Modeling the Role of Polysemy in Verb Categorization

Elizabeth Soper & Jean-Pierre Koenig*

Abstract. Recent work has indicated that static word embeddings can predict human semantic categories (Majewska et al. 2021). In this paper, we consider the role of polysemy in semantic categorization, by comparing sense-level embeddings with previously studied static embeddings in their prediction of human-produced categories. We find that the polysemy is crucial for predicting human categories; sense-level embeddings dramatically outperform static embeddings in predicting semantic categories. Our findings highlight the role of polysemy in semantic categorization that is exclusively based on linguistic input.

Keywords. Distributional semantics; polysemy; categorization; natural language processing.

1. Introduction. As we learn language and learn about the world, we acquire semantic knowledge. This knowledge comes both from perceptual input (what we experience first-hand) as well as linguistic input (what we hear and read about). We organize our knowledge of categories, and the words we use to denote them, based on connections we form between words and the things they denote in the world, as well as connections between words and other words. In this paper, we are interested in exploring two interrelated issues: (1) how much of human categories may come from what we hear or read? and (2) whether partitioning word representations into similar contexts of use (a proxy for fine-grained polysemy) improves the fit of linguistic knowledge to human categories?

Distributional semantic models are well-suited to help us answer these two questions as they use linguistic co-occurrence statistics from a corpus to create vector representations of word meanings. Words with similar meanings, which occur in similar contexts, end up near each other in space, while unrelated words, which occur in very different contexts, end up far apart in space. Word embeddings have become popular in recent years, in particular for their ability to accurately predict the similarity between words (Landauer & Dumais 1997, Pereira et al. 2016, Devlin et al. 2019). Since similarity is a primary criterion for categorization (Collins & Loftus 1975), word embeddings may also be good at predicting semantic categories.

Because word embeddings are trained on linguistic input only, with none of the perceptual input that humans receive, they are ideal for studying the unique role of language in semantic categorization. Additionally, because current word embeddings keep track of the contexts in which words are found, we can look specifically at how polysemy affects how semantic categorization can derive from linguistic knowledge. As words generally have multiple possible senses, categorization decisions may depend on which which sense of a word is being considered. Representing the distinct senses of polysemous words is thus likely to be important to how humans categorize sets of verbs' denotations and how these categories to denotations are used to approximate categories

*The authors gratefully acknowledge the audience at the Experiments in Linguistic Meaning conference, hosted by the University of Pennsylvania, for their useful comments and feedback. Authors: Elizabeth Soper, SUNY at Buffalo (esoper@buffalo.edu) & Jean-Pierre Koenig, SUNY at Buffalo (jpkoenig@buffalo.edu).

of objects. We compare different types of word embeddings when it comes to predicting human categories, and find that representing individual senses is indeed crucial for predicting semantic categories.

2. Background. There are two main classes of word embedding models: traditional static embeddings (Landauer & Dumais 1997, Mikolov et al. 2013, Pennington et al. 2014) represent each word type as a unique vector, while more recent contextual models (Peters et al. 2018, Devlin et al. 2019) generate a unique representation for every instance of a word in context. Since both static and contextual embeddings have been shown to model pairwise similarity between words well (Pereira et al. 2016, Chronis & Erk 2020), and since similarity is a primary criterion for categorization, it seems intuitive that word embeddings should predict categorization well. Some previous work supports this intuition; word embeddings have excelled at word sense disambiguation (Giulianelli et al. 2020, Soler & Apidianaki 2021, Chronis & Erk 2020) and topic modeling (Sia et al. 2020, Aharoni & Goldberg 2020), when cast as categorization problems.

As mentioned in the introduction, we are interested in the present paper in linguistically-based semantic category induction. Instead of grouping individual word tokens into distinct senses, or documents into topics, the goal of semantic categorization is to group unique words into semantically related clusters. This more abstract type of categorization has received less attention in the word embedding literature.

In order to understand what word embeddings can tell us about the role of polysemy in deriving semantic categories from linguistic input, it is important to note that there are at least two different approaches to explaining polysemy. One account holds that polysemous words have a single, under-specified meaning, and that context helps disambiguate between a defined set of senses (Pustejovsky 1998). Static models are analogous to this view of meaning, as they create a single representation which is meant to encompass all uses of a word form. A stronger claim has been made (for example, by Elman 2009; see also Marvel & Koenig 2015) that different senses of a word are not simply reflected in, but actually *created by* context. Proponents of this claim believe that word meaning is fundamentally context-dependent. Contextual language models like BERT implicitly take this view of word meaning, as they represent each instance of a word in a particular context as a unique embedding. We can then treat classes of contexts as equivalent to senses in these models.

Because static and contextual models align with different theoretical perspectives on polysemy, word embeddings are well-suited to test these two conceptions of polysemy and their effect on linguistically-based category induction; static embeddings are a proxy for the single-entry view, and contextual embeddings for the radically context-dependent view. Previous work using word embeddings to model semantic categorization has treated semantic categorization as categorization of word types, rather than word senses. By using static embeddings only, it implicitly assumed that people use a summary representation to categorize words. While summary or underspecified representations of word meanings is appropriate for many tasks, as Pustejovsky (1998) shows, much work in psycholinguistics suggests that when reading words in and out of context, native speakers often favor one sense or the other (see, among others, Brocher et al. 2018 for evidence and summary). In using linguistic input to derive (part of) their semantic knowledge, learners therefore most likely rely on the word sense instantiated in the input. As a result, categorization decisions

may depend on which sense of a word is being considered and any attempt to determine how much semantic knowledge may derive from linguistic input would be best served by taking polysemy into consideration. If this is the case, we would expect contextual embeddings to predict semantic categories better than static embeddings.

Interestingly, recent work evaluating different word embedding models on verb categorization suggests just the opposite. Majewska et al. (2021) found that contextual models perform poorly compared to older static models when approximating the verb categorization done by participants in their experiments. We argue that this result is due not to the irrelevance of context to categorization, but rather to the way the contextual embeddings were extracted from the model in Majewska et al. (2021). Although many of the words in Majewska et al. (2021)'s ground truth data are polysemous and are assigned to multiple categories by participants, they evaluate models in a one-representation-per-word-form manner. Even when evaluating BERT, which has been shown to encode sense-specific information (Chronis & Erk 2020, Soler & Apidianaki 2021), this information was thrown away, either by feeding words to the model in isolation or by averaging over all contexts. Because they use polysemous data to test representations which do not encode sense information, Majewska et al. (2021)'s results may not reflect the full potential of contextual architectures to model categorization.

In this study, we test the ability of static and sense-specific embeddings to predict human-produced categories. Our ultimate goal is to determine the role linguistic input plays in semantic knowledge; our more modest goal in this paper is to find out whether polysemy matters in modeling verb categorization. Our overall finding is that, as we predicted, sense-specific embeddings are much better at predicting human behavior than static embeddings.

3. Dataset. We use the Phase 1 data from SpA-Verb (Majewska et al. 2021), which contains 825 verbs sorted into 17 broad classes. 10 participants were asked to sort words by clicking and dragging each word into a circle (see Figure 1). There were no constraints on the number or size of groups participants could create; they were only asked to sort them according to their meaning. Table 1 gives an overview of the classes resulting from this sorting task. 116 verbs belong to more than one class. No words were assigned to more than 3 classes. On average, words are assigned to 1.14 classes.

The SpA-Verb dataset is particularly relevant for our goals because the categories reflect actual human behavior, rather than expert-curated categories. Additionally, since words were presented in isolation with no disambiguating context, using this dataset is a strong test of the role of polysemy. If people categorize according to individual senses rather than a summary representation of each word, even when shown words out of context, this would support a context-sensitive view of meaning and suggest that polysemy matters when building semantic knowledge from linguistic input.

4. Models. Next we describe the models we compared against our baseline of human performance.

4.1. WORD2VEC. The first model we evaluate is a word2vec model trained on part-of-speech-tagged data (Fares et al. 2017). Part-of-speech tagging allows the static model to distinguish between senses which have different parts of speech (e.g. *duck_NOUN* and *duck_VERB*), al-

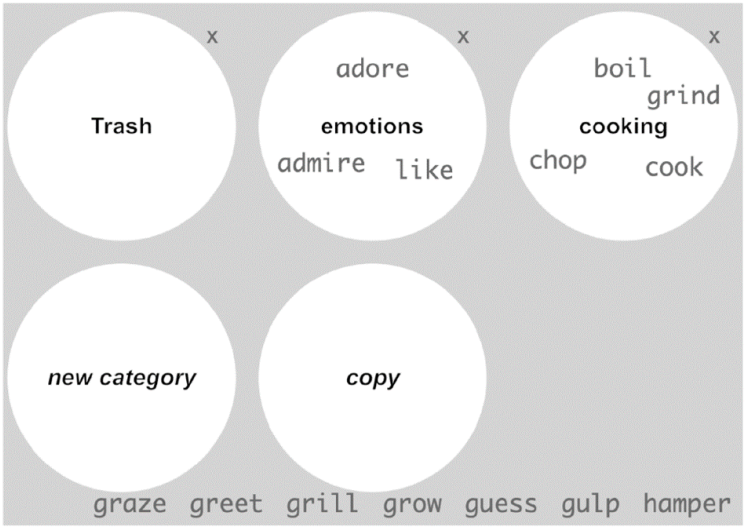

Figure 1: Screen interface for the SpA-Verb sorting task (Majewska et al. 2021)

Cluster label	**Example verbs**
movement	*wander, fly, glide, roam*
communication	*persuade, command, tell*
crime & law	*beat, abduct, abuse, shoot*
negative emotion	*offend, aggravate, enrage*
positive emotion	*admire, respect, adore, like*
cognitive process	*suppose, assume, realize*
cooking	*cook, slice, stew, boil*
possession	*belong, obtain, acquire*

Table 1: A sample of the 17 gold classes in SpA-Verb dataset (labels are given for descriptive purposes only)

though senses which have the same part of speech are still conflated into a single vector (e.g. *get#ACQUIRE* and *get#UNDERSTA-ND*). Skip-gram with negative sampling was used to train the model on Gigaword 5th Edition (Parker et al. 2011), with a context window size of 5 and 300 dimensions. Six words from SpA-Verb (*do, have, be, own, slurp*, and *snooze*) were not in the model's vocabulary (due either to removal of stop words from corpus or rare words not occurring in corpus) and were thus excluded from our analysis.

4.2. BERT. We evaluate three methods of extracting embeddings from the contextual BERT model: two baseline methods, which create one representation per word form, and a multi-prototype method which generates one representation per word *sense*. For all methods, we use BERT Base Uncased from HuggingFace's transformers package (Wolf et al. 2020).

4.2.1. DECONTEXTUALIZED (DECONT). First and most simply, we extract embeddings from BERT by feeding each word to the model in isolation. This creates a single, static embedding for each word. This strategy has been used previously as a way to easily extract 'context-free' representations from BERT (Liu et al. 2019, Vulić et al. 2020).

4.2.2. AGGREGATED (AGGR). Next, we create static embeddings from BERT by averaging a word's embeddings across 100 unique contexts. This aggregated approach still reduces a word to a single representation, but has been shown to produce higher quality representations than the decontextualized strategy (Bommasani et al. 2020).

4.2.3. MULTIPROTOTYPE (MPRO). Finally, to test whether sense-specific information is important to semantic categorization, we distill token-level BERT embeddings into multiple prototype embeddings. We use the method of Chronis & Erk (2020) to generate representations which correspond to different senses of a word, without collapsing every token into a single representation. Multi-prototype embeddings were generated as follows:

1. For each verb in the dataset, we sampled up to 100 sentences from the British National Corpus (BNC Consortium 2007), excluding non-verbal uses of the target word. A few words in the set occurred in the corpus fewer than 100 times. Four words (*broil*, *corrupt*, *exhale*, and *misspend*) did not occur as verbs at all in the corpus and were excluded from our analysis. The average number of occurrences sampled for a word was 95.6.
2. We extract BERT token embeddings for each collected occurrence of a word. For words which BERT tokenizes into multiple word pieces, we average over all component pieces.
3. We cluster the token embeddings for each verb. Like Chronis & Erk (2020), we use k-means clustering to group tokens into 'sense' clusters. We use the number of verb senses listed in WordNet (Miller 1995) to determine the appropriate k for each word. Verbs in the dataset had on average 5.9 senses. (min: 1, max: 59, for *buzz*).
4. After identifying clusters, we take the k cluster centroids for each word. These are the embeddings we evaluate against the SpA-Verb categorization data.

4.3. RANDOM BASELINE. Finally, we generate random vectors and evaluate them in order to establish a baseline for random chance performance.

5. Evaluation. To evaluate the performance of each model against a 'gold standard' set of human-generated categories, k-means clustering is used to group verbs into predicted classes. We use the

same metrics as Majewska et al. (2021): modified purity and weighted class accuracy are combined in an F1 score, calculated as their balanced harmonic mean. Modified purity is the mean precision of predicted clusters:

$$\text{MPUR} = \frac{\sum_{C \in Clust, n_{prev(C)} > 1} n_{prev(C)}}{\#test\_verbs} \quad (1)$$

where each cluster C from the set of all K_{Clust} induced clusters *Clust* is associated with its prevalent gold class, and $n_{prev(C)}$ is the number of verbs in an induced cluster C taking that prevalent class, with all other verbs considered errors. $\#test\_verbs$ is the total number of verbs in the dataset. While modified purity is a measure of precision, weighted class accuracy targets recall:

$$\text{WACC} = \frac{\sum_{C \in Gold} n_{dom(C)}}{\#test\_verbs} \quad (2)$$

where for each class C from the set of gold standard classes *Gold*, we identify the dominant cluster from the set of induced clusters having most verbs in common with C ($n_{dom(C)}$).

Because MPro BERT has multiple representations for a single word, the same word form may show up more than once within a single cluster. To prevent artificially inflating the recall in evaluating MPro BERT, we eliminate duplicates within each cluster before evaluation.

6. Results. Table 2 shows the results of each embedding type, compared to results reported in Majewska et al. (2021). The baseline models (Decont. and Aggr. BERT) perform comparably to previously reported results. Part-of-speech-sensitive word2vec model scores about 10 points higher than reported for a similar model architecture without part-of-speech information. MPro BERT, by contrast, performs dramatically better than other embeddings, achieving more than double the F1 score of the best previously reported BERT results. This suggests that polysemy does play an important role in modeling linguistically-based semantic categorization: human categories that can be derived from linguistic input correspond to semantically similar contexts of use of words.

Model	F1-optimal	F1-gold
Random baseline	0.204	0.161
Majewska word2vec	0.355	0.326
Majewska best BERT	0.340	0.322
POS-tagged word2vec	0.442	0.433
Decont. BERT	0.309	0.191
Aggr. BERT	0.398	0.346
MPro BERT	**0.743**	**0.687**

Table 2: Average F1 across models on coarse-grained categories. 'Gold' is for k=17, as in the ground truth. 'Optimal' is best result for k in the range (5, 50).

The benefit of sense-specific embeddings for modeling coarse-grained categoirzation is clear in the example of *freeze*. In the ground truth data, *freeze* belongs to just one class, related to cooking (along with words like *bake, fry, melt,* and *thaw*). *Freeze* has another figurative sense, meaning to stop or suspend. Because the word is polysemous, static embedding clusters struggle to categorize

it appropriately. In the aggregated BERT clusters, *freeze* appears in a cluster predominated by verbs related to violence (*whip, shoot, choke, crush, smash*). Decontextualized BERT puts *freeze* in a heterogeneous cluster with a few cooking words (*melt, stew, fry*) but also many seemingly unrelated words (*knit, greet, disturb, wander*). It appears that the different senses of the word skew its static representation and prevent accurate classification. MPro BERT, by contrast, puts *freeze* in two clusters: one related to cooking (as in the ground truth) and another cluster with words like *stop, delay, arrest* and *restrict*, which seems to correspond to the figurative sense of *freeze*. Thus factoring out different senses allows MPro BERT to give a more accurate and reasonable categorization.

7. Discussion. One might argue that the superior performance of MPro BERT embeddings is due simply to the increased total number of embeddings for this setting, compared to the static embeddings; because MPro BERT has multiple representations per word form, it has more 'chances' to correctly categorize each word. We do find that the MPro BERT clusters have more members on average: the average ground truth class has 55.5 members, while for our optimal MPro BERT results, the average cluster size was 96.0 words. On average, one word form appeared in 3.02 clusters, but only in 1.14 ground truth classes. All else being equal, the larger cluster size and the fact that words appear in more clusters should lead to higher recall scores and lower precision. In fact, we find that MPro BERT embeddings have higher precision *and* higher recall scores than the static embeddings, confirming that the difference we see between MPro BERT and static embeddings is not merely a fluke; sense-level embeddings really do seem to better capture broad semantic categories better.

In fact, MPro BERT might be better at modeling human participants than suggested by our F1 measure. This is because MPro BERT tends to capture more distinct senses per word than human participants did, as human participants generally focused on a single sense when categorizing. For example, the word form *jump* occurs in one MPro BERT cluster corresponding to violence (*jump#ATTACK*), another cluster corresponding to physical movement (*jump#HOP*), and a third one related to change (*jump#INCREASE*). In the ground truth data, *jump* only occurs once, in a class related to physical movement. Perhaps this is the most salient sense of the word *jump*, and therefore participants were more likely to be thinking of this sense during the word sorting task and ignore its other possible senses. Our F1 score is therefore a conservative measure of the success of MPro BERT in modeling human categorization exclusively from linguistic input, in that it counts these other non-dominant senses of *jump* against MPro BERT in our evaluation. However, the fact that embeddings for *jump* were assigned three separate clusters is not necessarily a weakness: the MPro BERT clusters are more thorough as they represent each sense of the word separately and appropriately assign them to separate clusters. MPro BERT clusters correspond to all possible ways of classifying sets of contexts (sense) words appear in to derive semantic categories; human participants seem to be more likely to grab onto fewer or even only one of set of contexts (sense).

As the previous example suggests, our conservative F1 scores may not give a full picture of the quality or reasonableness of the word embedding clusters, as it abstracts away from the fact that human participants might only attend to a few or even one sense of words they semantically categorize even when those words are presented in isolation. Our current results also abstract away from an important aspect of categorization, namely its flexibility and context-dependency:

categorization is a relatively flexible task. There may be many possible criteria for sorting a group of words, especially when given such a large set of words to sort (Tversky 1977, Barsalou 1982). The flexibility of human categorization might explain the low inter-annotator agreement on the task; Majewska et al. (2021) measured the agreement between two initial test participants on the verb sorting task, which was just 0.400 B-Cubed score. This low agreement suggests that humans don't perform very consistently in creating broad semantic categories from a large group of words. Even within a single participant, categories were not created based on consistent criteria. As a result, it's possible for induced categories from word embeddings to be reasonable, but still correlate poorly with our ground truth data. To fully determine how much semantic knowledge can be extracted from linguistic input, it is critical for models based on word embeddings to be able to mimic the flexibility and contextual-dependence of human categorization.

8. Conclusion and Future Work. Majewska et al. (2021) found that static word2vec embeddings were better at predicting human categorization of verbs than contextual BERT embeddings. In this paper, we challenged this result, comparing sense-specific embeddings against previously evaluated static representations, and found that the rich, sense-specific information present in BERT allows it to excel at predicting semantic categories. On a more general level, these results suggest that linguistic input encodes a great deal of information about semantic categories, independently of other perceptual input that humans receive, and that this category information can be extracted from embedding models which are trained on linguistic data alone. These results also suggest that word meanings that derive from linguistic input are better able to model human categorization if they correspond to semantically similar sets of contexts (the distributional equivalent of word senses). More broadly, our research supports the words-as-cues theory put forth in Elman (2009) in that it is similar sets of contexts of use that best predict human categorization. Future work is needed, though, to extend this analysis to nouns, which may behave differently, as well as to further explore the role of language in forming semantic categories, and whether models like those discussed here can model the flexible, goal-dependent nature of human categorization.

References

Aharoni, Roee & Yoav Goldberg. 2020. Unsupervised domain clusters in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7747–7763. https://doi.org/10.18653/v1/2020.acl-main.692.

Barsalou, Lawrence W. 1982. Context-independent and context-dependent information in concepts. *Memory & Cognition* 10(1). 82–93. https://doi.org/10.3758/bf03197629.

BNC Consortium. 2007. British National Corpus. *Oxford Text Archive Core Collection* .

Bommasani, Rishi, Kelly Davis & Claire Cardie. 2020. Interpreting pretrained contextualized representations via reductions to static embeddings. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* 4758–4781. https://doi.org/10.18653/v1/2020.acl-main.431.

Brocher, Andreas, Jean-Pierre Koenig, Gail Mauner & Stephani Foraker. 2018. About sharing and commitment: The retrieval of biased and balanced irregular polysemes. *Language, Cognition, and Neuroscience* 33(4). 443–466. https://doi.org/10.1080/23273798.2017.1381748.

Chronis, Gabriella & Katrin Erk. 2020. When is a bishop not like a rook? When it's like

a rabbi! Multi-prototype BERT embeddings for estimating semantic relationships. *Proceedings of the 24th Conference on Computational Natural Language Learning* 227–244. https://doi.org/10.18653/v1/2020.conll-1.17.

Collins, Allan M. & Elizabeth F. Loftus. 1975. A spreading-activation theory of semantic processing. *Psychological Review* 82(6). 407–428. https://doi.org/10.1037/0033-295X.82.6.407.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee & Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* 4171–4186. https://doi.org/10.18653/v1/N19-1423.

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

Fares, Murhaf, Andrey Kutuzov, Stephan Oepen & Erik Velldal. 2017. Word vectors, reuse, and replicability: Towards a community repository of large-text resources. *Proceedings of the 21st Nordic Conference on Computational Linguistics, NoDaLiDa, 22-24 May 2017, Gothenburg, Sweden* 131. 271–276.

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

Landauer, Thomas K & Susan T Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review* 104(2). 211. https://doi.org/10.1037/0033-295x.104.2.211.

Liu, Qianchu, Diana McCarthy, Ivan Vulić & Anna Korhonen. 2019. Investigating cross-lingual alignment methods for contextualized embeddings with token-level evaluation. *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)* 33–43. https://doi.org/10.18653/v1/k19-1004.

Majewska, Olga, Diana McCarthy, Jasper JF van den Bosch, Nikolaus Kriegeskorte, Ivan Vulić & Anna Korhonen. 2021. Semantic data set construction from human clustering and spatial arrangement. *Computational Linguistics* 47(1). 69–116. https://doi.org/10.1162/coli_a_00396.

Marvel, Aron & Jean-Pierre Koenig. 2015. Event categorization beyond verb senses. *Proceedings of the 11th Workshop on Multiword Expressions (MWE 2015)* https://doi.org/10.3115/v1/w15-0913.

Mikolov, Tomas, Kai Chen, Greg Corrado & Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings* 1–12.

Miller, George A. 1995. Wordnet: a lexical database for english. *Communications of the ACM* 38(11). 39–41. https://doi.org/10.1145/219717.219748.

Parker, Robert, David Graff, Junbo Kong, Ke Chen & Kazuaki Maeda. 2011. *English gigaword fifth edition*. Linguistic Data Consortium.

Pennington, Jeffrey, Richard Socher & Christopher D Manning. 2014. Glove: Global vectors for

word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* 1532–1543. https://doi.org/10.3115/v1/d14-1162.

Pereira, Francisco, Samuel Gershman, Samuel Ritter & Matthew Botvinick. 2016. A comparative evaluation of off-the-shelf distributed semantic representations for modelling behavioural data. *Cognitive Neuropsychology* 33(3-4). 175–190. https://doi.org/10.1080/02643294.2016.1176907.

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

Pustejovsky, James. 1998. The semantics of lexical underspecification. *Folia lingüística: Acta Societatis Linguisticae Europaeae* 32(3). 323–348. https://doi.org/10.1515/flin.1998.32.3-4.323.

Sia, Suzanna, Ayush Dalmia & Sabrina J Mielke. 2020. Tired of topic models? clusters of pretrained word embeddings make for fast and good topics too! *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* 1728–1736. https://doi.org/10.18653/v1/2020.emnlp-main.135.

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

Tversky, Amos. 1977. Features of similarity. *Psychological Review* 84(4). 327–352. https://doi.org/10.1037/0033-295x.84.4.327.

Vulić, Ivan, Simon Baker, Edoardo Maria Ponti, Ulla Petti, Ira Leviant, Kelly Wing, Olga Majewska, Eden Bar, Matt Malone, Thierry Poibeau et al. 2020. Multi-simlex: A large-scale evaluation of multilingual and crosslingual lexical semantic similarity. *Computational Linguistics* 46(4). 847–897. https://doi.org/10.1162/coli_a_00391.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz et al. 2020. Transformers: State-of-the-art natural language processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP): System Demonstrations* 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6.