What Company Do Semantically Ambiguous Words Keep? Insights from Distributional Word Vectors

Barend Beekhuizen
Dept. of Computer Science
University of Toronto
barend@cs.toronto.edu

Saša Milić
Dept. of Computer Science
University of Toronto
sasa@cs.toronto.edu

Blair C. Armstrong
Dept. of Psychology
University of Toronto
blair.armstrong@utoronto.ca

Suzanne Stevenson
Dept. of Computer Science
University of Toronto
suzanne@cs.toronto.edu

Abstract

The diversity of a word's contexts affects its acquisition and processing. Can differences between word types such as monosemes (unambiguous words), polysemes (multiple related senses), and homonyms (multiple unrelated meanings) be related to distributional properties of these words? We tested for traces of number and relatedness of meaning in vector representations by comparing the distance between words of each type and vector representations of various "contexts": their dictionary definitions (an extreme disambiguating context), their use in film subtitles (a natural context), and their semantic neighbours in vector space (a vector-space-internal context). Whereas dictionary definitions reveal a three-way split between our word types, the other two contexts produced a two-way split between ambiguous and unambiguous words. These inconsistencies align with some discrepancies in behavioural studies and present a paradox regarding how models learn meaning relatedness despite natural contexts seemingly lacking such relatedness. We argue that viewing ambiguity as a continuum could resolve many of these issues.

Keywords: lexical/semantic ambiguity; homonymy; polysemy; vector space models; contextual diversity

Introduction

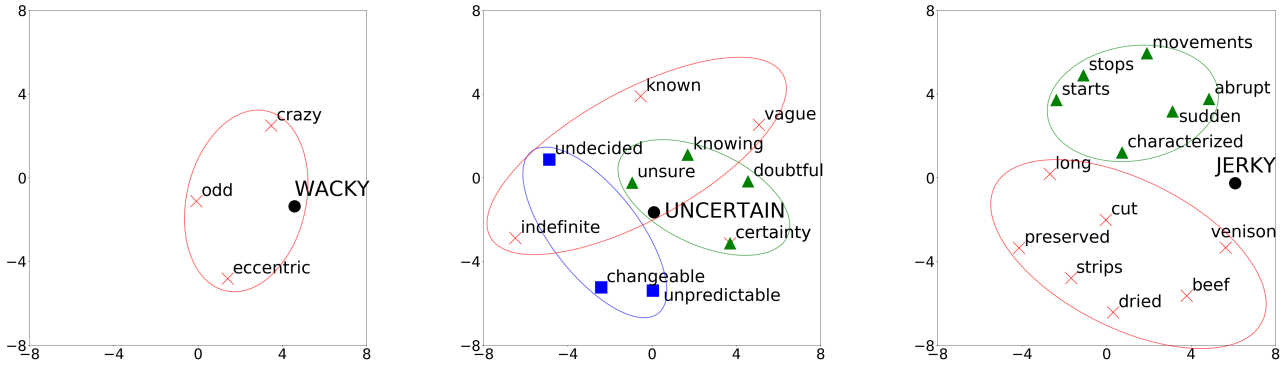
Extracting the statistical structure in a stream of words provides the observer—be it a human or a computational model—with important information about word meanings (e.g., Smith & Yu, 2008; Erk, 2012). Encountering a word in a number of diverse contexts permits the accumulation of information regarding the frequent versus idiosyncratic co-occurrence rates with other words, thereby revealing the word's meaning based on "the company that it keeps" (Firth, 1957). Indeed, varied contexts are necessary for the human or machine learner to pinpoint the consistent semantic aspects (across situations) that are relevant to a word, such that overall, contextual diversity helps the learner to predict a word's meaning (e.g., Kachergis et al., 2017). Moreover, contextual diversity is apparently reflected in the resulting learned representations of words, acting as a key principle of lexical organization (e.g., Jones et al., 2017).

Treating contextual diversity as a monolithic property, however, is a major simplification: We need a better understanding of how the interpretations of words can be shaped by diverse contexts that vary in partially systematic ways, and how this impacts their learned representations. For example, some of the contextual diversity for a *homonym* such as *bank* is due not only to using this word in varied contexts discussing MONEY, but also to using this word in varied and (relative to MONEY contexts) *distinct* contexts discussing RIVERS. Thus, the similarity structure of the contexts of a homonym should display a different topology or shape from that of the contexts of a relatively unambiguous word with essentially a single meaning (hereafter a *monoseme*). In a similar vein, a *polyseme*, with multiple related interpretations, such as a *chicken* referring to a FARM ANIMAL or to MEAT OF THAT ANIMAL, should also differ from a monoseme in the similarity structure of its contexts, even if those have some semantic overlap.

Given that ambiguous words make up the bulk of content words in language (Klein & Murphy, 2001), understanding how the interpretations of semantically ambiguous words are resolved by context, and the trace this process leaves in lexical representations, is key to advancing multidisciplinary research in this area. In particular, it may contribute to theoretical debates regarding some apparently inconsistent experimental results obtained using monosemes, polysemes, and homonyms. For example, some lexical decision experiments indicate an overall processing advantage for all ambiguous words, with no differences between polysemes and homonyms (e.g., Hino et al., 2006). If these results correlate with the differences in the topology of contexts noted above, they suggest that homonyms and polysemes occur in equally diverse contexts, and in more diverse contexts than monosemes. In contrast, other lexical decision experiments reveal a processing advantage only for polysemes, whereas homonyms show a processing *disadvantage* relative to monosemes (e.g., Armstrong & Plaut, 2016). These results suggest that the structure of the contexts – whether the diverse contexts of a word arise due to related or unrelated senses – may differently impact the representation of polysemes and homonyms.

Based on past work, it is therefore unclear how the representations of monosemes, polysemes, and homonyms differ. Here, we ask whether such representational differences are a by-product of systematic differences between the range of contexts in which each type of word is encountered—specifically, differences in the topology of each word's contexts. Do both types of ambiguous words occur in more distinctive contexts than monosemes? Is the relatedness of a polyseme's interpretations associated with a greater degree of similarity in its set of contexts as compared to homonyms?

To investigate these questions, we draw on the wealth of research in both psycholinguistics and computational modeling on vector-based representations of word meanings (e.g., Landauer & Dumais, 1997; Erk, 2012). These approaches use the aggregate contexts of a word to represent word meaning as a distributional semantic vector (DSV) in a high-dimensional space. Following Firth (1957), and given the above-noted



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

Figure 1: Multidimensional scaling plots for the GloVe vectors of three target words (in capitals) and of words in their dictionary definitions (cf. Exp. 1) 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 'x's, definition 2 in green triangles, definition 3 in blue squares). The same two-dimensional space was used for all words so variation in the distances can be compared across plots.

importance of contextual diversity in learning and organizing lexical meaning, it is perhaps not surprising that such context-based DSVs have been successful in modeling human behaviour in a number of semantic tasks, such as similarity judgments and analogy completion (e.g., McNamara, 2011; Baroni et al., 2014). Recent work has further examined whether the meanings of an ambiguous word can be usefully "extracted" from its DSV (e.g., Arora et al., 2016).

We build on such work with a new approach to using the spatial relations among DSVs to inform us about *the meaning structure* of a word. The overarching goal of the present work is quite straightforward: Do the properties of DSVs reflect the number and relatedness of a word's interpretations? To gain insight into this issue that is robust and relevant to a wide array of psycholinguistic and computational researchers, we examined the distance between a word's DSV and the DSVs of its contexts as defined in various ways.

Our Approach: Delineating Monosemes, Polysemes, and Homonyms

Above, we noted the convergence between research on contextual diversity as a key property of word *usages*, and the use of context-derived word *representations* (distributional semantic vectors, or DSVs). Our goal is to see whether the "traces" of contextual diversity that are encoded in DSVs are revealing about the properties of a word's semantics: especially whether the *similarity structure* of the contexts is encoded in a DSV. In particular, do properties of DSVs—which are created by aggregating over the contexts of usages of words—tell us something about the number and relatedness of interpretations that a word encodes?

To this end, we investigated how DSVs relate to relevant portions of the semantic space they occur in. Following much previous work (McNamara, 2011; Baroni et al., 2014), we assume that the semantic similarity between two DSVs is indicated by their relative positioning in the high-dimensional semantic space: DSVs that cluster in a similar region of the space are more semantically similar than those that are more spread out in the space.

With this in mind, Figure 1 illustrates the components of our main hypothesis here: (1) We assume that the DSV for a monoseme aggregates over relatively similar contexts; since its resulting vector representation has fewer distinctions to encode, the expected distance between it and (a vector representing) any of those contexts should be relatively small. (2) A DSV for a polyseme will be relatively more distant from (the vectors for) its contexts, since the varying contexts of its multiple senses pull its word vector representation somewhat away from any one particular context. (3) A DSV for a homonym will be the most distant from its contexts, since its encoding reflects contexts in which it has various interpretations with no overlap in semantics; the resulting DSV must encode and thus "sit between" these more distant, non-overlapping contexts.

To test this hypothesis, we require DSVs for a set of target words from our three *ambiguity types* (monosemes, polysemes, and homonyms), and DSVs that represent the contexts of these target words, so that we can measure the distance between them. We use standard, off-the-shelf DSVs whose usage is widespread in psycholinguistics and computational linguistics. Several models were included to explore whether any potential differences we find among the ambiguity types are robust to the particular methods for creating DSVs.

There are various ways to identify relevant contexts to compare these target items to. First, we consider—as an extreme example of disambiguating context—the **dictionary definition(s)** of a word. Such definitions have been carefully constructed to elaborate the distinctive semantics of a word, and as such, they serve as a proxy to a set of very clearly biased contexts that reflect all of the word's interpretations, and potentially their relationship to one another. In this way, the

definitions of a target word should serve as highly effective "probes" of whether and how the word's interpretations are captured in its DSV. Going back to the lay-out of monosemes, polysemes, and homonyms in semantic space, we expect these definitional contexts to accurately pinpoint salient spatial regions whose distances to the target are highly informative (cf. Figure 1).

Second, we consider the actual **linguistic usage contexts** of a word. 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 word vectors, they are a natural probe to measure the extent to which the resulting DSV of a word is closer to or further from its contexts. These results should help reveal how the DSV is related to actual contextual aspects of its meaning, in contrast to the definitional aspects.

Finally, we consider the context of the target DSV as it is situated within the semantic space; that is, the context is the target's most semantically-similar **neighbours** (cf. Burgess, 2001). 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 neighbours. Again, going back to the lay-out of monosemes, polysemes, and homonyms in semantic space, we hypothesize that the various neighbours of a target word in semantic space may occur closer to or further from the target depending on the variety of their shared semantic dimensions.

In all cases, to compare these different types of contexts to our targets, we aggregate the DSVs of the content words in each instance of a context to form a single DSV (cf. Schütze, 1998), and compare that to the target DSV. Our key experimental measure is the average cosine distance between the DSV of the target word, and the DSVs of each of its contexts of a certain type.

Experimental Set-up

Target words

We selected target words with the aim of maximizing the ability to detect ambiguity effects while simultaneously ruling out effects from other potentially confounding properties. Additionally, selection was constrained to facilitate the re-use of these items in future coordinated psycholinguistic experiments. We started by collecting the intersection of words common to the following sources: the **SUBTL word frequency database** (derived from movie/television subtitles, Brysbaert & New, 2009), the **CMU pronouncing dictionary** (Weide, 1998), the **Yarkoni et al. (2008) measures of orthographic neighbourhood**, and the **Wordsmyth dictionary** (Parks et al., 1998).

The Wordsmyth dictionary includes separate entries for unrelated meanings, with related senses grouped under a single entry. Manual inspection of the definitions suggests that in a small number of cases these definitions may not cover some meanings of a word, and some choices of senses as related (or not) may not be accurate. However, overall the meaning/sense counts have been found to correlate significantly with ambiguity effects in several prior behavioural experiments (e.g., Rodd et al., 2002). Thus, this source should be suitable for delineating between monosemes which are (relatively) unambiguous and which have only a single meaning/sense, polysemes which have multiple related senses, and homonyms which have multiple unrelated meanings (and possibly related senses within those, given the rarity of homonyms with only one sense per meaning). The eDom norms (Armstrong, Tokowicz, & Plaut, 2012) were also used to further filter the Wordsmyth homonyms in particular, because this resource includes a large set of pre-screened homonyms suitable for psycholinguistic experimentation, as well as norms on additional psycholinguistic properties of interest for later studies.

Table 1: Features used in matching and as covariates

Property	monosemes	polysemes	homonyms
# Unrelated Meanings	1 (0)	1 (0)	2.2 (0.02)
# Related Senses	1 (0)	5.72 (0.1)	7.48 (0.21)
# Noun Interp.	0.59 (0.02)	3.03 (0.07)	3.76 (0.11)
# Verb Interp.	0.14 (0.02)	2.21 (0.11)	3.06 (0.14)
# Adjective Interp.	0.16 (0.02)	0.42 (0.05)	0.58 (0.06)
# Letters	5.4 (0.07)	5.4 (0.07)	4.63 (0.06)
# Phonemes	4.5 (0.06)	4.5 (0.06)	3.86 (0.05)
# Syllables	1.6 (0.03)	1.6 (0.03)	1.3 (0.02)
ln(Word Freq. + 1)	0.89 (0.02)	0.94 (0.02)	0.95 (0.02)
Case of first letter	0.84 (0.02)	0.93 (0.01)	0.91 (0.01)
Coltheart's N. Orth.	3.45 (0.21)	3.71 (0.22)	6.85 (0.29)
OLD20	1.9 (0.02)	1.8 (0.02)	1.56 (0.02)
Coltheart's N. Phon.	7.14 (0.38)	7.19 (0.39)	12.12 (0.46)
PLD20	1.7 (0.03)	1.7 (0.03)	1.39 (0.02)
Pos. Letter Freq.	1183 (28)	1123 (28)	901 (26)
Pos. Bigram Freq.	155 (7)	145 (7)	115 (6)

_ Interp. = Interpretations associated with a part of speech. Case of first letter = Does the word most frequently appear with the first letter in uppercase (0)? Coltheart's N = Number of neighbours based on letter (Orth.) or phoneme (Phon.) substitution. OLD20/PLD20 = Orthographic/Phonological Levenshtein distance. Pos. Letter/Bigram = Freq. of a letter/bigram in a given position in a word. Per ambiguity type, mean and (variance) are reported

After combining these databases and removing words with less than two phonemes, we obtained 429 homonyms, 4672 polysemes, and 1229 monosemes. We then selected 429 polysemes and monosemes that were matched to the homonyms to the greatest extent possible at the item level on a number of psycholinguistic covariates (including, for polysemes, the number of senses), using the SOS stimulus optimization software (Armstrong, Watson, & Plaut, 2012). The covariates, along with their descriptive statistics, are presented in Table 1. Overall, the optimization created groups of monosemes, polysemes, and homonyms that are very similar—but not identical—in these statistics. Possible effects of the remaining imperfections in the matching were addressed in the analysis (see Statistical Methods).

Vector spaces

To evaluate the robustness of our findings and determine whether there are major differences across different implementations of word co-occurrence models and corpora, we replicated our computational experiments on three sets of pre-

trained vectors:[1] **LSA** (Landauer & Dumais, 1997), trained on the TASA corpus (Günther et al., 2015), and **GloVe** (Pennington et al., 2014) and **Word2vec** (Continuous Skipgram; Mikolov et al., 2013), both trained on English Wikipedia and Gigaword (Fares et al., 2017). The LSA vectors used here are a standard set that has been the subject of extensive research over 20 years. The GloVe and Word2vec sets represent two contemporary and very popular models trained on identical natural language corpora. All vectors have 300 dimensions.

Experiments

The distance from the target word to its set of contexts was defined as the mean of the cosine distances between the target DSV and each of its context DSVs. We ran 3 experiments, each using different context types:

Experiment 1: Each definition context DSV is formed from a single WordSmyth definition by averaging the DSVs of all gloss words (omitting stopwords; Bird et al. (2009)).

Experiment 2: Linguistic usages are lines of dialog containing the target word, extracted from the Subtlex corpus (Brysbaert & New, 2009). Each usage context DSV is formed by averaging the DSVs of all words in the corpus line (excluding stopwords and the target word itself).

Experiment 3: Neighbour contexts were the 20 DSVs with the lowest cosine distance from the target in the vector space.

Statistical methods

We used a stepwise multiple linear regression procedure to test for differences between ambiguity types in each experiment. First, we regressed out the effects of the psycholinguistic covariates in Table 1 (omitting Colheart's N Orth and N Phon to avoid collinearity with OLD and PLD). Then we tested for significant differences between the ambiguity types on the residual differences; these are the results reported below. In these analyses, the baseline level of ambiguity type was rotated to run all pairwise comparisons between types. The Type-I error rates in each experiment were held constant at $p < .05$ (2-tailed) using the Bonferroni-Holm procedure.

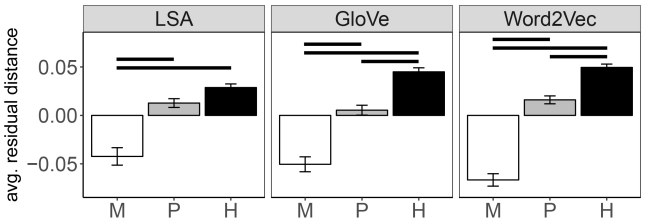
Results

To reiterate our hypotheses underlying the motivation for our experiments, we analyzed our experimental data to determine whether monosemes were closer to their contexts than polysemes, and whether polysemes were closer to their contexts than homonyms (cf. Figure 1) .

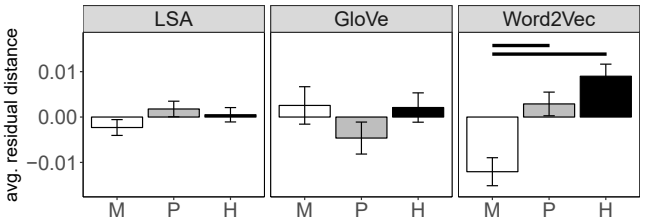
Experiment 1: Distance to dictionary definitions

The mean residual distances between the target DSVs and their definition DSVs for each of our ambiguity types (M = monoseme, P = polyseme; H = homonym), are presented in Figure 2a, with statistically significant comparisons noted.

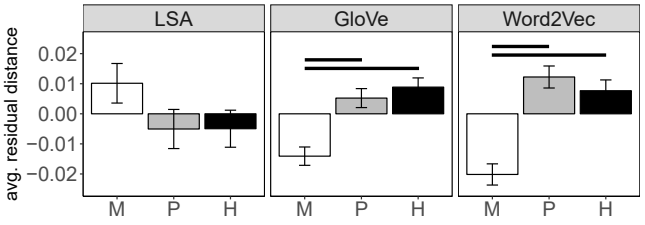
In line with our predictions, Glove and Word2Vec showed a significant 3-way distinction wherein the distance of the target



(a) Experiment 1: Dictionary definitions



(b) Experiment 2: Linguistic usage contexts



(c) Experiment 3: Semantic neighbourhoods

Figure 2: Ave. residual distance by ambiguity type for the different vector sets in each experiment. Error bars plot the standard error. Black lines denote significant differences between ambiguity types; $p < .001$ in all cases except Expt. 2, Word2vec M–P $p < .01$.

DSV to the definition DSVs was smallest for the monosemes and largest for the homonyms, with the polysemes falling in between. LSA also showed the 3-way trend numerically, but P vs. H was only trending ($p = .06$). The sample words from this experiment used to generate Figure 1 underscore this point: the definitions of the monoseme are most tightly clustered around the target word, followed by the polyseme, and with the homonym displaying the most dispersed set of definition words in vector space. This indicates that when a DSV is trained on samples of natural language text, probing it with dictionary definition "contexts" can reveal traces of the diversity of contexts in which the word was encountered. The similar patterns obtained across all three vector space models suggest that the overall co-occurrence structure in language drives these effects, given the differences in implementation and training corpora across our three vector sets.

Experiment 2: Distance to linguistic usage contexts

Figure 2b presents the mean residual distances between the target DSVs and their linguistic usage DSVs for each of our ambiguity types. It shows that for Word2vec, the contexts

[1] Gathered from `http://vectors.nlpl.eu/repository/` and `http://www.lingexp.uni-tuebingen.de/z2/LSAspaces/`

for both types of ambiguous words were significantly more distant from the target word than those of monosemes, but there was no difference as a function of the relatedness of an ambiguous word's interpretations. Given that homonym meanings are completely unrelated to one another, the fact that polysemes group with homonyms suggests that despite having related senses, each sense is, on average, encountered in just as distinct a context as an unrelated meaning. Thus, discussions of COOKING versus RAISING *chickens* may be as distinct as discussions of FINANCIAL versus RIVER *banks*.

Neither LSA nor GloVe showed significant effects across ambiguity types. Because LSA was trained on a different corpus than GloVe and Word2vec, we cannot infer whether its training corpus does not contain the same diversity of contexts needed to observe differences between ambiguity types, or whether this failure is due to how that algorithm creates DSVs.

However, GloVe was trained on the same corpus as Word2vec, so the differences between these two algorithms are attributable to how each model creates DSVs. A possible source of these differences is that in creating DSVs, GloVe optimizes over the entire word co-occurrence matrix at once, where Word2vec does iterative sentence-by-sentence training. Fleshing out the impact of these types of algorithmic differences is an active area of research (e.g., Rubin et al., 2014).

Experiment 3: Distance to semantic neighbourhood

Figure 2c presents the mean residual distances between the target DSVs and their neighbourhood DSVs for each of our ambiguity types. For Word2vec and GloVe, we observe a similar two-way split as for Word2vec in Experiment 2: homonyms and polysemes are not statistically different from one another, but both show larger distances than the monosemes. In other words, these results indicate that words with multiple interpretations have nearest neighbours that are more distant than those of monosemes, but the distance is not impacted by the relatedness of an ambiguous word's meanings. No significant effects were observed when analyzing the LSA vectors.

Supplemental Analysis: Extreme Polysemes

Our initial set of polysemes had slightly fewer senses than the homonyms did when matching for our broad set of psycholinguistic covariates. If the number of senses rather than their nature drives some of our results, we would expect polysemes with more senses to have a greater average distance to their contexts in the various experiments than both homonyms and polysemes. To evaluate this possibility, we repeated our experiments on an additional set of "extreme polysemes" that were matched on all covariates except the number of senses, and so had 38% more senses than the homonyms (10.4 vs. 7.5). No evidence for a confounding effect of number of senses was observed in Experiments 1 and 2, where numerically the extreme polysemes ranked somewhere in between the homonyms and polysemes. In Experiment 3 we did find the extreme polysemes had either the numerically largest (GloVe and Word2vec) or the smallest (LSA) distance to their semantic neighbourhood, suggesting that here the number of senses does affect the neighbourhood density differently in the case of different algorithms and/or corpora. Future studies are planned to better understand these phenomena.

General Discussion

Our 3 experiments tested for differences in the word vectors between the different ambiguity types in several types of "contexts": dictionary definitions that highlighted the defining or prototypical semantic dimensions of the words (Experiment 1), linguistic usage contexts that emphasized the co-occurrence relations of a word (Experiment 2), and neighbours in the vector space that measure the direct relation that a word has to related words (Experiment 3). Our key findings were as follows: In Experiment 1, all models generated a three-way distinction in which monosemes were further from their contexts than polysemes, and polysemes were further from their contexts than homonyms. In contrast, in Experiment 2 (Word2vec) and Experiment 3 (Word2vec & GloVe), we only observed a two-way split wherein ambiguous words were further from their contexts relative to monosemes. These findings support our highest-level intuition that (at least some types of) ambiguous words are encountered in more diverse contexts than monosemes. The contrast between natural contexts and the semantically more distinct contexts of dictionary definitions have interesting implications for theories and studies of ambiguous word representation, learning, and processing that transcend multiple disciplines of cognitive science.

The discrepancies across our experiments pose a paradox for theories of word learning. Whereas dictionary definitions allowed us to distinguish homonyms from polysemes, our two experiments based on naturally-occurring language did not – there, polysemes and homonyms are equally distant from their contexts. How is it that vectors that (as one test shows) are sensitive to meaning relatedness can be learned from contexts that (on average) do not reflect this factor? One potential answer is that the observed difference in Experiment 1 is not due to the target word vectors themselves, but to the definition words and how their vectors are spread out in vector space. This would resolve the paradox by denying that the homonym-polyseme distinction is 'contained' in the word vectors. Another possibility is that the natural language contexts and the semantic neighbourhoods are insufficiently strong probes into this difference; this would resolve the paradox by assuming that the homonym-polyseme distinction *is* captured in the word vectors, but that one needs very strong probes to reveal it. If the answer turns out to be that the distributional contexts *do not* encode the homonymy-polysemy distinction, this would raise a second question of how human language learners arrive at these distinct kinds of representations. In particular, such a result would suggest that, despite the success of word vectors based on linguistic context alone, they cannot capture all of the knowledge people have of word meaning and its organization.

Another factor that may have contributed to our discrepant results is how we divided words into three ambiguity types. The substantial amount of variance within each of the types

may reflect systematic variation better explained through an "ambiguity continuum" – in which the relatedness of senses varies continuously (e.g., Klepousniotou et al., 2008). Already, the adoption of such a graded view of ambiguity has been shown to modulate behavioural semantic ambiguity effects, such that polysemes assumed to have lower representational overlap across senses (e.g., metaphorical polysemes such as FILM vs. CELESTIAL *star*) produce ambiguity effects more similar to homonyms than polysemes assumed to have higher representational overlap across senses (e.g., metonymic polysemes such as *chicken*; Klepousniotou et al. 2008). Similarly, the inconsistency in finding two-way (Hino et al., 2006) or three-way (Armstrong & Plaut, 2016) distinctions may relate to how researchers divide their items into ambiguity types. Our computational approach makes it possible to evaluate whether and how other ways of measuring the number and relatedness of a word's meanings align with our measures here. It potentially reconciles these effects based on where words sit on the ambiguity continuum.

In conclusion, the alignment of the particular inconsistencies across our experiments with other discrepancies in the literature provides general support for our initial hypotheses and hints at a unifying account of these findings based on an ambiguity continuum. Thus, our simple approach of using three ambiguity types is clearly only a starting point, whereas future work should consider graded transitions in representational overlap, among many other factors (e.g., possible interactions with grammatical class and meaning frequency; Armstrong, Tokowicz, & Plaut 2012), as well as other ways of measuring the dispersion of vectors in semantic space. Our selection of a large set of stimuli that are suitable for psycholinguistic experimentation allows us to drill in on these possibilities in future analyses, as well as to evaluate how our computational results align with coordinated experiments using the same items. Thus, our interdisciplinary approach provides targeted directions for advancing the study of ambiguity by asking the question: can we know a word's ambiguity by the company that it keeps?

Acknowledgments: SS, SM, BB are supported by NSERC grant RGPIN-2017-06506; BCA by NSERC grant RGPIN-2017-06310.

References

Armstrong, B. C., & 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., & Plaut, D. C. (2012). eDom: Norming software and relative meaning frequencies for 544 English homonyms. *BRM*, *44*, 1015-1027.

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

Arora, S., Li, Y., Liang, Y., Ma, T., & Risteski, A. (2016). Linear algebraic structure of word senses, with applications to polysemy. *arXiv preprint arXiv:1601.03764*.

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings ACL.*

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

Brysbaert, M., & 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. *BRM*, *41*(4), 977–990.

Burgess, C. (2001). Representing and resolving semantic ambiguity: A contribution from high-dimensional memory modeling. In D. S. Gorfein (Ed.), *On the consequences of meaning selection: Perspectives on resolving lexical ambiguity* (pp. 233–261). American Psychological Association.

Erk, K. (2012). Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compass*, *6*(10), 635-653.

Fares, M., Kutuzov, A., Oepen, S., & 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.*

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

Günther, F., Dudschig, C., & Kaup, B. (2015, Dec 01). LSAfun - an R package for computations based on Latent Semantic Analysis. *BRM*, *47*(4), 930–944.

Hino, Y., Pexman, P. M., & Lupker, S. J. (2006). Ambiguity and relatedness effects in semantic tasks: Are they due to semantic coding? *JML*, *55*(2), 247–273.

Jones, M. N., Dye, M., & Johns, B. T. (2017). Context as an organizing principle of the lexicon. In *Psychology of learning and motivation* (Vol. 67, pp. 239–283). Elsevier.

Kachergis, G., Yu, C., & Shiffrin, R. M. (2017). A bootstrapping model of frequency and context effects in word learning. *Cognitive Science*, *41*(3), 590–622.

Klein, D. E., & Murphy, G. L. (2001). The representation of polysemous words. *JML*, *45*(2), 259–282.

Klepousniotou, E., Titone, D., & Romero, C. (2008). Making sense of word senses: The comprehension of polysemy depends on sense overlap. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(6), 1534–1543.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychol. Rev.*, *104*(2), 211.

McNamara, D. S. (2011). Computational methods to extract meaning from text and advance theories of human cognition. *Topics in Cognitive Science*, *3*(1), 3–17.

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

Parks, R., Ray, J., & Bland, S. (1998). *Wordsmyth English Dictionary-thesaurus [Retrieved September 2008 from wordsmyth.net]* (Vol. 1).

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings EMNLP.*

Rodd, J. M., Gaskell, G., & Marslen-Wilson, W. (2002). Making sense of semantic ambiguity: Semantic competition in lexical access. *JML*, *46*(2), 245–266.

Rubin, T., Kievit-Kylar, B., Willits, J. A., & Jones, M. N. (2014). Organizing the space and behavior of semantic models. In *Proceedings CogSci.*

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

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*(3), 1558–1568.

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

Yarkoni, T., Balota, D. A., & Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, *15*(5), 971-979.