcaption.

In the usage context experiments here, the polysemes and homonyms are not distinguishable. It is worth noting that the item matching was primarily constrained by the small set of viable homonyms, which means that the covariates of the extracted polysemes (and monosemes) are driven by the tighter range of these properties in homonyms. However, this may also mean that the samples of matched polysemes (and monosemes) may not fully represent their larger populations. Although controlling for the covariates with the item-matching procedure was intended as an additional measure of statistical control, the potential lack of representativeness, especially for the large population of polysemes, may be a potential downside of the matched dataset, and this trade-off should be borne in mind when using it.

C | EXPERIMENTS INCLUDING NUMBER OF SENSES

The weaker effect of sense relatedness observed in the item-matched experiments of Appendix B suggested further consideration. Our polyseme and homonym items were selected based on having only related senses versus unrelated meanings, respectively, and thus our ambiguity type variable was intended to capture this distinction between them. However, a key property that differs substantially between polysemes and homonyms in the full dataset, but

not in the item-matched dataset, is their total number of senses (see Table 1). Thus, we must consider whether the more robust findings of a difference between polysemes and homonyms in the full dataset are driven largely by the difference in the **number** of senses, rather than their **degree of relatedness**.[18] While the number of senses of polysemes and homonyms are much closer in the item-matched data, those latter analyses suffer from two additional issues: First, there is a substantial reduction in statistical power due to a much smaller number of items in the matched sets. Second, the item-matching process leads to a selection of matched polysemes that may not be representative of the wider population of polysemes (see discussion in Appendix B).

To study the effects of number of senses and their relatedness more directly, and avoid these disadvantages of the item-matched experiments, we ran follow-up analyses in which we directly assessed the role of **both** number of senses and ambiguity type (the latter being the categorical difference between polysemes and homonyms that we propose captures sense relatedness). Here we include all and only the polysemes and homonyms from the full dataset (omitting monosemes, which uniformly have number of senses equal to 1). We applied the same hierarchical regression set-up as elsewhere, with the same initial step of removing the contributions of the same psycholinguistic covariates as noted earlier. In the earlier experiments, the second step of this process consisted of a monovariate regression in which the residualized similarities (calculated by our various measures) were predicted only on the basis of ambiguity type. Here we instead ran a multivariate regression in which we predicted the residualized similarities on the basis of both the ambiguity type (polyseme vs. homonym) and the number of senses. If the difference between polysemes and homonyms can be fully explained by differences in their number of senses, without regard to their relatedness, ambiguity type should not be a significant factor in these analyses.

[18] Because number of senses is partially reflected in the way the levels of our independent variable are defined—monosemes having one sense, polysemes and homonyms more than one sense—we did not use it as a predictor in the hierarchical regression analysis, nor did we regress it out along with other covariates in finding the residual similarities for our main experiments.

The results of this series of analyses are presented in Table 4. Recall that, in our experiments on the full dataset (Section 4.1), which did not include the number of senses as a variable, we found a significant effect of ambiguity type of polysemes versus homonyms in each of Experiments 1, 2a, and 2b. Here we want to see whether this ambiguity type effect, which is intended to capture relatedness of senses, remains significant in these analyses even when accounting for the effect of number of senses. Indeed, we see in Table 4 that this is the case: **both** ambiguity type and number of senses are significant predictors in the first three experiments.[19]

As we found in our earlier analyses, the neighborhood similarities (Experiment 3) are not sensitive to relatedness—the polyseme versus homonym distinction. However, in both our main analyses and those here, we find strong evidence that the neighborhood similarities are sensitive to the number of senses: as single versus multiple senses (monosemes vs. polysemes and homonyms) in the original experiments, and as an integral number of senses in the experiments reported here. It is perhaps the case that, in this experiment, the categorical variable of ambiguity type is simply not sufficiently sensitive to pick up the relatedness distinction, with only the binary distinction of polysemes versus homonyms. Modeling a finer-grained distinction would, however, require large-scale ratings of sense relatedness. In the absence of such ratings, the categorical variable is all we have available for testing the effect of sense relatedness. An aim for the future is to collect large scale sense relatedness ratings that would enable creation of a dataset of polysemes and homonyms designed to tease apart the effects of number of senses versus their degree of relatedness. In the meantime, our results here establish that there is an effect of ambiguity type pertaining to the relatedness—and not just the number—of senses.

[19] In the case of Experiment 2a, the direction of effect of ambiguity type is curiously reversed. We believe this may not be meaningful: Firstly, the sign *is* negative (as expected) in the univariate experiment (cf. Figure 2b). Secondly, because number of senses and ambiguity type are correlated with one another and we are analyzing the effects of both of these predictors simultaneously, the slopes and *p*-values here for ambiguity type reflect only the smaller amount of unique variance explained only by it, after removing the variance that is also explained by number of senses (the same logic also applies to the number of senses predictor, but here we focus on the ambiguity type variable that is central to the present work). Not allowing any of the shared variance between ambiguity type and number of senses to be (at least partially) attributed to ambiguity type is a very conservative statistical test of our ambiguity type variable and could also have led to moderation or suppression effects. Given that the results in these supplemental analyses are mostly in agreement with our first set of analyses, we take the tentative position that, by and large, ambiguity type per se has the effects that we have claimed. This issue might be better probed in future work that selects an item set with the a priori aim of matching stratified samples of number of senses across the two ambiguity types, and then testing for this effect in the absence of any confounds.

experiment	ambiguity type	number of senses
Exp. 1: Definition words	-0.027 ***	-0.005 ***
Exp. 2a: Usage contexts (Wiki)	0.009 *	-0.004 ***
Exp. 2b: Usage contexts (SUBTL)	-0.007 *	-0.002 ***
Exp. 3: Semantic neighborhoods	-0.002 (n.s.)	-0.002 ***

TABLE 4 Slope and significance for number of senses and ambiguity type when predicting residualized similarities on each of the four experiments.

Disentangling of these nuanced effects in more detail will require future research.

D | COMPARISONS TO OTHER CONTEXT-BASED MEASURES

In Section 1, we noted three measures that have been proposed to assess degree of ambiguity of a word, namely Semantic Diversity or SemD (Hoffman et al., 2013), Contextual Distinctiveness (McDonald and Shillcock, 2001), and Contextual Diversity (Adelman et al., 2006). (For detailed discussions of these models, we refer to the respective papers.) Motivated by the distributional hypothesis, these measures consider the structure of the usage contexts of a word, without directly considering the distributional semantic representation of the word itself (as we do here). As such, it is worth comparing how well SemD, Contextual Distinctiveness, and Contextual Diversity capture the ambiguity structure in the dataset used in this paper – that is, can they make the distinctions among monosemes, polysemes, and homonyms that our approach can? Specifically, we compare each of the three measures to our method using probes consisting of usage contexts (Experiment 2), since that constitutes the most directly comparable scenario.

For SemD, we used off-the-shelf scores made available with the paper (Hoffman et al., 2013). For Contextual Distinctiveness, we calculate scores[20] for the same Wikipedia corpus we use (as in Experiment 2a), since those are the contexts used to train our DSVs and thus provide the

[20] Using a five-word window, and considering the top-500 most frequent words, as these settings obtained good results, as reported by McDonald and Shillcock (2001), and as a five-word window makes this approach maximally comparable to our hyperparameter settings for word2vec.

most direct comparison of the methods.[21] For Contextual Diversity (Adelman et al., 2006), we used the scores reported for the SUBTLEXus corpus reported in Brysbaert and New (2009).[22] We then determined the intersection of the words from our experiments and words for which we had SemD scores, and used that set of words as our set of items in the analyses below. This includes 738 monosemes, 3787 polysemes, and 317 homonyms, for a total of 4737 items.

D.1 | Correlations between Context-Based Measures

As a first comparison, we correlated our probe similarity measures using usage context probes, with SemD ($r = -0.40$), Contextual Distinctiveness ($r = 0.19$), and Contextual Diversity ($r = -0.10$) across all the items in the dataset here ($p \ll .001$). These low to medium correlations suggest that our approach reflects different properties than SemD, Contextual Distinctiveness, and Contextual Diversity, of the linguistic contexts and their relation to the target words. (Note that SemD and Contextual Distinctiveness have a correlation on the dataset here of $r = -0.27$, SemD and Contextual Diversity of $r = 0.30$ and Contextual Distinctiveness and Contextual Diversity of $r = -0.41$, all $p < .001$.)

The negative correlation of our measure with SemD is expected because SemD, as a measure of diversity, assesses dissimilarities among contexts, higher values of which would often correspond to a lower similarity as assessed by our measure (i.e., of context vectors and the representation of the target word in word2vec). The positive correlation with Contextual Distinctiveness is in line with the expectation that a target word-conditioned distribution over context words that deviates more from the global distribution (high Contextual Distinctiveness), and a higher average similarity between word2vec context vectors and the representation of the target word (low values on our measure), both reflect limited diversity in the set of contexts

[21] We also compared Contextual Distinctiveness trained on SUBTLEXus to our Experiment 2b on SUBTLEXus contexts as probes, but the results using Contextual Distinctiveness were worse than those on Wikipedia (and worse than our results on SUBTLEXus).

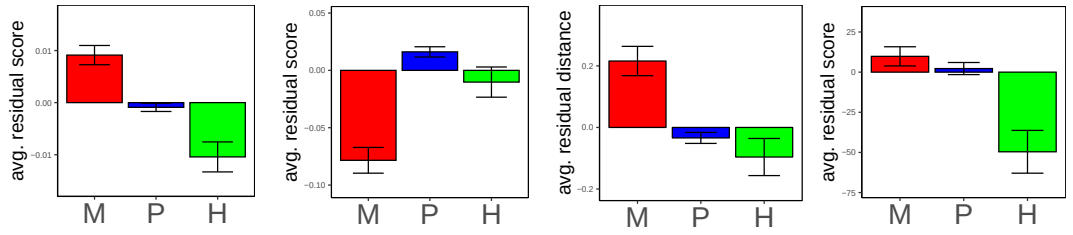
[22] These results are similar to our Experiment 2b on SUBTLEXus.

a word occurs in. Finally, although significant, the correlation between Contextual Diversity and our measures is small, suggesting that the two measures are tapping into different sources of information.

D.2 | Experimental Results with Other Context-Based Measures

Next, we reran our experiments using the above values of SemD, Contextual Distinctiveness and Contextual Diversity, with an identical set-up as in Section 3. Figure 7 presents the results for our corresponding results on usage based contexts (our Experiment 2a), along with those of SemD, Contextual Distinctiveness (calculated on the same corpus), and Contextual Diversity. Recapitulating our own findings, in Section 4.1.2 we established that when using aggregated context vectors from the Wikipedia corpus as our probes, our approach found distinctions between monosemes, polysemes, and homonyms, repeated here in Figure 7(a). In contrast, SemD and Contextual Distinctiveness each display the distinction between ambiguous and unambiguous words (as expected, as measures of degree of ambiguity), but fail to show the split between related and unrelated senses; see Figure 7(b) and (c). Contextual Diversity, finally, shows a distinction between homonyms and the other two groups, with homonyms having a lower diversity than monosemes and polysemes (Figure 7(c)), and thus also fails to show the pattern of distinctions between all three ambiguity types observed when using our measures.

We take this as further support that our approach of considering the target representation, and not only the contexts (as the three other measures do), is critical to evaluating the ambiguity structure of a word's distributional semantic representation. Whereas all these measures draw on the same information – bags of words making up the linguistic contexts of a target word – they differ in how they conceptualize such information, and thereby only display low to medium correlations with our measures, and obtain differing results on our experiments.



(a) Results from our Exp. 2a (on data subset Appendix D).

M > P > H

(b) Average residual SemD.

M < P, H

(c) Average residual Contextual Distinctiveness.

M > P, H

(d) Average residual Contextual Diversity.

M, P > H

FIGURE 7 Average residualized measures in the full dataset to each type of probe, by ambiguity type: monosemes (*M*), polysemes (*P*), and homonyms (*H*). Significant differences (if any) between *M*, *P*, and *H*, and the direction, are indicated in each subcaption.

E | IMPACT OF COVARIATE SELECTION

In Section 3.5, we discussed our experimental approach using hierarchical regression, in which we regressed out (from our similarity measures) some but not all of the pycholinguistic covariates listed in Table 1. This may raise issues about the robustness of our results: are they are dependent on the particular subset of covariates used in the regression, or is our selection representative of a set of measures that all capture the same underlying psychological constructs?

Recall that we partialed out the effects of the number of syllables, the number of phonemes, log word frequency, PLD20, OLD20, bigram probability, and unigram probability. We did not use Coltheart's N (Ort), Coltheart's N (Phon), and number of letters in the regression, because these covariates are known to tap into similar underlying constructs as, and be collinear with, the included variables (e.g., see Yarkoni et al., 2008, for discussion of the relationship between Coltheart's N and OLD20).

To study the robustness of this particular selection of covariates, we compared it to the results obtained when using the most explanatory factors that emerge from a Principal Com-

ponent Analysis (PCA). Specifically, we applied PCA to all 10 covariates, and kept the first N components whose cumulative relative variance summed to 0.90 or greater (i.e., those that explain at least 90% of the variance in the covariance matrix). In our case, that meant using the first $N = 5$ components, both in the full dataset as well as in the smaller, matched, dataset (cf. Appendix B).

The aim of this process is to identify estimates of latent constructs (the PCA components) that are tapped into by these psycholinguistic variables. For example, this process identified a principle component correlated with OLD, PLD, and Coltheart's N, and that appears to reflect a property akin to orthographic neighborhood size. Rather than using any particular combination of our psycholinguistic variables, whose multiple inclusion can cause other issues such as collinearity violations, we only enter the estimates of the components that our set of psycholinguistic variables tap into, that is: transforming each of the 10 covariate values for each datapoint into a 5-dimensional coordinate in the PCA space. An additional advantage of this procedure is that every component is uncorrelated with every other component, eliminating collinearity issues when including the components in the model. Furthermore, by virtue of the relatively small number of components needed to account for the bulk of the variance in the psycholinguistic covariates, the resulting regression model includes fewer independent variables and gains statistical power.

Using the transformed covariate values (within the 5-dimensional PCA space) in the regression model – instead of our set of psycholinguistic covariates reported in the main text – resulted in identical patterns of significance in our results, both for the full data set as well as for the matched data. This corroborates the finding reported in the main text and demonstrates that it is not due to an idiosyncratic choice of which covariates to include in our model.