Investigating the Extent to which Distributional Semantics Models Capture a Broad Range of Semantic Relations

Kevin S. Brown[1,2], Eiling Yee[3], Gitte Joergensen[3], Melissa Troyer[4], Elliot Saltzman[5], Jay Rueckl[3], James S. Magnuson[3,6,7], and Ken McRae[4]

[1] Department of Pharmaceutical Sciences, Oregon State University, Corvallis, OR, USA

[2] School of Chemical, Biological, and Environmental Engineering, Oregon State University, Corvallis, OR, USA

[3] Department of Psychological Sciences, University of Connecticut, Storrs, CT, USA

[4] Department of Psychology, University of Western Ontario, London, Canada

[5] Department of Physical Therapy, Boston University, Boston, MA, USA

[6] BCBL. Basque Center on Cognition, Brain, & Language, Donostia-San Sebastián, Spain

[7] Ikerbasque. Basque Foundation for Science, Bilbao, Spain

Keywords: distributional semantics models; semantic relations; thematic fit; event-based relations; function, shape, and color relations

Corresponding Author:
Kevin S. Brown
Department of Pharmaceutical Sciences and School of Chemical, Biological, and Environmental Engineering
Oregon State University
Pharmacy Building 317
1601 SW Jefferson Ave.
Corvallis, OR 97331
Email: kevin.brown@oregonstate.edu

Phone: 541-737-8251

Abstract

Distributional Semantics Models (DSMs) are a primary method for distilling semantic information from corpora. However, a key question remains: What types of semantic relations among words do DSMs detect? Prior work typically has addressed this question using limited human data that are restricted to semantic similarity and/or general semantic relatedness. We tested eight DSMs that are popular in current cognitive and psycholinguistic research (PPMI; GloVe; and three variations each of Skip-gram and CBOW using word, context, and mean embeddings) on a theoretically-motivated, rich set of semantic relations involving words from multiple syntactic classes and spanning the abstract-concrete continuum (19 sets of ratings). We found that, overall, the DSMs are best at capturing overall semantic similarity, but also can capture verb-noun thematic role relations and noun-noun event-based relations that play important roles in sentence comprehension. Interestingly, Skip-gram and CBOW performed the best in terms of capturing similarity, whereas GloVe dominated on thematic role and event-based relations. We discuss theoretical and practical implications of our results, make recommendations for users of these models, and demonstrate significant differences in model performance on event-based relations.

1. Introduction

A great deal of theoretical, empirical, and computational research in cognitive science has been conducted to gain a better understanding of the nature of human semantic knowledge, and word meaning in particular. A particularly challenging prerequisite to studying words experimentally or computationally concerns how to characterize their meanings. The two most frequently used methods for constructing distributed semantic representations involve collecting data from human participants, or analyses of text corpora. For example, semantic representations for words referring to living things (*cow*), nonliving things (*tractor*), and action verbs and nouns (*to walk*, *a walk*) have been constructed using data collected from humans, such as semantic feature production norms (Buchanan, Valentine, & Maxwell, 2019; McRae, Cree, Seidenberg, & McNorgan, 2005; Vinson & Vigliocco, 2008). Such norms have been used with substantial success as the basis for human experiments (Mirman & Magnuson, 2008; Papies, 2013; Pexman, Hargreaves, Siakaluk, Bodner, & Pope, 2008), connectionist models (Andrews, Vinson, & Vigliocco, 2009; Rabovsky & McRae, 2014; Rogers et al., 2004) and network science models (Hills, Maouene, Maouene, Sheya, & Smith, 2009; Stella, Beckage, & Brede, 2017). A marked disadvantage however, of collecting empirical data from humans using methods such as feature norming is that such methods involve collecting huge amounts of data and entail exceedingly labor-intensive coding of those data. Furthermore, methods such as feature norming do not extend readily to abstract nouns such as *equity*, abstract verbs such as *racialize*, adjectives such as *awkward*, and adverbs such as *quickly*.

Representations of word meaning can also be constructed using corpus-based Distributional Semantic Models (DSMs), and these models have also played a significant role in research on semantic memory. DSMs provide representations for all content words that occur in

samples (documents, sentences) of usable text or transcribed speech. There are now many types of DSMs, and this area of research has been vibrant, varied, and prominent (for an excellent recent review, see Lenci, 2018). One class of DSMs is *count-based* and relies on counting word-word or word-document co-occurrences. This class yields semantic representations either directly from co-occurrence counts, as in HAL (Lund & Burgess, 1996), or transformed counts (e.g., using global matrix factorization), as in Positive Pointwise Mutual Information (PPMI; Bullinaria & Levy 2007), Latent Semantic Analysis (LSA; Landauer & Dumais, 1997), and Global Vectors (GloVe; Pennington et al., 2014). A second class consists of passive co-occurrence models that use the accumulation of random vectors as a mechanism for semantic abstraction (BEAGLE; Jones, Kintsch, & Mewhort, 2006). A third class includes retrieval-based models in which semantic memory consists of an episodic word-by-context matrix (Dennis, 2005; Kwantes, 2005). A fourth class of DSM is prediction-based and relies on feedforward connectionist networks in either of two training modes (word2vec, Mikolov et al., 2013). One mode (Skip-gram) involves using a target word as input to predict its set of surrounding context words as output, and the other (CBOW: Continuous Bag of Words) is trained using a target word's set of surrounding context words as input to predict the target word as output. When comparing Skip-gram and CBOW to human performance, given that the input words are represented as one-hot (localist) vectors, researchers typically use the vectors of input-to-hidden unit weights ("word embeddings") as semantic representations.

A primary goal of all of these DSMs is to construct vector representations that approximate word meaning. These representations can then be used to investigate people's knowledge of semantic relatedness and to provide the foundation for constructing experimental

items, models of semantic processing (Andrews et al., 2009; Rotaru, Vigliocco, & Frank, 2018), and network science models of lexical knowledge (Utsumi, 2015).

The goal of the present article is to investigate the degree to which four of the currently best-performing DSMs—the count-based PPMI and GloVe models and the prediction-based Skip-gram and CBOW models—are able to capture a broad range of semantic relations and to provide insights into human language processing. Our testbed consists of 19 sets of human ratings, which includes ratings of semantic similarity, general relatedness, verb-noun thematic role relations that are central to sentence processing, and noun-noun event-based relations. The primary innovation of our research is to investigate the extent to which eight DSMs, PPMI, GloVe, and three variations each of Skip-gram and CBOW[1], can account for a broad, theoretically central, and well-defined range of word-word relations.

1.1. Semantic Similarity and DSMs

In general, in the human semantic memory literature, semantic similarity is discussed and operationalized in terms of shared features (cars and trucks are both made of metal, have wheels, are used for transportation, are found on roads, are driven by humans, and so on) and/or belonging to the same category (taxonomic relations; cars and trucks are similar because both are vehicles).

In its strongest form, the distributional semantics hypothesis states that when considering words, this kind of semantic similarity can be captured using distributional co-occurrences (and transformations of them) that are inherent to linguistic input. That is, similarity is reflected in the degree to which words co-occur with the same sets of words (and/or in the same documents or

[1] The three variations of Skip-gram and CBOW are defined by their use of word embeddings, context embeddings (vectors of hidden-to-output unit weights), or the mean of these two embeddings.

conversational topics) in spoken and written human language. The output of DSMs are vector representations that typically are operationalized as word meanings. The cosine between a pair of word vectors typically is used to estimate similarity between the two words' meanings. Overall, DSMs have been successful in capturing similarity among word meanings (Baroni et al., 2014; Levy, Goldberg, & Dagan, 2015; Mandera et al., 2017). There are a large number of articles in which DSMs have been tested on their ability to capture human semantic judgments, and in some cases, decision latency data (e.g., semantic priming tasks). Studies of the degree to which DSMs can capture people's intuitions about similarity have played a major role in evaluating, contrasting, and refining multiple types of DSMs. Overall, the fields of computational linguistics, psycholinguistics, and human cognition have learned a great deal about the extent to which human conceptual knowledge can be captured in co-occurrence statistics that are present in language (see Kumar, 2021; Kumar, Steyvers, & Balota, 2021; Lenci, 2018, for recent reviews).

1.2. Other Semantic Relations and DSMs

There is no doubt that similarity plays a major role in human conceptual processing in general, and in language comprehension and production in particular. However, semantic similarity is far from the sole important type of semantic relation. In this article, we use "semantic relation" as a superordinate term. Thus, semantic similarity is a type of semantic relation. Furthermore, we discuss multiple other types of semantic relations below, including a number of thematic role and event-based relations (Carlson & Tanenhaus, 1988; Estes, Golonka, & Jones, 2011).

In some studies, DSMs have also been tested on their ability to account for what typically is called associative relatedness (Griffiths, Steyvers, & Tenenbaum, 2007; Rapp, 2002;

Sahlgren, 2006). Associative relatedness is measured using word association norms in which participants produce a word (or sometimes a chain of words) in response to a cue word (Nelson, McEvoy, & Schreiber, 1998; De Deyne, Navarro, & Storms, 2013). For example, given *mop* as a cue in Nelson et al. 24% of participants produced the thematically-related *floor*. It is important to note that the operationalization of associative relatedness is, essentially, its definition; that is, two words are associated if participants produce a response word given a cue word (McNamara, 2005; McRae, Khalkhali, & Hare, 2012). A key issue with association norms, however, is that participants' responses constitute a mixed set of semantic relations that include, for example, similar concepts (*fox-wolf*), antonyms (*light-dark*), features of the cue such as a part of an object (*tricycle-pedals*), function relations (*drill-carpentry*), and other thematic role (verb-agent: *arrest-police*; verb-patient: *arrest-criminal*) and event-based relations (*election-candidate*). Therefore, when testing whether DSMs can account for human associative relatedness as operationalized by word association norms, it is unclear what semantic relations actually are, or are not, being captured by the models. We therefore tested DSMs' ability to capture specific types of relatedness that are important for word and sentence processing, although one set of general relatedness ratings was included because it is part of a dataset that has been used extensively by computational linguistics to compare models.

In addition to word association, several researchers have investigated DSMs' ability to capture thematic relations. Asr, Zinkov, and Jones (2018) and Kacmajor and Kelleher (2020) used Jouravlev and McRae's (2016) thematic production norms as a test bed. These norms consist of data collected from participants who were given concrete noun cues such as *dog* and were asked to produce the names of other living or nonliving things that interact in situations involving the cue concept, such as *leash*. Responses that clearly did not constitute a thematic

relation (e.g., *dog-animal* is a subordinate-superordinate relation) were removed because participants were not following the instructions, although this occurred rarely. Kacmajor and Kelleher tested twelve models on their ability to capture these thematic relations (the proportion of participants who produced a valid response), as well as the associative relations found in Nelson et al.'s (1998) word association data, and the noun- and verb-pair human semantic similarity ratings from SimLex (see Section 2.1). None of the models performed extremely well (maximum correlation of 0.42 when evaluated against the similarity ratings from SimLex), and in general, the DSMs they tested were better at capturing similarity relations than thematic relations. Their knowledge-based models (i.e., those based on WordNet) also performed better on similarity than thematic relations, but the difference was starker – they performed better than DSMs on similarity and much worse on thematic relations.

Asr et al. (2018) used Skip-gram to simulate Jouravlev and McRae's (2016) thematic production data, SimLex999 similarity ratings, and the similar and related pairs from WordSim353 (a set of human relatedness ratings for which the pairs have been divided post hoc by similarity vs. other semantic relations; see Section 2.1). They also used Skip-gram to predict these ratings using combinations of word embeddings (the weights from the input layer to the hidden layer) and context embeddings (the hidden to output layer weights). They found that cosine similarity using the mean of word and context embeddings was the "winner" when they considered all of the datasets that they used. Wingfield and Connell (2022) also found that Skip-gram and CBOW performed reasonably well on these same three datasets, and that larger windows increased performance for the related pairs and for the thematic production data. For these reasons, we tested both Skip-gram and CBOW on a number of types of semantic relations using word, context, and mean embeddings.

1.3. The Current Study

From both a computational and psychological perspective, the studies outlined in Section 1.2 have broadened our understanding of semantic and thematic relations, and how they might be learned. For example, psychologically, they have provided insights into how much can be learned about these types of relations from language alone, as well as the limitations of this type of learning. These studies have made researchers think more deeply about the roles of language in learning concepts, the types and amounts of linguistic input that people experience over their lifetimes (and the consequences of such input), the plausibility of multiple types of learning mechanisms, and the ways in which learning from language and learning from other perceptual input and action might reinforce one another. The goal of the current research was to expand these lines of investigation (e.g., Asr et al., 2018 Kacmajor & Kelleher, 2020; Wingfield & Connell, 2022; and Lapesa and Evert, 2013 & 2014; see Section 2.6) in multiple ways, including the set of DSMs tested, the use of word, context, and mean embeddings for Skip-gram and CBOW, and the breadth and specificity of the semantic relationships examined.

Broadening the types and specificity of semantic relationships under investigation is of particular interest because a large number of semantic and thematic role relations play key roles in language processing. These include, but are not limited to, relations such as: the degree of similarity (*chair - sofa*), how things are used (function: *couch*-[used for] *sitting* or *broom*-[used for sweeping a] *floor*), where people or things typically are found (location: *airport-pilot*, *barn-hay*), the types of actions typically performed by people or animals (agent-action: *artist - sketch*), and the types of things or people on which an action typically is performed (action-patient: *serve-customer*, *strum-guitar*). People's knowledge of these relations is key to language usage and to understanding the world in general. For example, semantic similarity has played an

important role in theories of the organization of semantic memory going back to Collins and Quillian (1969) and progressing through to representational similarity analyses of neural data (Mur et al., 2013). Object and entity properties such as its shape, its color, how it is used, and where it typically is found have been central to theories of conceptual processing and word meaning in the mind and brain for a long time (Martin, 2007). Relations between verbs and agents, patients, instruments, and locations are key to theories of how humans understand sentences (Carlson & Tanenhaus, 1988; Rabs et al., 2022). Finally, relations among components of events play central roles in schema theories of human knowledge (Ghosh & Gilboa, 2014), and in theories of event cognition (Zacks, 2020).

In the present study, we examined whether eight DSMs are able to capture a broad range of human ratings covering multiple semantic relations. That is, our goal was to understand the degree to which DSMs encode a number of relations that are important for understanding words and sentences. Our primary aim was to test the degree to which it might be possible to use a DSM as a viable representational basis of a processing model of word and sentence comprehension. Although we tested a number of the currently best-performing type-based (what Kacmajor & Kelleher, 2020, call "vanilla") DSMs, and systematically manipulated some of their parameters, our main purpose was to understand whether it is possible to use the distance between word vectors obtained from these DSMs to approximate human knowledge of a wide array of semantic relations. By type-based DSMs, we mean non-contextual models that employ relatively standard methods such as using counts to obtain a single vector for each word and using cosine similarity between those vectors as a measure of semantic relatedness. Skip-gram and CBOW are included in this category because they create one representation for each vocabulary word.

1.4 Transformer Models

In our study, we exclude a newer generation of substantially more powerful and complex models, particularly transformer-based models (Vaswani et al, 2017) like BERT (Devlin et al., 2019). These models, by virtue of their complex architecture, are able to form rich, context-sensitive representations ("token-based") and perform well on many natural language processing tasks. Transformer models generate multiple vectors for each word in their vocabulary, and they can be tuned to simulate specific tasks. For example, BERT decomposes sentence-level input into token, segment, and position embeddings. BERT learns some syntactic information (Tenney et al., 2019; Liu et al., 2019) and semantic roles (Ettinger, 2019; Tenney et al., 2019); however, this knowledge has been revealed using either cloze tasks (Taylor, 1953) or by learned classification using BERT's representations.

We exclude discussion of these models for a few reasons. For one, despite the recent appearance of these newer models, many psycholinguistic practitioners who use representations from DSMs as tools in their research continue to use models such as PPMI, GloVe, Skip-gram, and CBOW a great deal. There continues to be a great deal of research in cognitive psychology, cognitive neuroscience, and psycholinguistics on the recognition of isolated 'wordforms' (phonological or orthographic) and accessing the meanings of single words when they are not presented in sentential or discourse contexts. Isolated spoken word recognition continues to be a major area of research, with development of large performance databases (Goh, Yap, & Chee, 2020), and work focused on isolated spoken words is critical in current research on language development and disorders (Apfelbaum et al., 2023; Giovannone & Theodore, 2021; McMurray et al., 2022), language and cognitive decline in aging (Nitsan, Banai, & Ben-David, 2022), and unlocking the organization of bilingual lexical knowledge and processing (Desroches et al.,

2022). While many questions remain to be answered about form recognition, there is also active research on semantic processing in isolated words (Nenadić et al., 2022). Similar questions are addressed in the domain of isolated visual word recognition, including basic processes of form recognition (Wang et al., 2021) and how word meaning influences those processes (Connell & Lynott, 2014; Pexman et al., 2017), how word meaning is represented in the human brain (Fernandino et al., 2022; Poeppel & Idsardi, 2022), and how words are related to one another, such as research on associative relations (De Deyne et al., 2019), and semantic priming (Hutchison et al., 2013). Type-based DSMs mesh well in terms of the information that might influence human processing in these cases. For example, out of the total citation count for Mikolov's original Skip-gram papers (Mikholov et al., 2013a; Mikolov et al., 2013b), 24% of these citations have come in the last year and a half. Moreover, the data we use here to compare DSMs involve limited to no contexts; typically, participants are presented with a pair of words and are asked to rate various types of relations. Given this kind of input, it is unclear if a transformer-type model would outperform the type-based DSMs we consider here given the lack of explicit context. However, we note that some recent work in constructing type-based representations from these models suggests that they could be used to investigate this issue (Chronis & Erk, 2020; Bommasani, et al., 2020; Ethayarajh, 2019; but see also Lenci, et al., 2022). In summary, we tested an important class of DSMs that figure prominently in the field. We understand that new models will continue to emerge in the future–transformers are one class but others are likely to emerge as well. In addition to investigating DSMs, our hope is that the present study provides a template for how to investigate the characteristics of other (including not yet developed) models.

2. The Testbed: Relations and Ratings

We used 19 ratings datasets (13 of which were archival; see the Appendix for a description of the other 6) to investigate a rich and varied set of semantic relations that involve words from multiple major syntactic classes and have meanings that span the abstract-concrete continuum. See Table 1 for information about the datasets. Table 1 shows, for each semantic relation, (1) the type of semantic relation, (2) the total number of word pairs in the dataset, (3) the number of pairs in which both words appear in the WikiOS corpus, (4) corpus coverage, which is (3) divided by (2) expressed as a percentage, and examples of (5) related and (6) unrelated pairs. We correlated cosines from PPMI, GloVe, Skip-gram, and CBOW with human ratings for pairs of words. The ratings datasets that we used are described below [numbered in square brackets]. Note that although we use word pairs that are related strongly to illustrate each relation, the datasets for all relations include pairs that span the continuum for the specific relation from strongly related to unrelated.

2.1. Semantic Similarity

SimLex-999 is a large database of semantic similarity ratings with item pairs that span the concrete-abstract continuum (Hill, Reichart, & Korhonen, 2015). Hill et al. used instructions designed to focus participants on similarity rather than on other types of semantic relatedness (e.g., participants were told that *cup-mug* and *frog-toad* are similar, whereas *car-tire* and *car-crash* are related but not similar). Their instructions were successful in that, for example, *quick* and *rapid* were rated as highly similar, whereas *meat* and *sandwich* were not (even though meat is often a part of a sandwich). Hill et al. removed pairs that included a word with lower than a 75% tendency for belonging to a specific part of speech to reduce part-of-speech and semantic ambiguity. They also explicitly instructed participants to give antonyms low ratings, so that, for

example, the mean rating for *tiny-huge* was extremely low. We therefore manually removed all antonym pairs from the verb-pair subset of SimLex-999 that we call SimLex-222-V in Table 1. (We consider the issue of DSMs and participant instructions of this type in the Discussion.) We tested model performance on all 999 word pairs [1; not shown separately in Table 1), the 666 noun pairs [2; SimLex-666-N in Table 1), 222 verb pairs [3; SimLex-222-V), and 111 adjective pairs [4; SimLex-111-A).

Table 1: Size, relation type, corpus coverage, and example item pairs for the semantic relations used in this study.

rating set	description	number of pairs	pairs in corpus	coverage (%)	highest-scoring	lowest-scoring
SimLex-666-N	noun similarity	666	666	100	creator, maker (9.62)	container, mouse (0.3)
WordSim-Sim	similarity	203	202	99.5	journey, voyage (9.29)	king, cabbage (0.23)
Function	similarity in usage	159	159	100	apple, peach (7)	teapot, wheel (1)
Shape	similarity in shape	222	128	57.7	headphones, earmuffs (5.5)	axe, baby carriage (0.1)
Manipulation	similarity in motions made when using objects	222	128	57.7	helmet, crown (5.5)	piano, badminton racket (0.1)
Color	objects of similar color	393	393	100	puck, spider (7)	grasshopper, lipstick (1)
SimVerb-3500	verb similarity	3281	3279	99.9	repair, fix (9.96)	create, dive (0)
SimLex-222-V	verb similarity	192	192	100	vanish, disappear (9.8)	ignore, explore (0.4)
SimLex-111-A	adjective similarity	111	111	100	quick, rapid (9.7)	new, ancient (0.23)
WordSim-Rel	general (undifferentiated) relatedness	252	250	99.2	environment, ecology (8.81)	king, cabbage (0.23)
Verb/Agent/Patient	agent/verb and verb/patient relatedness	720	720	100	marry, bride (7)	execute, nun (1.3)
Verb/Instrument	actions and the objects used to perform them	248	248	100	sweep, broom (7)	eat, plier (1)
Verb/Location	actions and places where those actions are performed	277	276	99.6	wait, lineup (7)	fish, puddle (1)
Event-Person	events and the kinds of people found at them	538	535	99.4	execution, executioner (6.9)	recess, parent (2.8)
Event-Thing	events and objects associated with them	592	591	99.8	blizzard, snow (7.0)	accident, train (3.0)
Instrument-Person	instruments and the people who use them	504	503	99.8	mirror, barber (6.9)	lantern, shopper (1.95)
Instrument-Thing	instruments and the types of things on which they are used	356	356	100	freezer, ice (6.7)	saucepan, coffee (1.85)
Location-Animate	locations and the kinds of living things found there	288	285	99.0	university, student (6.95)	desert, gangster (1.75)
Location-Thing	locations and the kinds of nonliving things found there	440	438	99.5	hotel, baggage (7.0)	alley, blood (3.5)

SimVerb-3500 [5] is a set of similarity ratings for a diverse set of 3,500 verb pairs. Gerz, Vulic, Hill, Reichart, and Korhonen (2016) selected the verbs from Nelson et al.'s (1998) association norms using Nelson et al.'s part-of-speech statistics. Gerz et al. signaled to their

participants that the items were verbs by using the infinitival forms, as in *to reply-to respond*. As in SimLex-999, they focused human raters on similarity by instructing them to distinguish between similarity and other types of relatedness, and to provide low ratings to antonyms. Thus, as with SimLex-222-V, we manually removed antonym pairs from SimVerb-3500 for all subsequent analyses.

Finkelstein et al. (2002) developed WordSim-353. Participants rated 353 word pairs on the degree to which they were semantically related in any manner. The pairs consisted primarily of nouns. Thus, similar pairs such as *journey-voyage* and pairs that shared other semantic relations such as *closet-clothes* (a closet is a place where clothes are kept) were given high ratings. Agirre et al. (2009) used their intuition to manually divide the 353 word pairs into two categories: 203 similar pairs such as *journey-voyage* and *street-avenue*, versus all other types of semantic relations that were present in Finkelstein et al.'s remaining items, such as *closet-clothes* and *treatment-recovery*. For investigating semantic similarity, we used the WordSim-353 set of 203 similar word pairs, which we refer to as WordSim-Sim [6].

2.2. Similarity Rated on Specific Dimensions

We used four sets of noun-pair similarity ratings, each focused on a different aspect of similarity. The first set focused on similarity in terms of general function (i.e., purpose of use) of the items, as in "How similar are the typical uses of *glue* and *tape*?" [7] (Function in Table 1). These pairs were designed to be similar in function, shape, both, or neither. Thus, many of the pairs that were rated as highly similar in function also had similar shapes, and as a consequence were also manipulated similarly[2], such as *bed-cot* (Yee, Drucker & Thompson-Schill, 2010).

[2] Although the fact that these functionally related pairs were also similar in shape and manipulation meant that function could not be isolated, trying to isolate function would have resulted in very few pairs (and pairs for which function similarity was not particularly high). Fortunately, the second and third sets (described below) of items were similar in shape or manipulation, but not function, allowing us to detect the contribution of function.

The second and third sets come from shape and manipulation similarity ratings from a common set of nouns collected by Musz, Yee, and Thompson-Schill (2012). The second set consists of the items rated on shape similarity. Participants were given word pairs such as *softball-grapefruit* and *key-screwdriver* and asked to "Picture the things that the words refer to and rate them according to how similar their shapes are" [8]. The third set focused on manipulation similarity. Musz et al. asked different participants to "Consider the typical movements that you make when you use these objects and rate how similar the movements are" [9]. Importantly, the pairs for the second and third sets were designed such that (a) when they were similar in shape, they were dissimilar in manipulation, (b) when they were similar in manipulation, they were dissimilar in shape, and (c) they were all dissimilar in function. These careful controls make these ratings useful for evaluating shape or manipulation sensitivity in the absence of similarity on the other two features.

The fourth set of ratings focused on color similarity. Participants were presented with word pairs such as *basketball-tiger* and *cherry-banana* and were asked to "Picture the objects that the words refer to and rate them according to how likely they are to be the same color" [10]. These pairs were assembled using the stimuli from several studies (Yee, Ahmed & Thompson-Schill, 2012; Yee, Huffstetler & Thompson-Schill 2011; Musz, Yee & Thompson-Schill, 2012). One particular strength of the items from these studies is that, due to the selection criteria, pairs that were rated as highly similar in terms of color were unlikely to be similar in terms of function, shape, or manipulation.

Note that the function, manipulation, shape, and color word pairs used in the present study, as well as the verb-agent/patient, verb-instrument, verb-location, and the six event-based noun-noun relations, included a much larger set of pairs than were used in the on-line eye-

tracking, priming, or sentence comprehension studies in the original articles. The items used in those on-line studies consisted of the most strongly related pairs, plus unrelated control pairs that typically were formed by re-pairing the strongly related words. Word pair selection for the on-line studies was influenced by other factors as well, such as making sure that no participant experienced any word in the experiment more than once. In all cases, the items for those on-line studies were chosen based on ratings of much larger sets of items. In the present study, we were able to use the larger sets of rated items to create more of a continuum of similarity or relatedness, and because factors such as making sure that no word was used more than once were irrelevant to the present simulations.

2.3. General Relatedness

We used one set of word pairs that included a number of types of semantic relations involving primarily nouns. These were the 252 pairs from WordSim-353 that were classified by Agirre et al. (2009) as being related in some way other than semantic similarity [11; WordSim-Rel]. This dataset includes pairs that are semantically related in various ways, such as *money-deposit* (depositing is something that you do with money), *tennis-racket* (you use a racket to play tennis, and the two words form a collocation as well), *closet-clothes* (you keep clothes in a closet), *lawyer-evidence* (lawyers present evidence), and *cup-liquid* (you use a cup to contain and drink liquids, and you pour liquids into cups). Note that the 252 related pairs plus the 203 similar pairs adds up to 455 pairs rather than 353. Agirre et al. included a number of pairs as unrelated items in both sets that were neither similar nor related in any other manner, such as *king-cabbage*.

2.4. Verb-Noun Thematic Role Relations

Relations between verbs and nouns that play the roles of agent, patient, instrument, and location have played a major role in theories of how people understand language for many years (Dresang, Dickey, & Warren, 2019), as well as theories of event knowledge and memory (Zacks, 2020). These items were collected for sentence comprehension experiments because they reflect people's conceptually-based thematic role knowledge (often termed "thematic fit"; Ferretti, McRae, & Hatherell, 2001; McRae, Spivey-Knowlton, & Tanenhaus, 1998). The first dataset includes verb-agent relations as in *arrest-cop* and *serve-waitress* and verb-patient relations as in *arrest-criminal* and *serve-customer* [12; Verb-Agent/Patient]. Note that all potential agents and patients were rated in both conditions, so that the data include agenthood and patienthood ratings for all verb-noun combinations. For agenthood, participants provided ratings for questions such as "How common is it for each of the following to chase someone/something?" For patienthood, participants rated, for example, "How common is it for each of the following to be arrested by someone?" For the present analyses, because the DSMs that we investigated are not able to distinguish agents from patients, we combined the agenthood and patienthood ratings by choosing the highest of the two ratings for each verb-noun pair.

We also used verb-instrument thematic fit ratings from Ferretti et al. (2001). Ferretti et al. asked participants to rate, for example, "How common is it for each of the following to be used for the action of stirring?", with the items including words such as *spoon*, *straw*, *fork*, *scissors*, and *stick* [13; Verb-Instrument]. The final verb-noun thematic role relation involved locations [14; Verb-Location]. Participants provided ratings for sets of verb-noun pairs regarding the likelihood of an action taking place in various locations, as in "How common is it

for someone to sleep in each of these locations?", with the items including *bedroom*, *chair*, *bathtub*, *tent*, and *car*.

2.5 Noun-Noun Event-based Relations

Hare, Jones, Thomson, Kelly, and McRae (2009) collected production norms for multiple noun-noun event-based relations. These included: event nouns and the types of people (*robbery-burglar*) and things (*surgery-scalpel*) that might be part of those events; instruments and the types of people who use them (*knife-chef*), and things on which they are used (*scissors-hair*); and locations and the types of people or animals (*barn-cow*) and things (*farm-tractor*) that tend to be found at those locations. Because Hare et al. had participants generate responses rather than provide ratings, to create methodologically comparable data sets, we collected ratings for the present project (see the Appendix for details). We tested the DSMs on each relation separately: event nouns and the types of people (*robbery-burglar*) [15; Event-Person], and things that take part in those events (*surgery-scalpel*) [16; Event-Thing]; instruments and the types of people (*knife-chef*) who use them [17; Instrument-Person], and things on which they are used (*scissors-hair*) [18; Instrument-Thing]; and locations and the types of people or animals (*airport-pilot*, *barn-cow*) [19; Location-Animate] and things (*farm-tractor*) [20; Location-Thing] that tend to be found at those locations.

2.6. Previous Research using Thematic role and Event-based Relations

Thematic role and event-based relations have been investigated by Lapesa and Evert (2013). They used count-based DSMs to simulate human semantic priming data from Ferretti et al. (2001; verbs priming agents, patients, instruments, and locations), McRae et al. (2005; agents, patients, instruments, and locations priming verbs), and Hare et al. (2009; event-based noun-noun priming). Lapesa and Evert compared tens of thousands of models by varying

parameters such as dimensionality reduction, corpus, window size, presence versus absence of part of speech tags, and the measure used to calculate distance between the resulting vectors (including neighbor rank). Many of the tested models distinguished related prime-target pairs (e.g., *arresting-cop*) from unrelated pairs (e.g., *dining-cop*) with quite high accuracy (e.g., cosine was higher for the related than for the unrelated pair). A joint corpus (a concatenation of the corpora that they used) provided the best performance, as did their longest window (15 words), and no dimension reduction. Lapesa and Evert also found that most of their models performed rather poorly when predicting human latency priming effects (in ms), although some models did predict a moderate proportion of variance.

Lapesa, Evert, and Schulte im Walde (2014) tested the ability of a large number of count-based DSMs to differentiate related versus unrelated pairs that were formed on the basis of multiple semantic relations. They used data from the Semantic Priming Project (Hutchison et al., 2013) and distinguished among synonyms (*frigid-cold*), antonyms (*hot-cold*), category coordinates (*table-chair*), forward phrasal associates (*help-wanted*), and backward phrasal associates (*wanted-help*). Lapesa et al. also amalgamated the prime-target pairs from Ferretti et al. (2001), Hare et al. (2009), and McRae et al. (2005) into a set that they called *generalized event knowledge* pairs. Their models distinguished between related and unrelated prime-target pairs with high accuracy across the types of relations. Lapesa et al. concluded that the size of the context window and dimensionality reduction are important for DSM performance across the similarity-based and other relations that they tested.

The investigations of Lapesa and Evert (2013) and Lapesa et al. (2014) are important because they show that count-based bag-of-words DSMs can account for both similarity and other semantic relations (for related work, see also Wingfield & Connell, 2022). The present

research differed from their studies in four primary ways. First, Lapesa and Evert used the prime-target pairs that appeared in the priming studies of Ferretti et al. (2001), Hare et al. (2009), and McRae et al. (2005), whereas we used ratings for the pools of items that served as the bases for selecting the smaller pools of items for the actual priming studies. Our approach provided us with much larger sets of items, allowing us to report results separately for each relation. Second, we simulated human ratings rather than testing whether the models can differentiate between related and unrelated pairs or can predict decision latency differences in priming studies. Third, we used a larger set of similarity-based relations and datasets. Fourth, GloVe, CBOW, and Skip-gram were not evaluated in their studies.

Finally, given the importance of thematic role information in language comprehension, a number of studies have used syntax-based distributional models constructed from grammatically parsed, part-of-speech tagged, or semantic role labeled corpora to simulate judgments and influences of verb-noun thematic relations (i.e., verb-noun thematic fit or selectional preferences; Baroni & Lenci, 2010; Erk et al., 2010; Sayeed et al., 2016; Santus et al., 2017; Tilk et al., 2016). Generally, these models construct a prototype vector for a thematic role of a verb (e.g., the patient role of *cut*) by averaging the dependency-based vectors of its most typical role fillers (i.e., the words that appear as the patient of *cut* in a corpus). The similarity of a noun with the thematic role prototype is used as the estimate of its plausibility as a filler for that role (e.g., things that can be cut). This approach has been successful in capturing ratings of verb-noun thematic relations, particularly with respect to verb-agent and verb-patient argument relations. Our corpora were not parsed, tagged, or labeled, allowing us to assess whether metrics based on the distributional statistics alone can reflect thematic relations.

3. The Corpora

We created two English corpora, Wiki2018 and EngOS, and two versions of each corpus: one at a document level (used to train GloVe and PPMI) and one at a sentence level (for Word2Vec). For Wiki2018, we extracted a document-level November 2018 dump of English Wikipedia into JSON format, and the sentence-level corpus was constructed by splitting each document into sentences. For consistency, we refer to a Wikipedia article/page as a "document." In Wikipedia, because the final sentence of every document is generally index terms/redirects (an index term captures the essence of the topic of a document, whereas a redirect is a link that automatically sends visitors to another page), we removed the final sentence from each document. For EngOS, we downloaded a recent version of the English Open Subtitles database from OPUS (https://opus.nlpl.eu). The OPUS version has improved sentence alignment and better language checking compared to what exists on opensubtitles.org and is broken into sentences. To make a document-level ("chunked") version of EngOS, we grouped sets of 356 tokens together (the mean document length in the Wiki2018 document corpus).

All corpora had punctuation removed, with the exception of hyphens, which were replaced by a space (e.g., 'self-governed' became 'self governed' rather than 'selfgoverned'). We converted all words to lowercase and removed all stop words and words that occurred fewer than 100 times. Sentences that consisted of fewer than three words were dropped. For the Wiki2018 document corpus, any documents with fewer than three tokens were removed; these typically are stubs (a document deemed too short to provide encyclopedic coverage of a subject) and redirects. All corpora were lemmatized using spaCy (https://spacy.io).[3]

[3] In our two sentence-level corpora, we found a roughly power-law tail leading to very infrequent, but unreasonably long, sentences. There were 53K (0.06% of 90M) sentences of length greater than 100 tokens in Wiki2018, and 307 (0.0002% of 176M) in EngOS. In EngOS, the vast majority of these are (i) strings of hundreds to thousands of integers presumably having to do with video display, (ii) unicode gibberish, or (iii) song lyrics rendered with no

Finally, the documents or sentences in all four corpora (Wiki2018/EngOS X document/sentence) were shuffled randomly. We then created our final corpora, which we call WikiOS, by randomly interleaving Wiki2018-sent and EngOS-sent and, separately, Wiki2018-document and EngOS-document. The WikiOS corpus consists of roughly 1.9 billion tokens; other studies have used similarly blended corpora (e.g., Baroni & Lenci, 2010).

4. The DSMs

We tested the extent to which the vector representations from each of the eight selected DSMs are sensitive to the 19 sets of relation ratings. The DSMs were PPMI, GloVe, Skip-gram and CBOW word embeddings for both words in each pair, Skip-gram and CBOW context embeddings for both words, and Skip-gram and CBOW using the mean of the word and context embeddings for both words. We also investigated a number of parameter settings for each measure. Each model took roughly 4 hours to train on a multi-core Mac workstation.

Researchers have suggested that Skip-gram and CBOW use parameterizations that are optimized for discovering similarity relations (Asr et al., 2018; Levy et al., 2015). Furthermore, researchers have almost always used word embeddings (input to hidden weights) and discarded the context embeddings (hidden to output weights). However, Levy et al. and Asr et al. have argued (and tested to some extent) that although word embeddings may best capture similarity, it is possible that context embeddings (hidden to output weights) may best capture other semantic relations. Therefore, we used word embeddings, context embeddings, and the mean of the two.

sentence punctuation. In Wiki2018, most of the extremely long sentences are either improperly rendered tables or long quotes containing no periods. For example, among the longest Wiki2018 sentences are a run-on quote from Meher Baba, tables of Gaelic Football results, awards in an Australian national music competition won by Lyneham High School in Canberra, and a list of matches played by the English rugby player John Holmes. We therefore removed all sentences from both corpora longer than 42 tokens (Adams, 1979).

Based on prior results (Lapesa and Evert, 2013; Lapesa et al. 2014; Troyer and Kutas, 2020), we hypothesized that longer windows might be advantageous for capturing thematic and event-based relations, whereas capturing similarity relations might generally be insensitive to window size. Our intuition was that event-based relations often are expressed in language via more distal lexical co-occurrences, at least sufficiently frequently that model performance might be influenced by window size. Therefore, we included short (2 words on each side), medium (7 words) and long (13) windows for each model. Rather than varying all possible hyperparameters, we focused on those that seemed most likely to matter (like window size), or have been shown to impact performance in previous studies. We tested a total of 96 models.

4.1 CBOW/Skip-gram hyperparameters

All CBOW and Skip-gram models used 300-dimensional embeddings (i.e., the networks had hidden layers of size 300). In addition to testing short (S), medium (M), and long (L) windows, respectively, we varied (1) the threshold for random downsampling of high-frequency words (Y: threshold = 0.001; N: do not downsample) and (2) using negative sampling (Y) or simply using hierarchical softmax (N). With negative sampling, 10 randomly sampled noise words were drawn. In all cases, we trained for five epochs with an initial learning rate of 0.025. The shape parameter for the negative sampling distribution was 0.75. These combinations produced 12 models for each of the three types of embedding (i.e., 3 window sizes, 2 downsampling possibilities, and 2 negative sampling possibilities). Using the Python gensim library implementation of Word2Vec (Rehurek & Sojka, 2011), we thus generated 36 CBOW and 36 Skip-gram models.

4.2 PPMI Hyperparameters

PPMI models used the same window sizes (S = 2, M = 7, L = 13) as in our CBOW and Skip-gram models. We varied two other hyperparameters: (1) smoothing the context distribution when computing mutual information (Y: smooth with an exponent of 0.75; N: no smoothing) and (2) shifting PPMI values. Shifting is used to "zero out" small mutual information values, and is controlled by a shift parameter k. The final mutual information values are $\widehat{P_{ij}} = \max(P_{ij}, \log k)$. We chose either k = 5 (Y) or do not shift (N: k = 1). All words in the window were uniformly weighted when computing co-occurrences. We initially explored other weighting schemes but this manipulation made no discernible difference to performance in an initial set of analyses. We used our own Python package for PPMI to generate the 12 PPMI models – this package is available at https://github.com/thelahunginjeet/pyppmi.

4.3 GloVe Hyperparameters

GloVe models used 300-dimensional embeddings and the same window sizes. We varied two parameters related to GloVe's factorization of the co-occurrence matrix: (1) the weighting function parameter alpha (S = 0.75, L = 1.0) and (2) the weighting function cutoff x_{max} (S = 10, L = 100; see Equation 9 and Figure 1 in Pennington, Socher, & Manning, 2014). The hyperparameters that we did not vary were the number of passes in the decomposition (25) and the initial learning rate (0.05). Using C code available from the original authors (https://nlp.stanford.edu/projects/glove/), we generated 12 GloVe models.

Table 2 is a guide to our notation for models and hyperparameters. For example, a CBOW model using mean embeddings, with long windows, no downsampling of frequent words, and training with negative sampling is denoted as CBOW(m)-LNY. A GloVe model with short windows, a weighting function alpha of 1.0, and a weighting function cutoff of 10 is

denoted as GLoVe-SLS. A link containing sample python scripts and data relevant to this project is provided at https://osf.io/86dxs/.

Table 2: Guide to hyperparameters and shorthand notation for DSMs.

model	abbreviation	embedding	first parameter	second parameter	third parameter	example shorthand
word2vec (Skip-gram)	SG(x)	word, context, or mean (w,c,m)	window size (S=5,M=15, L=27)	down-sampling threshold (Y=0.001, N=none)	negative sampling (y) or hierarchical softmax (n)	SG(w)-LYY
word2vec (Continuous Bag Of Words)	CBOW(x)	word, context, or mean (w,c,m)	window size (S=5,M=15, L=27)	down-sampling threshold (Y=0.001, N=none)	negative sampling (y) or hierarchical softmax (n)	CBOW(m)-SNN
Global vectors for word representation	GloVe		window size (S=5,M=15, L=27)	weighting function parameter (S=0.75, L=1)	weighting function cutoff (S=10, L=100)	GloVe-SSS
Positive Pointwise Mutual Information	PPMI		window size (S=5,M=15, L=27)	context smoothing (Y=0.75, N=none)	MI shifting (Y=5, N=none)	PPMI-MNY

5. Results & Discussion

5.1 Corpus Coverage and Context Embeddings

Table 1 indicates that we have excellent corpus coverage, usually exceeding 99%. The only notable exceptions are the shape and manipulation relations for which lower coverage is due to multi-word phrases that were used in the human rating studies (e.g., baby carriage and badminton racket). Note that there are fewer than 222 and 3,500 pairs in SimLex-222-V and SimVerb-3500, respectively, because we manually removed all antonym pairs, as discussed

above. We retained the names SimLex-222-V and SimVerb-3500 to remain consistent with the original rating studies.

We begin with one result that is independent of relation type. All of the six Skip-gram and six CBOW models using only context embeddings that were trained without negative sampling failed to account for any of the ratings (all correlations were close to zero). This was not true of word embeddings or mean embeddings trained without negative sampling but was specific to context embeddings. Other work has demonstrated that the success of negative sampling in word2vec models relies on the underlying distributional structure of the language/corpus, and not the prediction mechanism in the model (Johns, et al. 2019). However, there is no difference in the corpus we used for the three embeddings so this does not seem to be the explanation here. In any case, we refrain from further discussion of these 12 models, and focus only on the remaining 84. In all cases, we scored the models as follows. For a given pair of words rated for a particular semantic relation, we computed a cosine between the two word vectors of each model. We did this for all pairs for that relation, and then we calculated the Spearman correlation between model cosines and human ratings.

5.2 Similarity

5.2.1. SimLex. Figure 1 shows the results for SimLex-999, divided into nouns (SimLex-666-N), verbs (SimLex-222-V), and adjectives (SimLex-111-A). Because we show many comparisons of this type, we describe Figure 1 in detail. For each set of ratings, we show a color-coded barplot. CBOW models are on an orange spectrum, with word/context/mean embeddings having successively darker colors. SG models are similar, but on a purple spectrum. GloVe models are in yellow, and PPMI in black. In each bar plot: (1) the height of the bar corresponds to the median Spearman correlation over all sets of hyperparameters for that type of

model; (2) the plus sign on top of the bar signifies the single best hyperparameter combination within a model class, and (3) dotted lines are drawn at the minimum correlation that would be significant at $p = .05$ with df = number of item pairs minus two.

Figure 1 shows that model performance on nouns is quite good, consistent with previous work and with our WordSim-Sim results in Section 5.1.2. Substitutional similarity of nouns seems to be encoded across models (i.e., is well-captured by all of the DSMs), and there are no major differences among the types of models. Interestingly, the models do almost as well on adjectives, with CBOW and Skip-gram performing somewhat better than GloVe and PPMI. Capturing verb similarity is substantially worse, with correlations dropping by factors of two or more, with the exception of CBOW using context embeddings (medium orange bar; see 5.1.4 for further discussion of verb similarity). Although CBOW(c) is particularly good at capturing verb similarity, and comparable with the best performing models for nouns and adjectives, it will become apparent that it performs poorly on many other relations. On the contrary, the median correlations for GloVe and PPMI are below the significance level for verb similarity, but as we will see, GloVe's performance on thematic and event-based relations is strong.

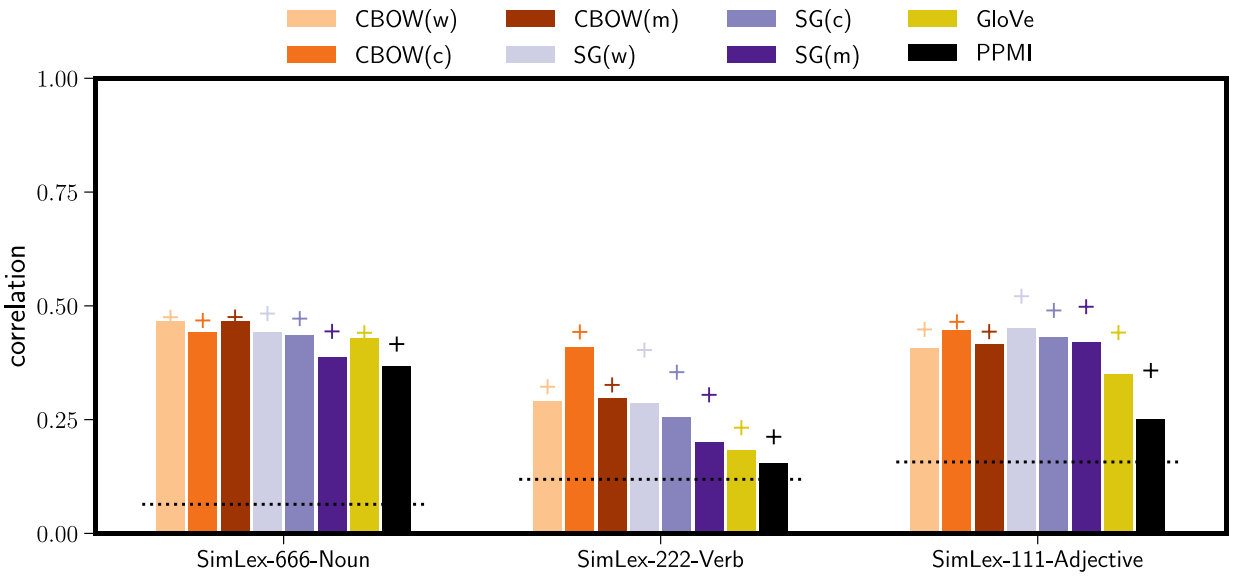


Figure 1. Summary of Spearman correlations between model-derived similarities and human ratings for SimLex (666 noun pairs, 192 verb pairs, 111 adjective pairs). For each of the eight model types (model + embedding), the bar height is the median correlation across the hyperparameter sets and the plus symbol denotes the maximum correlation obtained across the set. Dotted lines illustrate the size of the minimum $p = .05$ significant correlation given the number of pairs in the rating set (see Table 1 for number of pairs). For CBOW(c) and SG(c), models that did not use negative sampling performed consistently poorly and they are excluded.

5.2.2. Nouns and WordSim-Sim. Figure 2 shows noun similarity correlations for SimLex-666-N (repeated from Figure 1 to facilitate comparison) and WordSim-Sim. All models other than CBOW(c) capture WordSim-Sim ratings extremely well. In this study and others (Levy, Goldberg, & Dagan, 2015; Baroni et al., 2014), these relations are well-captured by a variety of DSMs, and show generally low hyperparameter sensitivity, as evidenced by the small difference between the median correlation over all sets of hyperparameters and the single best hyperparameter combination [the + sign].

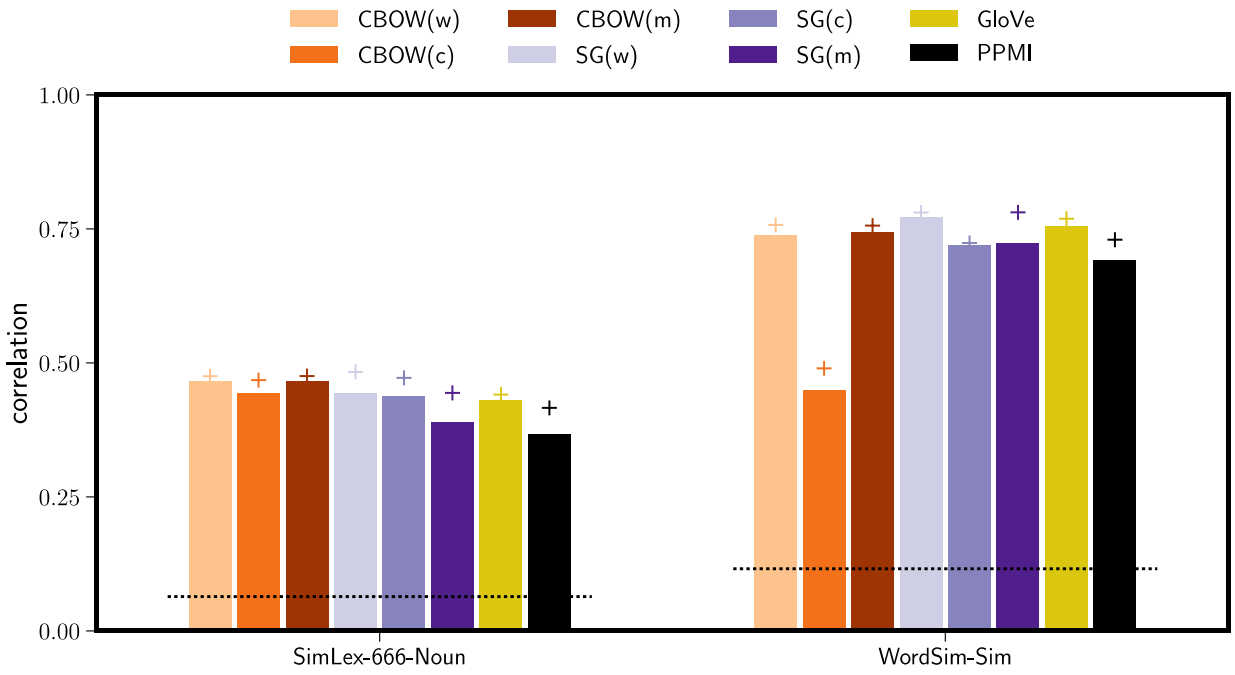


Figure 2. Model correlations with noun similarity relations (SimLex-666-N and WordSim-Sim); all annotations and symbols are as in Figure 1.

5.2.3 Similarity on Specific Dimensions. Figure 3 shows performance on similarity ratings that focus on four dimensions: function, shape, manipulation, and color. All DSMs capture concrete objects' similarity of function, which likely can be primarily attributed to the overlap between function and overall similarity in concrete objects in the real world, and the fact that things with similar functions are somewhat substitutional in language. Table 1 shows that one of the highest rated pairs in this set is *apple-peach*, both of which people eat (function). In addition, apples and peaches also share other types of features (e.g., round, grown on trees, sold in supermarkets), and this similarity presumably is reflected in shared linguistic contexts. In contrast, you have to go down to a pair tied for 57th in rating to find two words with similar function that are dissimilar on other dimensions (*birdcage-doghouse*), and which seem unlikely to participate in shared linguistic contexts.

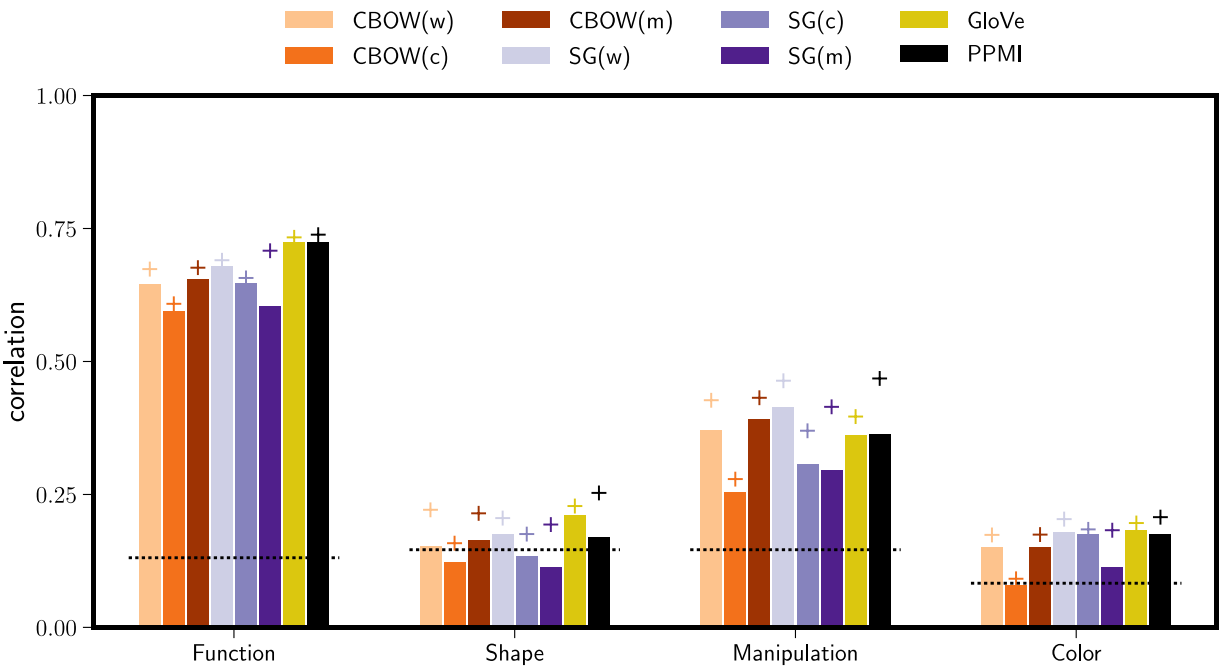


Figure 3. Model correlations with noun-noun similarity rated on specific dimensions (Function, Shape, Color, and Manipulation); all annotations and symbols are as in Figure 1.

Manipulation is analogous to some degree in that the linguistic contexts of the highest scoring pairs are likely to share some words (e.g., "wear" for *helmet* and *crown*, and "keys" for *piano* and *typewriter*). More surprisingly, most models pick up on shared color (significantly, albeit weakly), in which little else appears to discriminate high and low scoring pairs (e.g., *puck-spider*, *peas-frog*). This sensitivity to color similarity may be due to color words appearing in linguistic contexts (e.g., *black* appearing with both *puck* and *spider*). Unexpectedly, many of the models also produce significant correlations with shape ratings (although these correlations are weaker than those with color), with the best PPMI and GloVe models approaching a Spearman correlation of .25 or .20. Note that all of the function, shape, manipulation, and color relation pairs consist entirely of object nouns. Recently, "Multimodal DSMs" have been created by combining the standard textual input with images (Lazaridou, Pham, & Baroni, 2015). It would be interesting to test the degree to which multimodal models better capture color and shape relations.

5.2.4. Verb Similarity. None of the DSMs capture verb similarity particularly well (Figure 4, with SimLex-222-V repeated from Figure 1 to facilitate comparison with SimVerb-3500) as compared to overall noun, function, and manipulation similarity. The models perform somewhat better on SimVerb-3500. This occurs even though 170 pairs are in both SimLex-222-V and SimVerb-3500, with the Spearman correlation for human ratings of those pairs equal to 0.91. An interesting exception is CBOW with context embeddings. For SimLex-222-V, the median CBOW(c) model is better than the best model of any other type (even though CBOW(c) was anomalously poor at WordSim-Sim, though not SimLex-666-N).

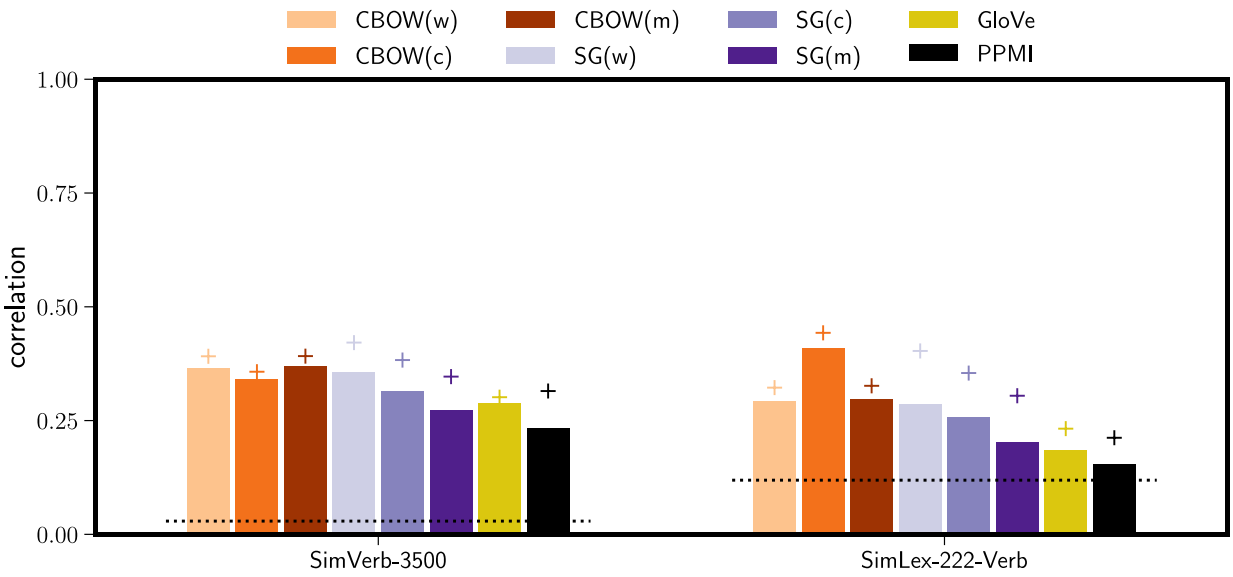


Figure 4. Model correlations with verb similarity relations (SimLex-222-V and SimVerb-3500); all annotations and symbols are as in Figure 1.

The DSMs may perform poorly on verb similarity ratings because verbs tend to have a greater number of meanings and senses. Because verb meaning is more malleable, they occur in more disparate linguistic contexts than do nouns, in particular concrete nouns (Gentner & Boroditsky, 2001; Kersten & Earles, 2004). Note that all of the DSMs used herein compute a single vector for a word, regardless of multiple senses and/or meanings. Furthermore, for verbs there may be additional variability in the human ratings because individuals may differ with respect to the verb meaning or sense that they consider when rating similarity.

We investigated a representative model (SG(m)-SYY, the best overall performing DSM; see Figure 9) by considering items that participants rated as highly similar but the DSMs rated as low. A number of these pairs include a verb that is often considered as a light verb (e.g., *do*, *get*, *give*, *have*, *make*, *take*). Light verbs are used frequently in constructions in which they seem to have little meaning, such as "I'll get better", "She made a comment", or "She gave a talk". DSM representations will be influenced by the frequent light verb usages, whereas when humans rate, for example, the similarity of *make* and *build*, *make* in the context of *build* presumably allows

people to ignore the frequent light verb occurrences of *make*. We investigated this using SG(m)-SYY by removing all verb pairs that included *do*, *get*, *give*, *have*, *make*, and *take* from SimLex-222-V (23 pairs) and SimVerb-3500 (130 pairs). Removing these items modestly increased the Spearman correlation from .30 to .39 for SimLex-222-V, and from .35 to .37 for SimVerb-3500. Thus, this type of ambiguous verb is challenging to some degree for DSMs.

5.2.5. Adjectives. Similarity between pairs of adjectives is captured almost as well as noun similarity by the majority of the DSMs (see Figure 1). GloVe is slightly below the CBOW and Skip-gram models, and PPMI performs the most poorly.

5.3 General Relatedness

Figure 5 shows model correlations with WordSim-Rel (general relatedness word pairs and instructions). Two observations are notable. First, CBOW(c) performs poorly relative to the other models. Second, the correlations with human ratings are quite strong and comparable for all other DSMs, although the models do not capture general relatedness as well as they capture WordSim-Sim similarity (see Figure 2).

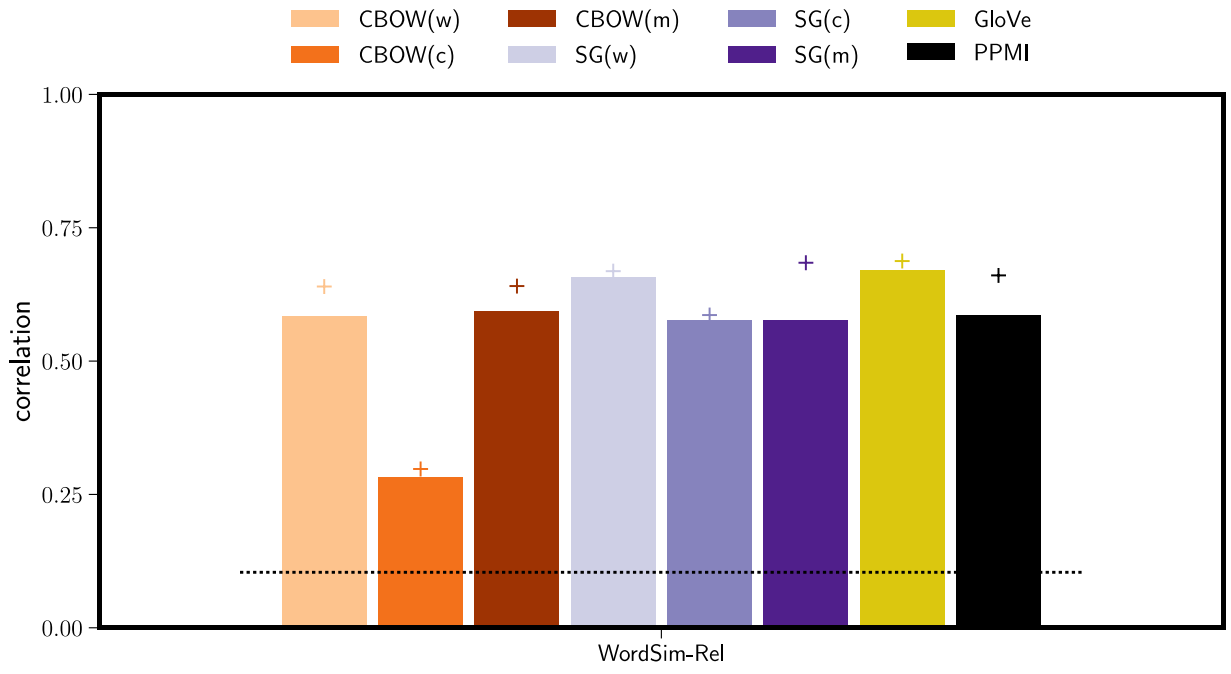


Figure 5. Model correlations with general relatedness (WordSim-Rel); all annotations and symbols are as in Figure 1.

5.4 Verb-Noun Relations

Figure 6 shows model performance for verb-agent/patient, verb-instrument, and verb-location relations. Interestingly, GloVe and PPMI in particular capture these thematic relations much more strongly than they do verb similarity (see Figure 4). This may occur because a verb is paired with a concrete noun in each item (although locations tend to be rated as somewhat less concrete than are types of things and people; Brysbaert, Warriner, & Kuperman, 2014).

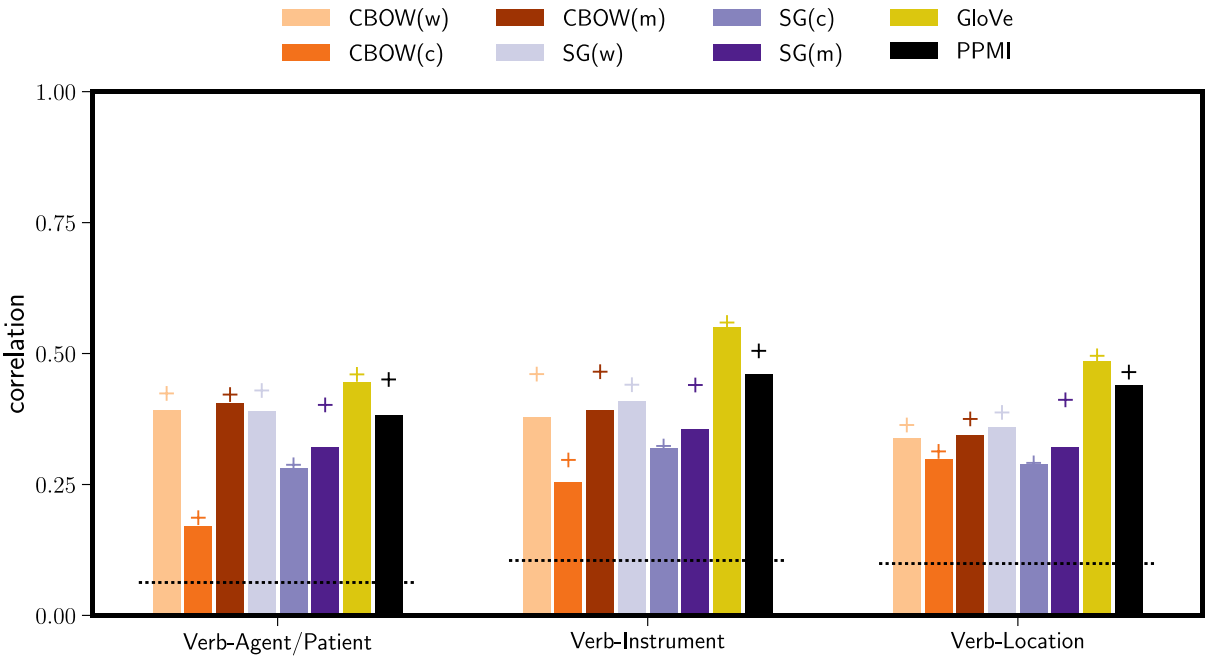


Figure 6. Model correlations with verb-noun thematic role relations (Verb-Agent/Patient, Verb-Instrument, Verb-Location); all annotations and symbols are as in Figure 1.

Although the verbs themselves may occur in quite variable contexts, the agent, patient, instrument, and location nouns tend to be less ambiguous than verbs, which might make capturing the verb-noun ratings easier. Furthermore, the related verb-noun combinations tend to occur together in language (and the world), whereas verb-verb similarity is more strongly tied to substitution in context. Also note that the human ratings differ substantially in that participants

rated the similarity between pairs of verbs for SimVerb-3500 and SimLex-222-V, whereas the verb-noun relation ratings involved judging, for example, how common it is for an action to be performed by someone, or how common it is for an action to occur at a certain location. Overall, the DSMs account reasonably well for these thematic role relations. In fact, GloVe and PPMI, and to a slightly lesser extent, CBOW with mean or word embeddings, and Skip-gram with word embeddings, perform as well on the verb-noun relation ratings as they do on SimLex-666-N (noun pair similarity).

A final potentially important factor underlying the DSMs' better performance on verb-noun than verb-verb relations follows from the fact that the verb-noun pairs were chosen as candidates for word-word priming studies. That is, because the experimenters (who include the final author on the present article) were testing whether it is possible for these verbs and nouns to prime one another outside of any other context, they selected verb-noun pairs with verb sense (as well as noun sense) in mind. Therefore, they chose verb-noun pairs in which the dominant sense of the verb matched the most strongly related nouns that followed. In contrast, the SimLex-222-V and SimVerb-3500 pairs were not selected with these constraints in mind.

5.5 Event-Based Relations

Figure 7 shows model performance on the six noun-noun event-based relations. One surprising feature of these results is that the models perform quite well on relations that involve people and animals, but they perform poorly on relations that include things (inanimate objects). Looking more carefully into model performance at the pair level provides a couple of potential explanations for the discrepancy between people/animals and things.

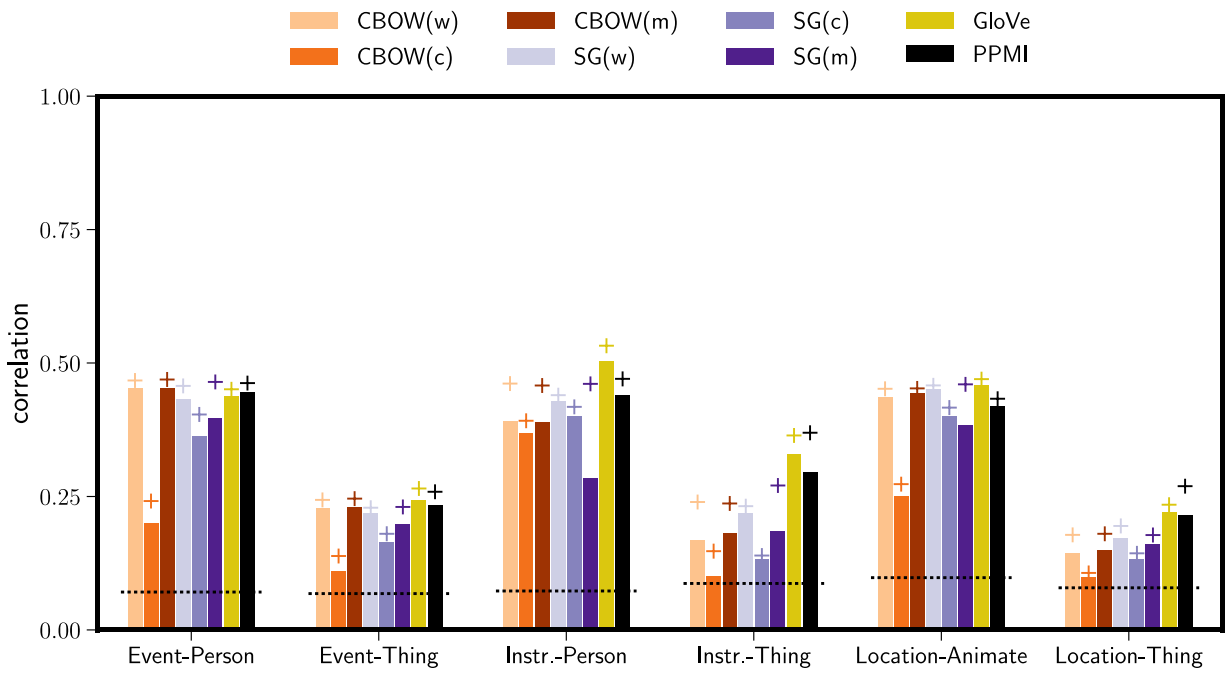


Figure 7. Model correlations with noun-noun event-based relations (Event-Person, Event-Thing, Instrument-Person, Instrument-Thing, Location-Animate, Location-Thing). All annotations and symbols are as in Figure 1.

Figure 8 shows scatterplots of the model pair similarities against the human ratings, for a representative DSM, SG(m)-SYY. There are a few differences among the relations. In terms of the human ratings themselves, Figure 8 shows that they are reasonably evenly distributed across the one to seven scale for all of the relations except event-thing and location-thing. In both of these cases, the ratings are skewed toward the high end, so that event-thing and location-thing ratings will be difficult for any DSM to predict. The red boxes in the event-thing and location-thing scatterplots are meant to draw the reader's attention to the fact that in these data, pairs of concepts that are deemed highly related by humans have a huge spread in model similarities, from completely unrelated to highly related. When we look at the event-thing and location-thing model ratings for pairs that have human ratings of 6.75 or above, we find what we call a *birthday-food* problem.

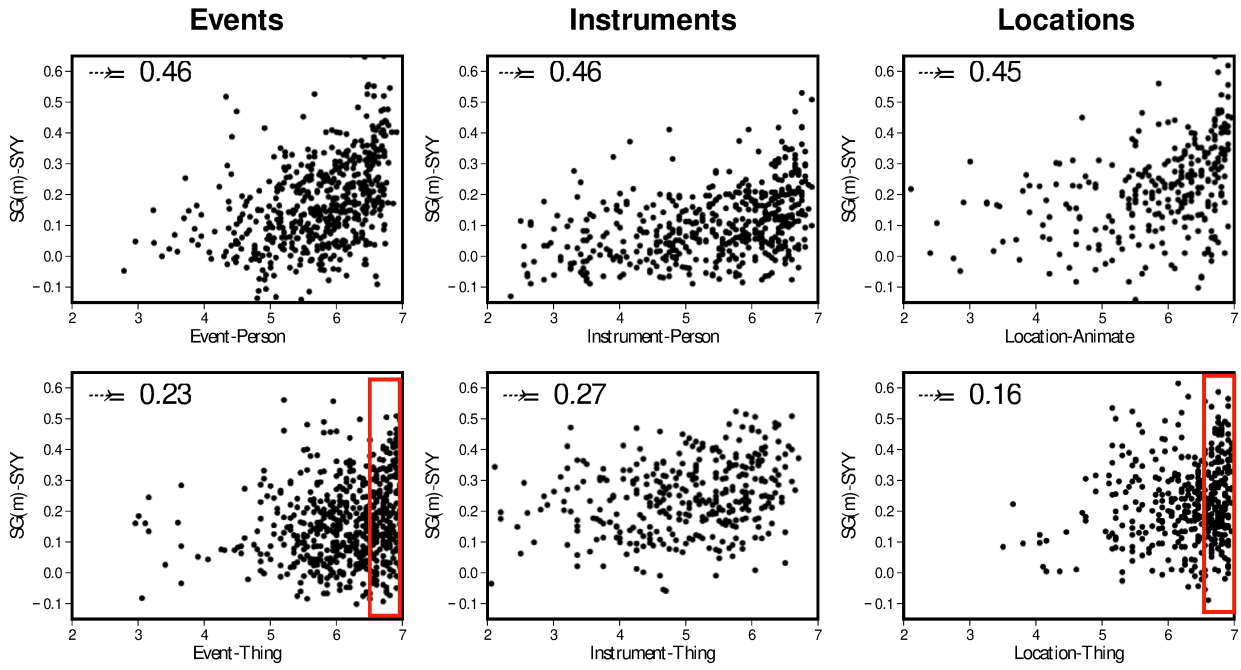


Figure 8. Scatter plot of word pair cosine similarities from a representative DSM (Skip-gram using mean embeddings, short windows, random downsampling of frequent words, and negative sampling) against human ratings, on an item-by-item basis, for the six sets of noun-noun thematic relations. Red boxes highlight items which are given very high ratings (6.5 or higher out of 7) by human participants.

Because food is not associated specifically with a single or a small set of real world events or locations (people grow, buy, prepare, eat, and talk about food in many places), words like *food* appear in a large number of diverse linguistic contexts, and DSM scores can be lower for pairs including these words. Furthermore, for *birthday-food*, the real world event-based conditional probability is relatively high in one direction, P(food|birthday), but low in the other P(birthday|food). That is, given that a person is at a birthday party, the probability of food being present is high, whereas given that there is food present, the probability of being at a birthday party is very low. DSM scores, in a sense, depend on both conditional probabilities as reflected in linguistic input. In contrast, human ratings can be influenced in ways that the models are not. When participants were asked to rate "How likely is each type of thing to be found in this situation", *birthday-food* had a mean rating of 6.90 out of 7 because there pretty much always is

food at a birthday party. For a rating task such as this, participants apparently ignore the low conditional probability of a birthday party given the presence of food. That is, the fact that people eat food in many other instances other than birthday parties did not strongly influence people's ratings. In other cases, the thing that was used in the rating task refers to an object that occurs in only one or a few locations, or at one or a few types of events, such as *gravestone, ballot,* or *runway*. Because those words were paired in the rating task with the strongly related event or location, the model cosines and human ratings tended to be much more concordant.

In the event-people and location-animate relations, there is less variability in model scores for the most highly human-rated items. This occurs generally because the words used to refer to types of people in these rating tasks tended to be more situationally specific, as in *waiter, teacher, cashier, lawyer,* and *bride*, as opposed to say, *child, grandparent, person,* and *woman*. For the event-people pairs that have mean human ratings greater than or equal to 6.75, SG(m)-SYY appropriately produced high scores to *wedding-bride*, *olympic-athlete*, and *trial-defendant* (situationally specific people), although it did produce lower scores to *cruise-captain, baptism-baby*, and *sale-shopper* (more situationally diverse people). An important point, however, is that the people and animate items were dominated by the former type, thus resulting in higher correlations for those relations.

These issues are not specific to SG(m)-SYY. As one can see in Figure 7, all classes of models perform worse on event-based relations involving inanimates (e.g., event-thing) than on the corresponding set involving animates (e.g., event-person). We therefore investigated model behavior on items with high participant ratings (> 6.75) and low scores in three additional models: CBOW(m)-LNY, GloVe-SSL, and PPMI-SYN. These were chosen because they are the best performing models of their class overall (as discussed in section 5.5). All of the models

tended to generate relatively low similarity scores for pairs in which conditional probabilities are highly asymmetric. In addition to *birthday-food*, other pairs characterized by this pattern include *forest-stick*, *restaurant-chair*, *tavern-glass*, *anniversary-food*, *party-music*, and *store-sign.*

The item-level analyses on the event relations and verb similarity data yield an important point (but perhaps an obvious one) that is often overlooked in investigations of DSMs (or simulations of human data in general, regardless of the type of modeling or computational analyses involved). Model performance can be strongly influenced by the manner in which items are chosen for a study, and by the instructions that are given to participants. These aspects of human studies can have a larger effect and be a more critical factor in understanding model quality (or lack thereof), than the specific model or set of hyperparameters chosen for a given model. Another very different recent approach using "socially-based" models (Johns, 2021) demonstrated that making encodings that are aware of how words are communicated across people and discourses might also provide representations that lean toward encoding non-similarity relations.

5.6 Overall Performance

We used a rank aggregation method to investigate and summarize overall model performance. (Rank aggregation code is available at https://github.com/thelahunginjeet/pyrankagg.) We treated the 19 ratings instruments as raters and the 84 models as the objects to be ranked. For a given rating type (e.g., WordSim-Sim), the model with the highest correlation was given rank 1, and the one with lowest correlation was ranked 84. We used a Robust Rank Aggregation method (Kolde et al., 2012) to produce scores. Although these scores can be used to produce voted/aggregate model ranks, we considered the scores directly because they allow us to see not only which models are best, but how close they are to each other.

Figures 9-11 show model scores from the ranking algorithm considering all 19 relations (we excluded SimLex-999 while including each part-of-speech based SimLex component; Figure 9), similarity relations only (Figure 10, 9 ratings; SimLex-666-N, SimLex-222-V, SimLex-111-A, WordSim-Sim, Function, Shape, Manipulation, Color, SimVerb-3500), and (3) relatedness, thematic, and event-based relations only (Figure 11, 10 ratings; WordSim-REL, Verb-Agent/Patient, Verb-Instrument, Verb-Location, and the six noun event-based relations).

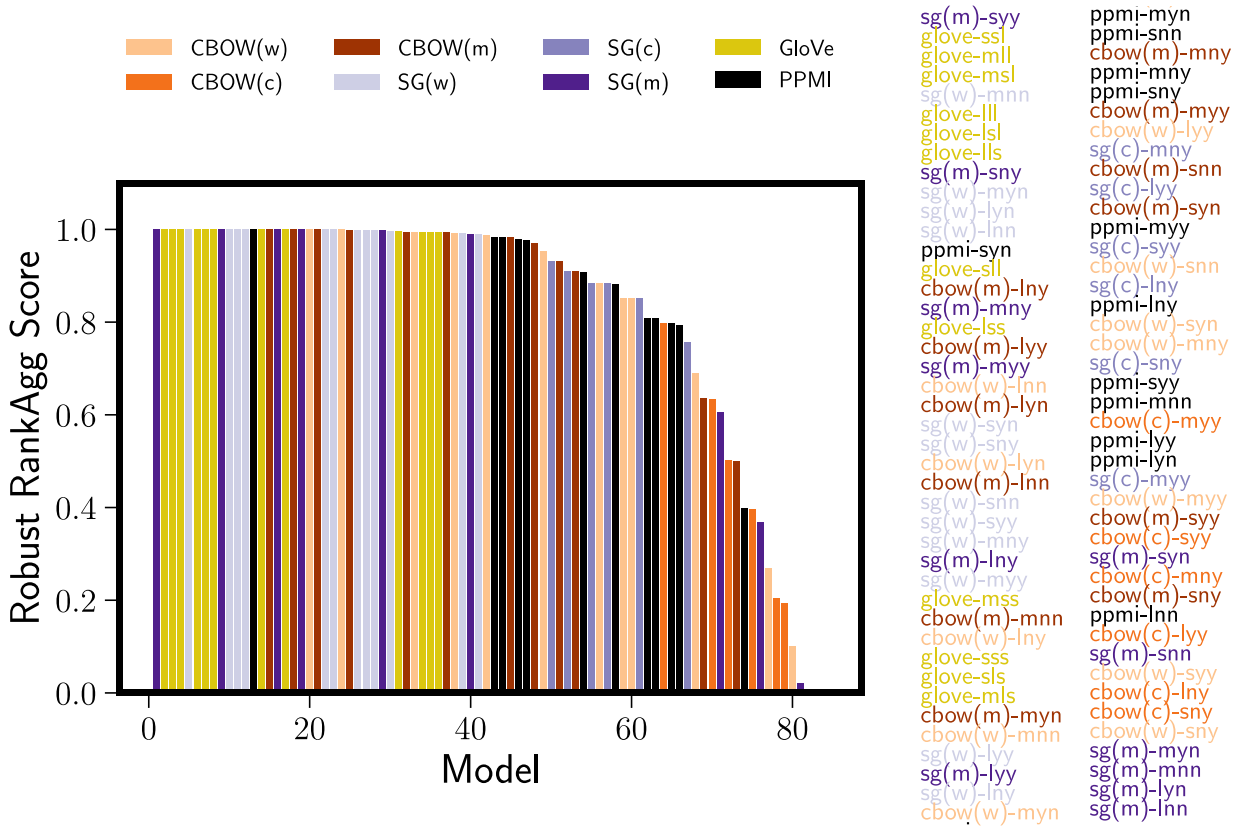


Figure 9. Ranking of 84 models using their performance over the entire set of relations considered in this study. The bar height is the sorted robust rank aggregation score, which is used to determine rank ordering. Skip-gram models with both word and mean embeddings and GloVe cluster near the top. Note that we have used all-lowercase shorthand for the model abbreviations, for better legibility.

One observation is that the rankings are crowded at the top. Although it is possible to crown a "best" model in each class, in many cases that best model is very close to multiple runners-up. In Figure 9, one must extend to at least the 40th best model to see a noticeable decrease. This

flatness is encouraging because it means, for example, that although GloVe models with a large weighting function cutoff ("GloVe-xxL" models) are overall the best, there are Skip-gram, CBOW, and a small number of PPMI models that, given appropriate hyperparameter choices, perform nearly as well. We use the word "encouraging" because if, for example, a cognitive psychologist, cognitive neuroscientist, or psycholinguist previously had used one of these models to control for a semantic variable in their stimuli in an experiment, there is no need for them to second-guess their model choice because a number of models with slightly different hyperparameters would also have worked reasonably well.

A second observation is that there is a striking difference in the kinds of models that are best at capturing similarity (Figure 10) versus those that are best at capturing other types of semantic relatedness (Figure 11). Relatedness is dominated by GloVe; the top 10 models are all GloVe (with the eleventh best model a PPMI model, another co-occurrence model). This is consistent with Pennington, Socher, and Manning's (2014) claims that GloVe was developed with the goal of learning longer-range semantic information. Moreover, GloVe does this over a range of hyperparameter combinations, although Figure 11 shows that medium and long windows are generally better for capturing thematic relations. Following the GloVe models are a mixture of PPMI models and SG(w) models with medium and long windows. Conversely, for similarity, SG(w) and SG(m) models with short and medium window sizes are the best performing.

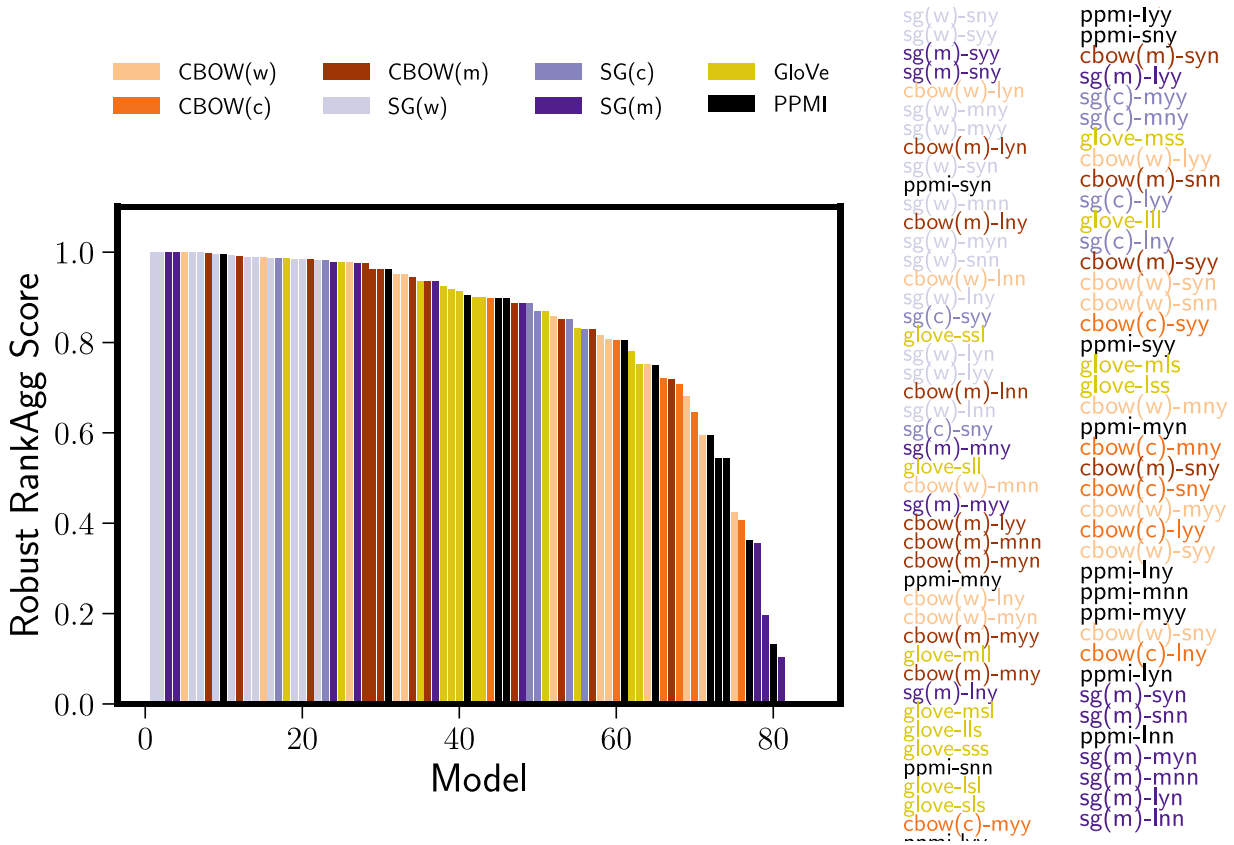


Figure 10. Ranking of 84 models using their performance over the similarity relations (SimLex, WordSim-Sim, Function, Shape, Manipulation, Color, SimVerb-3500). The bar height is the sorted robust rank aggregation score, which is used to determine rank ordering. Skip-gram models, particularly those using word embeddings, are the best performing models. Note that we have used all-lowercase shorthand for the model abbreviations, for better legibility.

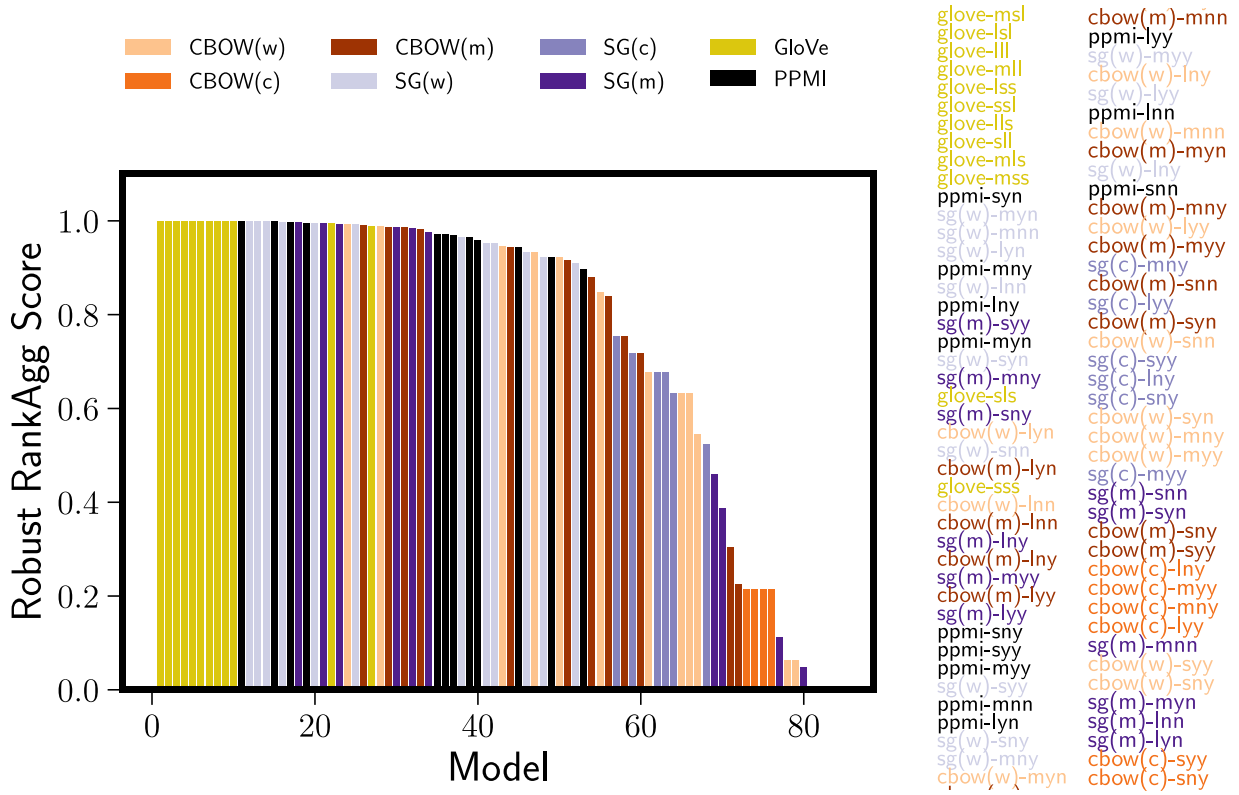


Figure 11. Ranking of 84 models using their performance over the relatedness, thematic, and event-based relations (WordSim-Rel, Verb-Agent/Patient, Verb-Instrument, Verb-Location, Event-Person, Event-Thing, Instrument-Person, Instrument-Thing, Location-Animate, Location-Thing). The bar height is the sorted robust rank aggregation score, which is used to determine rank ordering. GloVe models perform best here, followed by Skip-gram. Note that we have used all-lowercase shorthand for the model abbreviations, for better legibility.

One of the primary conclusions from these analyses is that model class is generally more important for achieving good overall performance than are specific hyperparameter values. For example, in Figure 11, our original hypothesis that long windows will perform better in capturing thematic relations is not borne out by the results. GloVe models of all window sizes are better than other models with long windows, and the best non-GloVe model (PPMI-SYN) has short windows. Furthermore, many models with long windows are among the worst performing. As another example, the best performing embedding type is similarly mixed between word and mean embeddings.

Are there any obvious differences in model architecture that would explain these broad patterns of differential model performance, specifically GloVe's superior performance on thematic relations, and Skip-gram's on similarity relations? For GloVe, a useful comparison can be made with PPMI. PPMI and GloVe are reasonably similar models, and certainly more similar to each other than to either of the word2vec models. Both models begin with co-occurrence matrices, and GloVe's word vectors are produced from a factorization of the PPMI matrix. This factorization, while not an orthogonal decomposition, may act in a similar manner to decompositions like those used in principal components analysis. In PCA, an eigendecomposition of the covariance matrix is able to find new, effective degrees of freedom that can be complex combinations of the original ones (coordinate axes, in this case). In a similar manner, GloVe's matrix decomposition may be able to produce combinations of local co-occurrence patterns to produce word vectors that can be similar in direction despite not consistently occurring in close proximity in text.

The default implementation of GloVe (Pennington et al, 2014) uses as its representations a combination of "word" and "context" vectors; the factorization it performs decomposes the log-count matrix into the product of two matrices (plus bias terms) and rows/columns of both matrices are combined to form GloVe's type-level representation of a word. Figure 7 shows that mean embeddings for both CBOW and Skip-gram perform better on thematic relations across the board; the best performing mean embedding CBOW and Skip-gram models (plus signs in Figure 7) are much closer to GloVe than word or context embeddings alone. Levy et al. (2015) point out that the effect of adding context to word vectors converts the cosine similarity function into a weighted sum of two terms: one (first order similarity) that is high when one word appears in the context of another, and another (second order similarity) that is high when the two words

are replaceable. The addition of context vectors adds first order similarity, and since thematic relations are about shared context and not substitutability, models that incorporate both (whether GloVe or Skip-gram/CBOW) perform better on those relations.

Conversely, notice in Figure 10 that the best models on similarity relations are predominantly Skip-gram and CBOW, and they overwhelmingly use word embeddings alone. Since similarity relations are more dominated by substitutability, it makes sense that letting second order similarity dominate (as is the case when context embeddings are omitted) would lead to better performance on these relations. Finally, we have no obvious answer as to why simply including context embeddings in Skip-gram and CBOW's representations is insufficient to obtain performance on thematic relations comparable to GloVe. However, we do note that Levy et al. (2015) point out that the bias parameters in GloVe are learned during training, giving GloVe additional degrees of freedom when compared to Skip-gram and CBOW, which could also contribute to better performance in the tasks we consider.

5.7. Clustering in Model and Relations Space

Finally, we present an additional analysis of overall model performance. Figure 12 shows a heatmap of the Spearman correlations of each of the 84 models for each of the 19 relations. The matrix has been hierarchically clustered in both model and relations space using unweighted average linkage (Sokal and Michener, 1958) and Euclidean distance, with cluster numbers being determined using the Gap* statistic (Mohajer, Englmeier, & Schmid, 2010). Relations are clustered on the x-axis, with the three major clusters indicated by the background color (green, magenta, or gray) of the x-axis labels. The y-axis is organized by model similarity, with the 6 major clusters of models indicated by the background color (brown, gray, orange, green, blue, or cyan) of the y-axis labels.

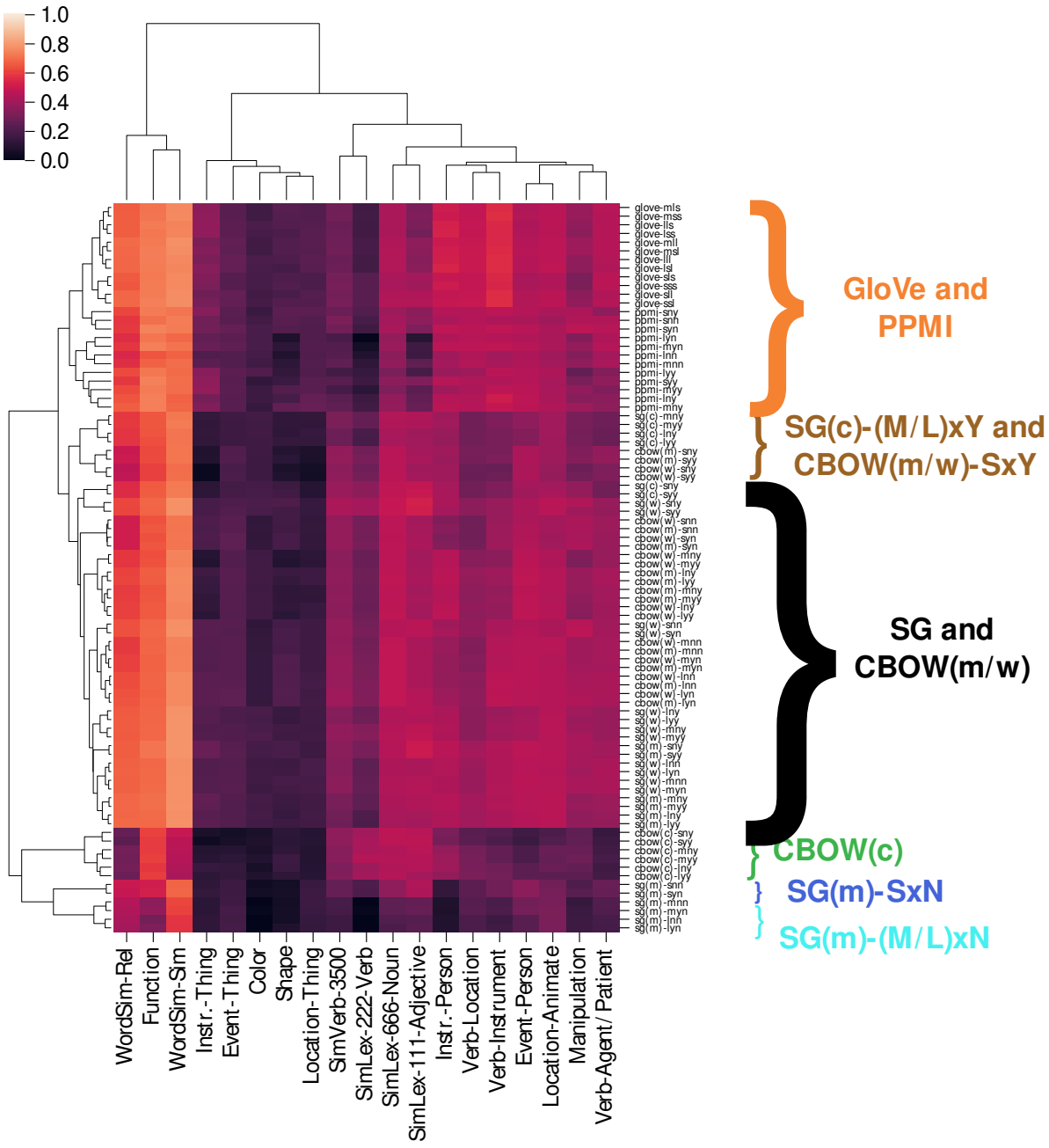


Figure 12. Clustering in model and relations space. Shown here is the (clustered) matrix of Spearman correlations of all 84 models we tested against all 19 relations. Hierarchical clustering with unweighted average linkage was performed on both relations (columns) and models (rows). Each dendrogram was cut to obtain clusters using the Gap* statistic (see text). Cluster identity is denoted by the colored boxes on the axis labels; there is no correspondence between coloring for model (row) and relation (column) clusters. Guides to the membership of the model clusters (the row clusters) are given in corresponding colored text at the right side of the plot. Note that we have used all-lowercase shorthand for the model abbreviations, for better legibility.

On the x- and y-axes we have used colored boxes to highlight discrete clusters of both relations (x-axis, 3 clusters) and models (y-axis, 6 clusters).The three relations clusters have green, gray, and magenta boxes, and the six model clusters use orange, brown, gray, green, blue, and cyan coloring.

When looking at the grouping by relation, one can see that the models do particularly well on noun similarity relations (green box, x-axis; WordSim-Sim, WordSim-Rel and Function) and poorly on verb similarity (gray box, x-axis; SimVerb-3500 and SimLex-222-V), thematic relations involving inanimate objects (gray box, x-axis; Location-Thing, Instrument-Thing, Event-Thing), and Color and Shape ratings (gray box, x-axis). This is consistent with the bar graphs shown previously. Performance on thematic relations involving animals and people (magenta box, x axis; Event-Person, Instrument-Person, Location-Animate) is substantially better than on thematic relations involving inanimate objects (gray box, x axis), which was discussed above as the "birthday/food" problem. Finally, it is clear that the models are better at capturing verb-noun relations (magenta box, x axis; Verb-Instrument, Verb-Location, Verb-Agent/Patient) than verb similarity relations (gray box, x axis; SimVerb-3500 and SimLex-222-V). This difference was discussed in Section 5.3. We also note that this performance hierarchy is generally preserved across models and hyperparameters, with the exception of some SG models that do not use negative sampling (blue box, y axis) and CBOW using context embeddings (green box, y axis).

Clustering by model shows little mixing of the types of DSMs in the clusters; co-occurrence models (GloVe and PPMI) occupy a single cluster (orange box, y axis) . None of the six clusters mix word2vec and co-occurrence based models. In contrast, models with different hyperparameters generally *do* cluster together so long as they are the same DSM type.

Irrespective of hyperparameters, all of the GloVe and PPMI models are in a single cluster (orange box, y axis). The vast majority of CBOW and SG models are also in a single cluster (gray box, y axis). An exception to this pattern is that any SG or CBOW models with mean or context embeddings that do not use negative sampling are in separate clusters (short window SG(m) models in the blue box, y axis, medium/long window SG models in the cyan box, y axis, and CBOW(c) models in the green box, y axis). This is consistent with our observations that negative sampling improves performance of SG and CBOW models that use context embeddings, and that CBOW(c) models are generally poor. Looking across the columns, one can see this clear drop in correlation for the blue, green, and cyan groups (all on the y axis) when compared to the orange and gray groups (both on the y axis). In particular, SG(m) models that use negative sampling separate from those that do not; the former are in the gray box, y axis, and the latter in the blue and cyan boxes, y axis.

6. General Discussion

The goal of the present article was to investigate the degree to which a sample of currently best-performing count and predict DSMs capture a range of semantic relations that are important for language processing. The primary novel contribution lies in testing DSMs on semantic relations that were divided into 19 datasets based on semantic relations and part of speech. These included similarity for nouns, verbs, and adjectives, as well as function, shape, manipulation, and color similarity for nouns. We also tested verb-based thematic relations, noun-noun event-based relations, and a dataset that includes multiple semantic relations. Finally, we investigated CBOW and Skip-gram context embeddings, and the average of the word and context embeddings, which has rarely been done.

Note that our theoretical approach and expectations contrast with those of Lake and Murphy (2021). As part of their studies, they judged DSMs in terms of their ability to capture exclusively semantic similarity. Lake and Murphy were interested in the degree to which DSMs capture superordinate category membership; they considered cases in which DSMs captured other semantic relations to be a failure. In contrast, because our interest was spurred by a desire to use DSMs as a basis for networks that might be used to simulate semantic and language processing more generally, we view DSM sensitivity to multiple types of semantic relations as a feature, rather than a bug.

Noun similarity, echoing other studies, is clearly the easiest kind of relation for these models to encode. To a great extent, the models were constructed to capture this relation. Correlations with human ratings approach 0.75 for both WordSim-Sim and word pairs sharing a functional relationship, and approach 0.5 for SimLex-666-N. Surprisingly, we found weak but significant correlations for perceptually similar items (manipulation, shape and color, see Lewis et al, 2019 for a similar result for shape and color). In general, the models are better at verb-noun relations than verb similarity, although some of this difference is almost certainly due to specific instructions given to SimLex and SimVerb study participants that would be difficult to emulate with these models (i.e., to ignore antonyms, although they clearly are related). Some models, GloVe in particular, do almost as well on verb/noun relations that are important components of sentence comprehension as they do on noun similarity. Finally, all models to some degree match human ratings on the event-based relations, performing better on sets involving people and animals than those involving things, an issue we discussed in detail above and revisit subsequently.

In terms of differentiating among models, there are two main takeaways: (1) there is no winner-take-all model, and (2) model class/type is generally more important than hyperparameter values. With respect to (1), when looking at overall rankings, in terms of similarity ratings, Skip-gram and CBOW models dominate (Figure 12), so if a researcher's goal is to construct vectors that are sensitive to similarity, Skip-gram, and to a lesser extent CBOW, appears to be the best choice. For the conglomeration of thematic and event-based relations, the top 10 models are all GloVe, with 8 of those using medium (15 word) or long (27 word) windows (Figure 13). Therefore, if a researcher's goal is to construct vectors that are sensitive to these other relations, especially those that play key roles in sentence processing or event cognition, GloVe appears to be the best choice. When we ranked the models based on their performance on all of the relations we examined, both GloVe and the word2vec models (CBOW and Skip-gram) appear near the top (Figure 11). Moreover, all rankings we produced (Figures 11-13) are crowded at the top; the difference between the best model and the 10th-best model can be in the third or fourth decimal place (in terms of score).

Returning to (2) above, the relative importance of model class over hyperparameter settings, even parameters that one would think should have a strong influence - for example, needing longer windows to capture thematic relations - are less important than model type. GloVe is dominant on non-similarity semantic relations, despite using a mix of medium (15 word), long (27 word), and short (5 word) windows. Conversely, word2vec models with medium-length windows are competitive with short windows near the top of the similarity rankings. Finally, even negative sampling, while generally favored, is not absolutely required for good model behavior (except in the case of using solely the context embeddings from Skip-gram and CBOW), as there are two CBOW models (CBOW(w)-LYN and CBOW(m)-LYN) and one

Skip-gram model (SG(w)-SYN) without negative sampling in the top 10 for similarity relations, and two models (SG(w)-MNN and SG(w)-MYN) in the top 10 overall.

Finally, the item-level analysis we conducted to explain model performance on SimLex-222-V and SimVerb-3500 and the noun-noun event-based ratings suggest two additional lessons. The verb similarity results indicate that there can be important subtleties to consider when comparing DSMs to human data; model performance *may* reflect the inability of a particular model to capture a certain kind of relation, but it may also have to do with specific instructions given to the human participants that the models either are not, or cannot, be exposed to. Regarding the event-based relations, our analysis points toward features of human event knowledge that none of the DSMs we tested can capture. In looking at items with low DSM scores but high participant ratings, a pattern emerges: they are pairs in which the conditional probabilities of the two items are highly asymmetric. For example, returning to the *birthday/food* example, given that you are at a *birthday*, the likelihood of encountering *food* is quite high. However, solely given that you are eating *food*, the likelihood that you are at a *birthday* is extremely low. In our rating study, the human participants were given the event name (*birthday*) and asked to rate the likelihood that a set of things would be found there, which led them to indicate that *food* is highly related to *birthday*. Because all of the DSMs we considered align words in vector space when they co-occur in highly similar contexts, the DSMs tend to score words as highly related only when both of the items are highly likely to imply the other. For example, given that you are in a *cemetery*, it is highly probable you will see *gravestones*, and given you are looking at *gravestones*, it is also highly likely you are in a *cemetery*. *Cemetery/bench*, on the other hand, is more akin to *birthday/food*.

What, then, can we say about the relation between DSMs and human semantic knowledge? This is a complex question that, in our view, consists of at least three parts, not necessarily independent: (1) do DSMs learn word representations that show some functional equivalence to what humans learn, (2) do they acquire that knowledge in a way that is similar to the way in which humans acquire it, and (3) can they be queried for that knowledge in the way we can query a human? Based on our findings and those cited in the Introduction, it would be unwise to claim that DSMs are complete models of human semantic knowledge. Lake and Murphy (2021) detail many ways in which DSMs (including models like BERT and GPT) fall short of humans, among which are: they have no grounding in, or influence from, perception; they acquire no actual knowledge about the world; they have difficulty with conceptual combinations ("butter knife"); and they cannot respond to instructions. To that, we would add that our own findings on event relations suggest the models also have serious difficulty in understanding relations with highly asymmetric conditional probabilities (what we called the *birthday/food* problem) and in relations involving words (like light verbs) that can occur in many, drastically different contexts. The issue of instructions (our (3) above) is clearly relevant to what we found as well; all the models perform suboptimally on color and shape similarity, likely because objects similar to each other only in one of these two dimensions are highly unlikely to occur in consistent contexts. However, if you asked human participants "how similar are *puck* and *spider*?" instead of "how similar in **color** are *puck* and *spider*," you would get a dramatically different answer. The fact that 1) these DSMs (and many other NLPs) cannot be given these kinds of instructions and 2) we cannot unpack their latent variables to separate different types of semantic relations often means that it can be hard to determine if the problem

is the representation they form or the inability to instruct them to focus on a particular aspect of that representation.

What about the good? If talking specifically about the models we consider here, they clearly do learn some word representations that are functionally similar to human representations, at least as far as relations between concepts are concerned. All of the models we consider perform well on noun and function similarity. As far as how they learn, there is at least one way in which this is realistic: the huge statistical learning literature has made it clear that humans can learn from regularities in the environment, even when those regularities are weak and the environment is relatively impoverished. The classic experiment of Saffran, Aslin, and Newport (1996) showed that young babies can differentiate "words" from "nonwords" when trained on a continuous stream of syllables in which the conditional probabilities have been manipulated to make certain syllabic patterns more frequent. This occurs without any acoustic markers of word or phrasal boundaries, or anything like grammar, in the input stream. DSMs do something very much like this; they learn from regularities of word patterns. We note that one of the weaknesses of the type-based DSMs, not shared by BERT or GPT, is that input is not structured; sentences are decomposed into bags of words in which order is arbitrary. It is thus clear why type-based DSMs cannot discriminate between agent-verb and verb-patient relations (and our analyses did not require them to do so).

As to whether this structure is learned in the same way as humans do (our (2) above), the models are weaker in that regard. DSMs are not necessarily a functional account of semantic knowledge, but there may be some overlap. An important aspect of learning language appears to be learning co-occurrences. When people rate various semantic relations, knowledge of co-occurrences derived from language input likely influences their ratings to some degree. This

may be particularly important for types of information that are not present in people's physical real world environment. On the other hand, it is unclear that the human mechanism for doing this would be direct (as in a DSM), rather than being an emergent property of some other system (for example, next word prediction). A similar tension between neurobiological and algorithmic accounts has recently been discussed in a different context (network models versus DSMs; Kumar et al, 2021), in which the authors recognized the power of learning co-occurrences but were similarly wary of claiming that DSMs are good models for human cognition. Rather than hope that any one of these models is actually a complete model of human semantic knowledge, researchers should instead use them pragmatically depending on what kinds of materials and relations they wish to consider.

Despite these drawbacks, in the absence of a method to empirically generate the basis for human semantic relation judgments at large scales, DSMs of various stripes remain the best current way to predict or fit human semantic knowledge at a large scale. Even if no single DSM provides a general account, the fact that DSM performance is somewhat relation-specific gives us two things. First, this yields an automated way of predicting human judgements for a variety of relations, which can be useful for experimental control in a variety of settings. Second, it gives researchers a set of levers to try to unpack aspects of relational knowledge. For example, if the algorithm and parameters of one particular DSM provide the best fit to shape similarity, whereas those of another DSM provide the best fit to manipulation similarity, what hypotheses can be generated about the basis for human ratings on these dimensions?

7. Conclusions

Quantifying human semantic knowledge is a long-standing and multifaceted challenge. The ability to translate theories of human semantic knowledge into numbers that specify the

relation between any arbitrary pair of words, no matter how concrete or abstract they are, has been a holy grail in this area. This ability would provide researchers with the means to compare theories (do the values that follow from one theory account for more variance in human performance than those of another?) and to control experimental item selection, as well as many other applications.

DSMs have a long history of use in cognitive science (Landauer & Dumais, 1997; Lund & Burgess, 1996), have increased in use as technological innovations have produced more sophisticated models (e.g., Mikolov et al., 2013), and appear to provide surprisingly robust characterizations of semantic relatedness. DSMs produce quantified representations that provide a measure of distance between any pair of concepts, they can be applied to any word (not just concrete words), and they can be derived automatically from large text corpora.

But to what degree do DSMs capture different, specific aspects of semantic relations that we can measure in experiments with human subjects? Our goal in this article was to compare currently best-performing DSMs on a broad range of semantic relations. Despite the large amount of research that has been done with DSMs to date, we believe that the comparisons of DSMs and the range of human performance data presented here will provide a unique resource for researchers who are interested in using DSMs as a basis for constructing models of language comprehension.

Acknowledgements

This research was supported in part by grants NSF 2043903 (PIs KB and JM), as well as by the Basque Government through the BERC 2022-2025 program and by the Spanish State Research

Agency through BCBL Severo Ochoa excellence accreditation CEX2020-001010-S and project PID2020-119131GB-I00 (BLIS) (JM). Support was also provided by a Natural Sciences and Engineering Research Council of Canada grant 05652 to KM.

References

Adams, D. (1979). The Hitchhiker's Guide to the Galaxy. Pan Books, London, United Kingdom.

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Pasca, M., & Soroa, A. (2009). A study on similarity and relatedness using distributional and wordnet-based approaches. *In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, Stroudsburg, PA, USA, NAACL '09 (pp. 19–27).

Andrews, M., Vigliocco, G. & Vinson, D. P. (2009). Integrating experiential and distributional data to learn semantic representations. *Psychological Review, 116*(3), 463-498.

Apfelbaum, K. S., Goodwin, C., Blomquist, C., & McMurray, B. (2023). The development of lexical competition in written-and spoken-word recognition. *Quarterly Journal of Experimental Psychology*, *76*(1), 196-219.

Asr, F. T., Zinkov, R., & Jones, M. (2018, June). Querying word embeddings for similarity and relatedness. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1: Long Papers* (pp. 675-684).

Baroni, M., & Lenci, A. (2010). Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, *36*, 673-721.

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics; Volume 1: Long Papers* (pp. 238-247).

Bommasani, R., Davis, K., & Cardie, C. (2020). Interpreting pretrained contextualized representations via reductions to static embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 4758-4781).

Buchanan, E. M., Valentine, K. D., & Maxwell, N. P. (2019). English semantic feature production norms: An extended database of 4436 concepts. *Behavior Research Methods*, *51*, 1849-1863.

Bullinaria, J., & Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, *39*, 510-526.

Carlson, G. N., & Tanenhaus, M. K. (1988). Thematic roles and language comprehension. *Syntax and Semantics*, *21*, 263-288.

Chronis, G., & Erk, K. (2020). When is a bishop not like a rook? When it's like a rabbi! Multi-prototype BERT embeddings for estimating semantic relationships. In *Proceedings of the 24th Conference on Computational Natural Language Learning* (pp. 227-244).

Connell, L., & Lynott, D. (2014). I see/hear what you mean: semantic activation in visual word recognition depends on perceptual attention. *Journal of Experimental Psychology: General*, *143*(2), 527.

De Deyne, S., Navarro, D. J., Collell, G., & Perfors, A. (2021). Visual and Affective Multimodal Models of Word Meaning in Language and Mind. *Cognitive Science*, *45*(1), e12922.

De Deyne, S., Navarro, D. J., & Storms, G. (2013). Better explanations of lexical and semantic cognition using networks derived from continued rather than single-word associations. *Behavior Research Methods*, *45*, 480-498.

De Deyne, S., Navarro, D. J., Perfors, A., Brysbaert, M., & Storms, G. (2019). The "Small World of Words" English word association norms for over 12,000 cue words. *Behavior Research Methods*, *51*, 987-1006.

Dennis, S. (2005). A memory-based theory of verbal cognition. *Cognitive Science, 29*, 145-193.

Desroches, A. S., Friesen, D. C., Teles, M., Korade, C. A., & Forest, E. W. (2022). The dynamics of spoken word recognition in bilinguals. Bilingu*alism: Language and Cognition*, *25*(4), 705-710.

Erk, K., Pado, S, & Pado, U. (2010). A flexible, corpus-driven model of regular and inverse selectional preferences. *Computational Linguistics*, *36*, 723–763.

Estes, Z., Golonka, S., & Jones, L. L. (2011). Thematic thinking: The apprehension and consequences of thematic relations. *Psychology of Learning and Motivation: Advances in Research and Theory*, *54*, 249–294.

Ethayarajh, K. (2019). How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 29th Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (pp. 55-65).

Ettinger, A. (2019). What BERT is not: lessons from a new suite of psycholinguistic diagnoses for language models. *arXiv:1907.13528*.

Hill, F., Reichart, R., & Korhonen, A. (2015). Simlex-999: Evaluating semantic models with genuine similarity estimation. *Computational Linguistics*, *41*, 665–695.

Fellbaum, C. (1998). WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

Fernandino, L., Tong, J. Q., Conant, L. L., Humphries, C. J., & Binder, J. R. (2022). Decoding the information structure underlying the neural representation of concepts. *Proceedings of the National Academy of Sciences*, *119*(6), e2108091119.

Ferretti, T. R., McRae, K., & Hatherell, A. (2001). Integrating verbs, situation schemas, and thematic role concepts. *Journal of Memory and Language*, *44*, 516-547.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., & Ruppin, E. (2002). Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on world wide web, ACM*, New York, NY (pp. 406–414).

Gerz D, Vuli I, Hill F, Reichart R, Korhonen A. (2016). SimVerb-3500: A large-scale evaluation set of verb similarity. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 2173–2182). Stroudsburg, PA: Association of Computational Linguistics.

Giovannone, N., & Theodore, R. M. (2021). Individual differences in lexical contributions to speech perception. *Journal of Speech, Language, and Hearing Research*, *64*(3), 707-724.

Goh, W. D., Yap, M. J., & Chee, Q. W. (2020). The Auditory English Lexicon Project: A multi-talker, multi-region psycholinguistic database of 10,170 spoken words and nonwords. *Behavior Research Methods*, *52*(5), 2202-2231.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review, 114*, 211-244.

Hare, M., Jones, M. N., Thomson, C., Kelly, S., & McRae, K. (2009). Activating event knowledge. *Cognition*, *111*, 151-167.

Hills, T. T., Maouene, M., Maouene, J., Sheya, A., & Smith, L. (2009). Longitudinal analysis of early semantic networks: Preferential attachment or preferential acquisition? *Psychological science*, *20*, 729-739.

Hutchison, K. A., Balota, D. A., Neely, J. H., Cortese, M. J., Cohen-Shikora, E. R., Tse, C. S., ... & Buchanan, E. (2013). The semantic priming project. *Behavior Research Methods*, *45*, 1099-1114.

Johns, B. T., Mewhort, D. J. K., & Jones, M. N. (2019). The role of negative information in distributional semantic learning. *Cognitive Science*, 43, e12730.

Johns, B. T. (2021). Distributional social semantics: Inferring word meanings from communication patterns. *Cognitive Psychology*, 131, 101441.

Jones, M. N., Kintsch, W., & Mewhort, D. J. K. (2006). High-dimensional semantic space accounts of priming. *Journal of Memory and Language, 55*, 534-552.

Jouravlev, O., & McRae, K. (2016). Thematic relatedness production norms for 100 object concepts. *Behavior Research Methods*, 48, 1349-1357.

Kacmajor, M., & Kelleher, J. D. (2020). Capturing and measuring semantic relations. *Language Resources & Evaluation*, *54*, 645-682.

Kolde, R., Laur, S., Adler, P., & Vilo, J. (2012). Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics*, *28*, 573-580.

Kumar, A., Steyvers, M., & Balota, B. (2021). DA critical review of network-based and distributional approaches to semantic memory structure and processes. *Topics in Cognitive Science*, *0,* 1-24.

Kwantes, P. J. (2005). Using context to build semantics. *Psychonomic Bulletin & Review, 12,* 703-710.

Landauer, T. K., & Dumais, S. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211-240.

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

Lapesa, G., & Evert, S. (2013). Evaluating neighbor rank and distance measures as predictors of semantic priming. In *Proceedings of the ACL Workshop on Cognitive Modeling and Computational Linguistics* (*CMCL 2013*), pp. 66-74.

Lapesa, G., Evert, S., & Schulte im Walde, S. (2014). Contrasting syntagmatic and paradigmatic relations: Insights from distributional semantic models. In *Proceedings of the Third Joint Conference on Lexical and Computational Semantics* (*SEM 2014*), pp. 160-170.

Lazaridou, A., Pham, N. T., & Baroni, M. (2015). Combining language and vision with a multimodal Skip-gram model. arXiv:1501.02598.

Lenci, A. (2018). Distributional models of word meaning. *Annual review of Linguistics*, *4*, 151-171.

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

Lewis, M., Zettersten, M., & Lupyan, G. (2019). Distributional semantics as a source of visual knowledge. *Proceedings of the National Academy of Sciences*, 116 (39), 19237-19238.

Liu, N.F., Gardner, M., Belinkov, Y., Peters, M. E., & Smith, N. A. (2019). Linguistic knowledge and transferability of contextual representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers),* pp. 1073-1094.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavioral Research Methods, Instrumentation, and Computers, 28*, 203-208.

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, *92*, 57-78.

McMurray, B., Apfelbaum, K. S., & Tomblin, J. B. (2022). The slow development of real-time processing: Spoken-word recognition as a crucible for new thinking about language acquisition and language disorders. *Current Directions in Psychological Science*, *31*(4), 305-315.

McNamara, T. (2005). *Semantic Priming.* New York, NY: Psychology Press.

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

McRae, K., Khalkhali, S., & Hare, M. (2012). Semantic and associative relations: Examining a tenuous dichotomy. In V. F. Reyna, S. B. Chapman, M. R. Dougherty, & J. Confrey (Eds.), *The Adolescent Brain: Learning, Reasoning, and Decision Making* (pp. 39-66). Washington, DC: APA.

McRae, K., Spivey-Knowlton, M. J., & Tanenhaus, M. K. (1998). Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language*, *38*, 283-312.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013a). Efficient estimation of word representations in vector space. arXiv:1301.3781.

Mikholov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems, 28* (2013)

Miller, G. A. (1995). WordNet: A Lexical Database for English. Communications of the ACM Vol. 38, No. 11: 39-41.

Mirman, D., & Magnuson, J. S. (2008). Attractor dynamics and semantic neighborhood density: processing is slowed by near neighbors and speeded by distant neighbors. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 65-79.

Mohajer, M., Englmeier, K-H., & Schmid, V. J. (2010). A comparison of Gap statistic with and with-out logarithm function. Technical Report No. 096, Department of Statistics, University of Munich.

Musz, E., Yee, E., & Thompson-Schill, S. L. (2012). *Mapping the similarity space of concepts in sensorimotor cortex*. Poster presented at the 2012 Meeting of the Cognitive Neuroscience Society, Chicago, IL.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, *36*, 402–407.

Nenadić, F., Podlubny, R. G., Schmidtke, D., Kelley, M. C., & Tucker, B. V. (2022). Semantic richness effects in isolated spoken word recognition: Evidence from massive auditory lexical decision. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Advance online publication. https://doi.org/10.1037/xlm0001208

Nitsan, G., Banai, K., & Ben-David, B. M. (2022). One size does not fit all: examining the effects of working memory capacity on spoken word recognition in older adults using eye tracking. *Frontiers in Psychology*, *13*.

Padó, S. (2007). *Cross-lingual annotation projection models for role-semantic information*. Saarland University.

Papies, E. K. (2013). Tempting food words activate eating simulations. *Frontiers in Psychology*, *4*, 838.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (EMNLP), Doha, Qatar: Association for Computational Linguistics (pp. 1532–1543).

Pexman, P. M., Hargreaves, I. S., Siakaluk, P. D., Bodner, G. E., & 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., & Yap, M. J. (2017). The Calgary semantic decision project: concrete/abstract decision data for 10,000 English words. *Behavior research methods*, *49*, 407-417.

Poeppel, D. & Idsardi, W. (2022). We don't know how the brain stores anything, let alone words. *Trends in Cognitive Sciences*, *26*(12), 1054-1055. https://doi.org/10.1016/j.tics.2022.08.010

Rabovsky, M., & McRae, K. (2014). Simulating the N400 ERP component as semantic network error: Insights from a feature-based connectionist attractor model of word meaning. *Cognition*, *132*, 68-89.

Rapp, R. (2002). The computation of word associations: Comparing syntagmatic and paradigmatic approaches. In *Proceedings of the 19th International Conference on Computational Linguistics*, *1*, 1–7.

Rehurek, R., & Sojka, P. (2011). Gensim–python framework for vector space modelling. NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic, 3(2).

Rogers , T. T. , Lambon Ralph, M. A., Garrard, P., Bozeat, S., McClelland, J. L., Hodges, J. R., and Patterson, K. (2004). The structure and deterioration of semantic memory: A neuropsychological and computational investigation. *Psychological Review*, *111*, 205-235.

Rotaru, A. S., Vigliocco, G., & Frank, S. L. (2018). Modeling the structure and dynamics of semantic processing. *Cognitive Science*, *42*, 2890-2917.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, 274 (5294), 1926-1928.

Sahlgren, M. (2006). *The word-space model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high dimensional vector spaces*. Ph.D. thesis, University of Stockholm.

Santus, E., Chersoni, E., Lenci, A., & Blache, P., (2017). Measuring Thematic Fit with Distributional Feature Overlap. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 659-669.

Sayeed, A., Greenberg, C., & Demberg. V. (2016). Thematic fit evaluation: An aspect of selectional preferences. In Proce*edings of ACL Workshop for Evaluating Vector Space Representations for NLP*.

Sokal, R., & Michener, C. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, *38*, 1409-1438.

Stella, M., Beckage, N. M., & Brede, M. (2017). Multiplex lexical networks reveal patterns in early word acquisition in children. *Scientific Reports*, *7*, 1-10.

Taylor, W. L. (1953). Cloze procedure: a new tool for measuring readability. *Journalism Bulletin, 30 (4)*, 415-433.

Tenney, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R. T., Kim, N., Van Durme, B., Bowman, S. R. Das, D., & Pavlick, E. (2019). What do you learn from context? Probing for sentence structure in contextualized word representations. *International Conference on Learning Representations 2019.*

Tilk, O., Demberg, V., Sayeed, A., Klakow, D., & Thater, S. (2016, November). Event participant modelling with neural networks. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*(pp. 171-182).

Troyer, M., & Kutas, M. (2020). To catch a Snitch: Brain potentials reveal variability in the functional organization of (fictional) world knowledge during reading. *Journal of Memory & Language*, *113*, 104-111.

Utsumi, A. (2015). A complex network approach to Distributional Semantic Models. PLoS ONE 10(8): e0136277. doi:10.1371/journal.pone.0136277

Van der Maaten, L.J.P, & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research, 9,* 2579-2605.

Van Paridon, J., & Thompson, B. (2021). Subs2vec: Word embeddings from subtitles in 55 languages. *Behavior Research Methods*, *53*, 629-655.

Vasmani, A., Shazeer, N., Parmar., N., Uszkoreit, J., Jones, L., Gomez, A. N., Laiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *NIPS 2017*.

Vinson, D. P., & Vigliocco, G. (2008). Semantic feature production norms for a large set of objects and events. *Behavior Research Methods, 40*, 183-190.

Wingfield, C., & Connell, L. (2022). Understanding the role of linguistic distributional knowledge in cognition. *Language, Cognition and Neuroscience*, *37*(10), 1220-1270.

Yee, E., Ahmed, S. Z., & Thompson-Schill, S. L. (2012). Colorless green ideas (can) prime furiously. *Psychological Science*, *23*(4), 364-369.

Yee, E., Drucker, D. M., & Thompson-Schill, S. L. (2010). fMRI-adaptation evidence of overlapping neural representations for objects related in function or manipulation. *NeuroImage*, *50*, 753-763.

Yee, E., Huffstetler, S., & Thompson-Schill, S. L. (2011). Function follows form: Activation of shape and function features during object identification. *Journal of Experimental Psychology: General*, *140*, 348-363.

Acknowledgements

This research was supported in part by grants NSF 2043903 (PIs KB and JM), as well as by the Basque Government through the BERC 2022-2025 program and by the Spanish State Research Agency through BCBL Severo Ochoa excellence accreditation CEX2020-001010-S and project PID2020-119131GB-I00 (BLIS) (JM). Support was also provided by a Natural Sciences and Engineering Research Council of Canada grant 05652 to KM.

Appendix

Events, Locations, and Instruments Ratings Methods and Descriptive Results

Methods

Participants

One-hundred, forty-one students (mean age = 19, range = [17-24]) participated for partial course credit. Approximately 80% of participants identified as female; 19% identified as male; and 1% identified as non-binary. All participants gave informed consent under the approval of the institutional review board at the University of Connecticut. The number of participants who completed each list differed somewhat by cue-target type (see Materials & Procedures), which included events-people (N = 41), events-things (N = 20), locations-animate (N = 20), locations-things (N = 20), instruments-animate (N = 20), and instruments-things (N = 20). The number of participants for the events-people ratings was approximately double that of the other cue-response types due to experimenter error (we ran the condition twice), and we elected to include all data.

Materials & Procedures

The items were taken from Hare, Jones, Thomson, Kelly, and McRae (2009). In that study, participants were given cue words and produced up to five responses of a certain type, according to six different sets of instructions:

1. Events-people: Participants were presented with 45 event nouns such as *sale* and were asked to provide types of people that typically are found at those events (e.g., *clerk, shopper*).

2. Events-things: Participants were presented with 53 event nouns such as *banquet* and were asked to provide things that typically are found at those events (e.g., *wine*, *flowers*).

3. Locations-people and animals: Participants were presented with 31 nouns describing locations (e.g., *barn*) and were asked to provide people and/or animals that typically are found at those locations (e.g., *farmer*, *cow*).

4. Locations-things: Participants were presented with 61 location nouns (e.g., *bathroom*) and were asked to provide things that are typically found at those locations (e.g., *toothbrush*, *toilet*).

5. Instruments-people: Participants were presented with 45 nouns describing instruments (e.g., *wrench*) and were asked to respond with people that typically use the instrument (e.g., *plumber*, *carpenter*).

6. Instruments-things: Participants were presented with 43 instrument nouns (e.g., *scissors*) and were asked to respond with the things that people typically act upon with each (e.g., *hair*, *paper*).

Hare et al.'s resulting data contained 5854 unique cue-response pairs. For the present purposes, we selected a subset of these data subject to several constraints. A weighted score for

each response was computed based on the number of participants who provided each response in first, second, third, fourth, or fifth position in Hare et al. (with first position given five times the weight as the fifth, and so forth). First, we excluded all cue-target pairs with a weighted score of less than 3. Second, we excluded all cues with fewer than seven unique responses across Hare et al.'s participants. Finally, two authors used intuition to select approximately half of the remaining cue-response pairs to use for the present study. This included eliminating pairs with generic responses (e.g., "woman," "man"). This resulted in 2,717 pairs: events-people: 538; events-things: 592; locations-people and animals: 504; locations-thingss: 356; instruments-people: 287; instruments-things: 440).

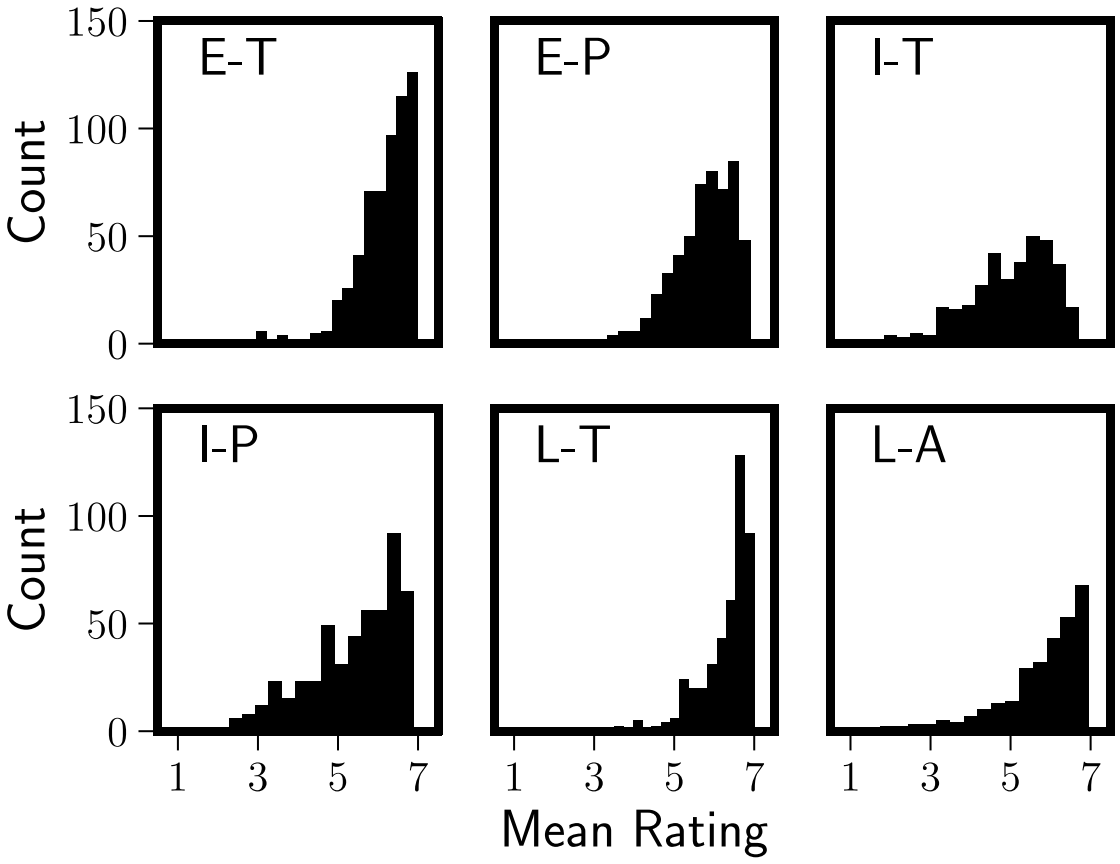
A separate list was created for each of the 6 cue-response types. A separate group of participants responded to each list. Within each list, each cue word was presented on a separate screen along with all of its responses (mean = 12.5; range = 6-20) on the same screen. Participants rated the event-based relations on a scale of 1 (Not at all likely) to 7 (Extremely likely). They were informed that "In this study, you will be presented with a list of events [or locations or instruments]. For each, we would like you to imagine situations that could take place during that event" [or location or instrument]." Specific instructions varied according to each cue-response type:

1. Events-people: How likely is each type of person to be found in this situation?
2. Events-things: How likely is each type of thing to be found in this situation?
3. Locations-people-animals: How likely is each type of person/animal to be found at this location?
4. Locations-things: How likely is each type of thing to be found at this location?
5. Instruments-people: How likely is each type of person to use the instrument?

6. Instruments-things: How likely is the instrument to be used on each type of thing?

Results

Overall, ratings were relatively high, collapsing across all six list types ($M = 5.73$, $SD = 0.47$, *range* = 1-7). Ratings varied somewhat across the six conditions: events-people ($M = 5.73$, $SD = 0.76$, *range* = 1-7), events-things ($M = 6.15$, $SD = 0.71$, *range* = 1-7), locations-people and animals ($M = 5.78$, $SD = 1.07$, *range* = 1-7), locations-things ($M = 6.30$, $SD = 0.64$, *range* = 1-7), instruments-people ($M = 5.41$, $SD = 1.12$, *range* = 1-7), and instruments-things ($M = 5.01$, $SD = 1.02$, *range* = 1-7). Histograms for each list are provided in Supplemental Figure 1.



Supplemental Figure 1. Histograms of mean ratings (by item) for each set. Abbreviations are as follows: Events-Things, E-T; Events-People, E-P; Instruments-Things, I-T; Instruments-People, I-P; Locations-Things, L-T; Locations-Animals-People, L-A.