Majewska, Olga, Diana McCarthy, Jasper J. F. van den Bosch, Nikolaus Kriegeskorte, Ivan Vulić, and Anna Korhonen. 2021. Semantic Data Set Construction from Human Clustering and Spatial Arrangement. *Computational Linguistics*, uncorrected proof.

Semantic Data Set Construction from Human Clustering and Spatial Arrangement

Olga Majewska
Language Technology Lab
University of Cambridge
om304@cam.ac.uk

Diana McCarthy
Language Technology Lab
University of Cambridge
diana@dianamccarthy.co.uk

Jasper J. F. van den Bosch
School of Psychology
University of Birmingham
vandejjf@bham.ac.uk

Nikolaus Kriegeskorte
Zuckerman Institute
University of Columbia
nk2765@columbia.edu

Ivan Vulić
Language Technology Lab
University of Cambridge
iv250@cam.ac.uk

Anna Korhonen
Language Technology Lab
University of Cambridge
alk23@cam.ac.uk

Research into representation learning models of lexical semantics usually utilizes some form of intrinsic evaluation to ensure that the learned representations reflect human semantic judgments. Lexical semantic similarity estimation is a widely used evaluation method, but efforts

Submission received: 20 April 2020; revised version received: 23 August 2020; accepted for publication: 3 December 2020.

https://doi.org/10.1162/COLI_a_00396

have typically focused on pairwise judgments of words in isolation, or are limited to specific contexts and lexical stimuli. There are limitations with these approaches that either do not provide any context for judgments, and thereby ignore ambiguity, or provide very specific sentential contexts that cannot then be used to generate a larger lexical resource. Furthermore, similarity between more than two items is not considered. We provide a full description and analysis of our recently proposed methodology for large-scale data set construction that produces a semantic classification of a large sample of verbs in the first phase, as well as multi-way similarity judgments made within the resultant semantic classes in the second phase. The methodology uses a spatial multi-arrangement approach proposed in the field of cognitive neuroscience for capturing multi-way similarity judgments of visual stimuli. We have adapted this method to handle polysemous linguistic stimuli and much larger samples than previous work. We specifically target verbs, but the method can equally be applied to other parts of speech. We perform cluster analysis on the data from the first phase and demonstrate how this might be useful in the construction of a comprehensive verb resource. We also analyze the semantic information captured by the second phase and discuss the potential of the spatially induced similarity judgments to better reflect human notions of word similarity. We demonstrate how the resultant data set can be used for fine-grained analyses and evaluation of representation learning models on the intrinsic tasks of semantic clustering and semantic similarity. In particular, we find that stronger static word embedding methods still outperform lexical representations emerging from more recent pre-training methods, both on word-level similarity and clustering. Moreover, thanks to the data set's vast coverage, we are able to compare the benefits of specializing vector representations for a particular type of external knowledge by evaluating FrameNet- and VerbNet-retrofitted models on specific semantic domains such as "Heat" or "Motion."

1. Introduction

Recent advances in representation learning have transformed the NLP landscape, introducing new powerful architectures that achieve unprecedented results on a plethora of natural language tasks (Peters et al. 2018; Devlin et al. 2019; Liu et al. 2019c; Radford et al. 2019; Yang et al. 2019, inter alia). Although high performance in downstream tasks may be the ultimate goal,[1] intrinsic evaluation benchmarks continue to provide a useful intermediary test of representation quality, with the advantages of simplicity and speed of execution. Estimation of lexical semantic similarity has been widely used as an intrinsic evaluation task, where the quality of word embeddings is assessed through comparison of distances between words in the embedding space against human judgments of semantic similarity and/or relatedness (Finkelstein et al. 2002; Agirre et al. 2009; Bruni, Tran, and Baroni 2014; Hill, Reichart, and Korhonen 2015). Further progress relies on the availability of high-quality evaluation benchmarks, challenging enough to test the limits of models' capacity to capture word semantics. However, these are still limited to a small number of typically well-resourced languages. Moreover, they predominantly focus on nouns, and less attention has been paid to the challenges posed to natural language models by the complex linguistic properties of verbs. Due to the verbs' central role in sentence structure as bearers of information pertaining to both structural and semantic relationships between clausal

1 Another goal may be modeling of human language reflecting cognitive performance for scientific purposes.

elements, attaining accurate and nuanced representations of their properties is essential to decrease the gap between human and machine language understanding (Jackendoff 1972; Levin 1993; McRae, Ferretti, and Amyote 1997; Altmann and Kamide 1999; Resnik and Diab 2000; Sauppe 2016, inter alia).

While recent efforts resulted in a large verb similarity data set for English, SimVerb-3500 (Gerz et al. 2016), the demand for challenging, wide-coverage lexical resources targeting verb semantics has not yet been fully met. Expert-built lexicons encoding rich information about verbs' semantic features and behavior such as FrameNet (Baker, Fillmore, and Lowe 1998) or VerbNet (Kipper Schuler 2005; Kipper et al. 2006) are still only available in a handful of languages, and noun-focused benchmark data sets are prevalent (Finkelstein et al. 2002; Agirre et al. 2009; Bruni et al. 2012; Hill, Reichart, and Korhonen 2015). In light of these considerations, in this article we present novel methodology which promises to mitigate the evaluation data scarcity problem and help overcome the bottleneck of slow and expensive manual resource creation.

We present our novel approach to collecting semantic similarity data by means of a two-phase design consisting of (1) *bottom–up semantic clustering* of verbs into theme classes, and (2) *spatial similarity judgments* obtained via a multi-arrangement method so far used exclusively in psychology and cognitive neuroscience research and with visual stimuli (Kriegeskorte and Mur 2012; Mur et al. 2013; Charest et al. 2014). We demonstrate how it can be adapted for the purposes of a large-scale linguistic task with *polysemous lexical stimuli* and yield wide-coverage verb similarity data. The method's promise lies in the intuitive nature of the task, where relative similarities between items are signaled by the geometric distances within a two-dimensional arena, as well as a user-friendly drag-and-drop interface. These properties of the annotation design significantly facilitate and speed up the task, as many concurrent similarity judgments are expressed with a single mouse drag. What is more, no classification structure or criteria are pre-imposed on the annotators, and similarities between individual verbs are judged *in the context of all other verbs* appearing in the arena, rather than in isolation. Crucially, the method enables word clustering and registers pairwise semantic similarity scores *at the same time*, which can be especially beneficial as a means of rapid creation of evaluation data to support NLP.

The final resource comprises 17 theme classes and *SpA-Verb*, a large intrinsic evaluation data set including 29,721 unique pairwise verb (dis)similarity scores for 825 target verbs. The SpA-Verb scores are Euclidean distances corresponding to dissimilarities between words, assembled in the representational dissimilarity matrix (RDM) (Kriegeskorte, Mur, and Bandettini 2008).[2] It surpasses the largest verb-specific evaluation resource previously available (SimVerb, with 3,500 pairwise similarity scores) by a significant margin. Thanks to its scale and vast coverage, as well as its inclusion of complete matrices of pairwise similarities for all possible pairings of verbs within a given class, SpA-Verb offers a wealth of possibilities for nuanced analyses and evaluation of semantic models' capacity to accurately represent concepts pertaining to different meaning domains and displaying different semantic properties. We demonstrate the resource's utility by evaluating a selection of state-of-the-art representation learning architectures on two tasks, corresponding to the two phases of our design: (1) clustering, using Phase 1 classes as gold truth, and (2) word similarity, using pairwise scores from

2 This effectively means that lower scores are assigned to similar verbs, and larger scores to dissimilar verbs.

the entire SpA-Verb (29,721 pairs) and the thresholded subset of 10k+ pairs, as well as selected subsets focusing on different semantic characteristics.

In our preliminary work (Majewska et al. 2020), we introduced the two-phase annotation design, described the interface and task structure, and discussed the key differences between our approach and the pairwise rating-based method used to create SimVerb-3500. This article substantially extends Majewska et al. (2020), provides a full and in-depth description and analysis of the entire annotation protocol, and also makes the following key contributions:

- We carry out cluster analysis on the output of Phase 1, applying a network analysis approach to the rough clustering data in order to scrutinize the emerging semantic classes and gain insight into annotator decisions (Section 4.2), and we discuss how this analysis is used to support the production of semantic classes for Phase 2.
- We present an in-depth examination of the semantic information captured by the two phases, rough clustering and spatial arrangements of lexical stimuli, by means of qualitative and quantitative comparative analyses with two lexical resources, FrameNet and VerbNet (Section 7). Our analyses revealed that annotators are able to differentiate between a range of semantic relations by means of relative item placements. What is more, the encouraging overlap observed with VerbNet classes suggests the method could help incorporate new verbs into the existing data set, or support creation of similar resources from scratch for other languages.
- We demonstrate the utility of the resource by evaluating a selection of representation models on two tasks, semantic clustering and word similarity, and illustrate its potential to enable nuanced, focused analyses targeting specific semantic properties and meaning domains (Section 8). In particular, our analyses reveal the primacy of static word embeddings incorporating external linguistic knowledge over state-of-the-art unsupervised Transformer-based architectures (Devlin et al. 2019) on both word-level semantic similarity and clustering. Our findings provide additional evidence in support of the vast potential of drawing on external linguistic information to help vector representations better reflect fine-grained semantic relations present in the mental lexicon: Thanks to the data set's large coverage, we can contrast performance of embeddings specialized for VerbNet and FrameNet classification information on focused semantic domains such as "experiencing/causing harm" or "applying/absorbing heat."

The article is organized as follows. Section 2 discusses related work, existing data sets, and alternative annotation protocols. Section 3 presents the structure and motivation of our annotation design and discusses the challenges involved in adapting the spatial arrangement method from visual to lexical stimuli. The two phases of our protocol are analyzed in Sections 4 and 5, respectively. The results of the inter-annotator agreement analysis are discussed in Section 6, and Section 7 presents an in-depth examination of the semantic information captured in each phase. Section 8 presents the results of the evaluation of a diverse selection of representation models on our data set.

2. Related Work

The availability of high-quality evaluation resources plays a crucial role in spurring advances in word representation learning, where the demand for challenging benchmarks to test the growing ability of models to reflect human semantics continues to rise. Lexicographic resources such as WordNet (Miller 1995; Fellbaum 1998), VerbNet (Kipper Schuler 2005; Kipper et al. 2006), or FrameNet (Baker, Fillmore, and Lowe 1998) encode a wealth of semantic, syntactic, and predicate–argument information for English words, but their reliance on experts makes them expensive and time-consuming to create. Meanwhile, crowd-sourcing has allowed us to leverage non-expert native-speaker intuitions about word meaning through a range of annotation tasks, commonly adopted as a quicker and more cost-effective alternative to produce evaluation data. Data sets consisting of human similarity ratings collected for sets of word pairs have been particularly popular (Baroni, Dinu, and Kruszewski 2014; Levy and Goldberg 2014; Pennington, Socher, and Manning 2014; Schwartz, Reichart, and Rappoport 2015; Wieting et al. 2016; Bojanowski et al. 2017; Mrkšić et al. 2017).

While word similarity data sets have been routinely used for intrinsic evaluation of general-purpose representation models, different views of what constitutes "semantic similarity" underlie their design, and there is no consensus on what meaning relationship word embeddings should capture, and what kind of signal to disregard. These varying perspectives are reflected in the different terminologies that have been adopted to refer to semantic proximity.

The term *semantic relatedness* has been used to describe words linked by any kind of semantic relation (Gentner 1983; Budanitsky and Hirst 2001, 2006; Turney and Pantel 2010), including but not limited to synonymy (*puzzle-bemuse*), meronymy and holonymy (*peel-fruit*), as well as antonymy (*light-dark*). Similarity defined as association, that is, the mental activation of a term when another is presented (Chiarello et al. 1990; Lemaire and Denhiere 2006) (e.g., *butter-knife, hammer-nail*), has been estimated in terms of how frequently the two words co-occur in the same contexts in language (and the physical world) (Turney 2001; Turney and Pantel 2010; McRae, Khalkhali, and Hare 2012; Bruni et al. 2012). One type of such associative relationship is **thematic relatedness**, which involves relations between concepts playing complementary roles in the same event or scenario (Lin and Murphy 2001; Estes, Golonka, and Jones 2011; Kacmajor and Kelleher 2020) (e.g., *dog, bark, leash, bone*).

Associative relatedness contrasts with a concept of semantic similarity defined in terms of shared superordinate category (Lupker 1984; Resnik 1995) (**taxonomical similarity** [Turney and Pantel 2010]) or common semantic features (Tversky 1977; Frenck-Mestre and Bueno 1999; Turney 2006). Viewed this way, similarity is quantified in terms of degree of overlap in semantic properties, such as shared function or physical features, with synonyms occupying the top region of the similarity scale (e.g., *fiddle-violin* [Cruse 1986]).

In this article, we reserve the term (semantic) similarity for this latter definition of closeness of meaning, as distinct from the more general *relatedness*, which includes semantic relations and taxonomical similarity, as well as (thematic) association, following previous work (Resnik 1995; Resnik and Diab 2000; Agirre et al. 2009; Hill, Reichart, and Korhonen 2015; Gerz et al. 2016; Kacmajor and Kelleher 2020). In Section 7, we explore how this distinction is captured by native speaker judgments in the two tasks constituting our annotation design, rough semantic clustering and spatial arrangements of words, through qualitative and quantitative analysis with reference to existing lexical resources.

Although widely useful, most of the data sets used for intrinsic evaluation are restricted in size and coverage, many conflate similarity and relatedness, and only a few target verbs specifically. Among the English-language resources used for intrinsic evaluation of semantic models, word pair data sets such as WordSim-353 (Finkelstein et al. 2002; Agirre et al. 2009), comprising 353 noun pairs, and SimLex-999 (Hill, Reichart, and Korhonen 2015), comprising 999 word pairs out of which 222 are verb pairs, have been prominent. Resources focused exclusively on verbs include YP-130 (Yang and Powers 2006) (130 verb pairs) and the data set of Baker, Reichart, and Korhonen (2014) (143 verb pairs), with the more recent addition of SimVerb (Gerz et al. 2016) providing pairwise similarity ratings for 3,500 English verb pairs.

While pairwise rating data sets have been regularly resorted to in intrinsic evaluation, alternative annotation methodologies and types of data sets have been proposed to address some of their limitations. Examples include *best-worst scaling* (Louviere and Woodworth 1991; Louviere, Flynn, and Marley 2015; Avraham and Goldberg 2016; Kiritchenko and Mohammad 2016, 2017; Asaadi, Mohammad, and Kiritchenko 2019), where annotators perform relative judgments of several items to decide which displays a given property to the highest and which to the lowest degree, and *paired comparisons* (Dalitz and Bednarek 2016), where the task is to determine which of the two items at hand has more of a given property. Another example is the task of outlier detection from clusters of semantically similar words (Blair, Merhav, and Barry 2017). Further, as an alternative to the words-in-isolation approach, data sets composed of judgments of similarity in context have been constructed (Huang et al. 2012; Pilehvar and Camacho-Collados 2019; Armendariz et al. 2020), where target words are presented in sentential contexts triggering a specific meaning of each word. Representation models have also been evaluated on synonym detection data sets using English as foreign language test data (Landauer and Dumais 1997; Turney 2001), word games (Jarmasz and Szpakowicz 2003), where the aim is to identify one correct synonym of the target word among 4 candidates, and on analogy (Mikolov et al. 2013a; Gladkova, Drozd, and Matsuoka 2016) and semantic relation data sets (Baroni and Lenci 2011).

The most extensive verb-oriented data set available to date, SimVerb-3500 (hereafter SimVerb), is a product of a crowd-sourcing effort with over 800 raters, each completing the pairwise similarity rating task for 79 verb pairs. In this article, we describe an alternative novel approach that enables an annotator to implicitly express multiple pairwise similarity judgments by a single mouse drag, instead of having to consider each word pair independently. This allowed us to scale up the data collection process and, starting from the same sample of verbs as those used in SimVerb, generate similarity scores for over eight times as many verb pairs. Consideration of multiple items concurrently also provides some context for ambiguous words while not relying on sentential contexts, which give rise to issues of sparsity and coverage, and are therefore less amenable to building larger lexical resources. Moreover, our approach also yields thematic item classes thanks to a precursor semantic clustering method, within which the similarity judgments are made.

3. Multi-Arrangement for Semantics

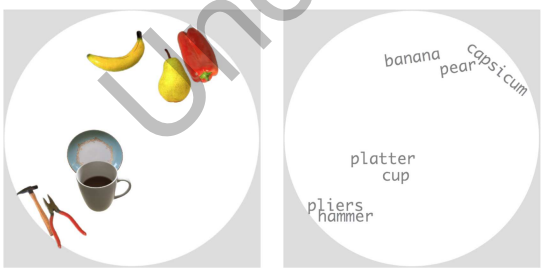
3.1 Spatial Arrangement Method (SpAM)

The spatial arrangement method (SpAM) has been previously used to record similarity judgments through geometric arrangements of visual stimuli in psychology and cognitive neuroscience (Goldstone 1994; Levine, Halberstadt, and Goldstone 1996;

Kriegeskorte and Mur 2012; Hout, Goldinger, and Ferguson 2013; Mur et al. 2013; Charest et al. 2014). However, its potential and applicability to *semantic similarity between lexical stimuli* has not yet been studied.

In the commonly used pairwise rating method (e.g., utilized to produce SimVerb) a rater is presented with a pair of words at a time and the number of possible pairwise combinations of stimuli grows factorially as the sample size increases. For a sample of n stimuli there are $n(n-1)/2$ pairwise combinations possible. However, in SpAM a subject arranges multiple stimuli simultaneously in a two-dimensional space (e.g., on a computer screen), expressing (dis)similarity through the relative positions of items within that space: Similar items are placed closer together and dissimilar ones further apart. The inter-stimulus Euclidean distances represent pairwise dissimilarities and all stimuli are considered in the context of the entire sample presented to the user. Each placement simultaneously signals the similarity relationship of the item to all other items in the set. Figure 1 illustrates this comparison.

SpAM leverages the spatial nature of humans' mental representation of concept similarity (Lakoff and Johnson 1999; Gärdenfors 2004; Casasanto 2008) and allows for a freer, intuitive expression of similarity judgments as continuous distances, rather than necessitating assignment of discrete numerical ratings. The latter, although omnipresent in intrinsic evaluation of representation models as a handy approximation of the strength of lexical relations, have been shown to have a number of limitations (Batchkarov et al. 2016; Faruqui et al. 2016; Gladkova and Drozd 2016; Kiritchenko and Mohammad 2017). Rather than reflecting semantic factors, annotators' judgments of isolated word pairs are often found to be biased by word frequency, prototypicality, order of presentation and speed of association. Moreover, subtle meaning distinctions and degrees of similarity between words are very difficult to quantify and translate onto a discrete scale without context or points of reference, in the form of other related words. As a result, the collected judgments are prone to inconsistencies, both across annotators and within the same annotator. SpAM helps address shortcomings of the absolute pairwise ratings by allowing repeated multi-wise, relative continuous similarity judgments, which produce evaluation data capturing the complexity of lexical relations in continuous semantic space.



Word1	Word2	Rating	Word1	Word2	Rating
hammer	pliers	4	platter	pear	0
pliers	cup	1	banana	capsicum	4
platter	cup	4	pear	capsicum	4
hammer	cup	1	hammer	capsicum	0
platter	pliers	1	pliers	capsicum	0
banana	cup	0	platter	capsicum	0
banana	pliers	0	cup	capsicum	0
banana	hammer	0	cup	pear	0
banana	platter	0	hammer	pear	0
pear	banana	4	hammer	platter	1
pear	pliers	0			

Figure 1
Comparison of the SpAM method with visual and lexical stimuli, and the pairwise rating approach, on a toy set of concrete real-world concepts. The 7-item sample generates 21 unique pairings of items in the pairwise rating method (example numerical ratings are given for illustrative purposes). In SpAM, placements of items express relative similarities: artefacts *pliers*, *hammer*, *platter*, *cup* are closer together than the fruit; Within the fruit group, *capsicum* is closer to *pear* than *banana*, while *pliers* and *hammer*, and *plate* and *cup*, form two smaller clusters of similar items. Images used in the diagram courtesy of the MRC Cognition and Brain Sciences Unit (University of Cambridge) and the Open Images Data set (Kuznetsova et al. 2020).

In this work, we adapt the multi-arrangement method proposed by (Kriegeskorte and Mur 2012), which uses inverse multidimensional scaling to obtain a distance matrix from multiple spatial arrangements of subsets of items within a 2D space. The participants are presented with subsets of items designed by an adaptive algorithm aimed at providing optimal evidence for the dissimilarity estimates. They are asked to drag and drop the stimuli within a circular **arena** on the computer screen, placing items perceived as similar close together and those dissimilar further apart (see Figure 1 again). The method has been shown to have high test–retest reliability (Spearman's $r = 0.93$, $p < 0.0001$) and to yield similarity data which strongly correlate with those acquired by means of the traditional pairwise similarity judgment approach (Spearman's $r = 0.89$, $p < 0.0001$) (Kriegeskorte and Mur 2012).

The first arrangement of all items within a sample provides an initial estimate of the RDM. Subsequently, the individual continues work on subsets sampled from the entire stimuli set. The adaptive subset selection algorithm elicits repeated judgments on items placed close together in the previous trial to ensure enough evidence is collected for the relative distances between the similar items and for each possible pairing (Figure 4). The process can be terminated at any time after the first arrangement onward, but an earlier termination entails a potentially noisier final RDM. The participant is instructed to use the entire space available for each consecutive arrangement. This allows them to spread out items previously clustered together, thus reducing bias from placement error. The relative inter-item distances, rather than the absolute screen distances, represent dissimilarities between the items from trial to trial. The RDM estimate is updated after each trial and the collected evidence is statistically combined to yield the final RDM (for details of the algorithm see Kriegeskorte and Mur [2012]). The thus obtained pairwise dissimilarity scores for each class are normalized by scaling each distance matrix to have a root mean square (RMS) of 1 to guarantee inter-class consistency Equation (6).

In order to adapt the multi-arrangement approach for the purposes of our task we had to address two key challenges, previously unsolved by SpAM-based methods: *scalability* and *semantic ambiguity*. So far, cognitive science research has applied SpAM to fairly small stimuli sets (≈100 items). Moreover, our preliminary tests revealed that larger samples are technically and cognitively difficult for humans. First of all, the size of the arena within which the items are arranged is restricted by the dimensions of the computer screen (Figure 4). With samples larger than 100 items the arena becomes overcrowded, which makes it difficult to distribute the items as needed. What is more, longer sessions increase participant fatigue and thus affect judgment quality and consistency. Second of all, lexical stimuli are semantically ambiguous: Without multiple sense labels, annotators consider different word senses, which results in different similarity judgments.

In the following sections, we describe a new SpAM-inspired framework that resolves both the issue of scalability and semantic ambiguity, and demonstrate how these key challenges are addressed by our proposed two-phase study design.

3.2 Two-Phase Design

The annotation process is structured as follows: first, in a *rough clustering phase* (Phase 1), our large starting sample is divided into smaller, broad classes of semantically similar and related verbs. Second, in a *spatial multi-arrangement phase* (Phase 2) the verbs placed in the classes in the previous phase are repeatedly arranged within the 2D space.

The two-phase design enables us to overcome the challenges posed by ambiguity and scale discussed in the previous section (Section 3.1). It splits the large sample into

manageable theme classes, which can be accommodated by most computer screens without negatively affecting legibility. Furthermore, the two-phase solution handles the issue of ambiguity by providing a functionality that enables annotators to copy verb labels to capture different word senses in Phase 1. The rough clustering phase ensures that each verb is presented in the context of related verbs in the arena in Phase 2, a necessary prerequisite for meaningful similarity judgments in psychology (Turner et al. 1987).[3] The sense of any given word is implied by the surrounding related words, which helps prevent discrepancies in similarity judgments between participants for ambiguous verbs. Moreover, it avoids the common issue of ambiguous low similarity scores (Milajevs and Griffiths 2016) that conflate similarity ratings of antonyms (*agree - disagree*) and completely unrelated notions (*agree - broil*), and elicits judgments between comparable concepts.

3.3 Data

In order to evaluate the scaling-up potential of our method and to enable direct comparisons with the standard pairwise similarity rating methods, we chose the 827 verbs from SimVerb (Gerz et al. 2016) as our sample (with two verbs, *tote* and *pup*, removed because of their very low frequency as verbs, producing an 825-verb final sample). The sample poses a considerable challenge due to its size, being seven times as numerous as the largest stimuli sets so far used in SpAM research, and spans a wide range of verb meaning, with each top-level VerbNet class represented by three or more member verbs.

3.4 Interface and Task Structure

The two tasks constituting our annotation design were set up on an online platform that allows users to save progress and resume annotation work after breaks as required.[4] Phase 1 and Phase 2 were set up consecutively as separate studies and participants were recruited for each individually. The guidelines for both phases were embedded in the experiment structure, available both prior to and during the task. The annotators' understanding of the instructions for each phase was tested in a short qualification task simulating the full experiment, which consisted in clustering (Phase 1) and spatially arranging (Phase 2) seven verbs. The average time spent on the qualification task was 1.5 minutes for Phase 1 and 7 minutes for Phase 2.

4. Phase 1: Rough Clustering

The goal of Phase 1 was to classify 825 English verbs into groups based on their meaning, so as to form broad (thematic) semantic classes. The guidelines instructed the annotators to group similar and related words together. While the exact number and size of the classes were left unspecified, the annotators were asked to aim for broad categories of roughly 30–50 words. Deviations from this guideline were allowed in case some smaller or larger semantically coherent groupings of verbs were identified.

The Phase 1 task interface presents the participants with a scrollable alphabetic queue of 825 verbs at the bottom of the screen and three white circles, "new category,"

3 According to Turner et al. (1987, page 46), "stimuli can only be compared in so far as they have already been categorised as identical, alike, or equivalent at some higher level of abstraction."

4 `www.meadows-research.com`.

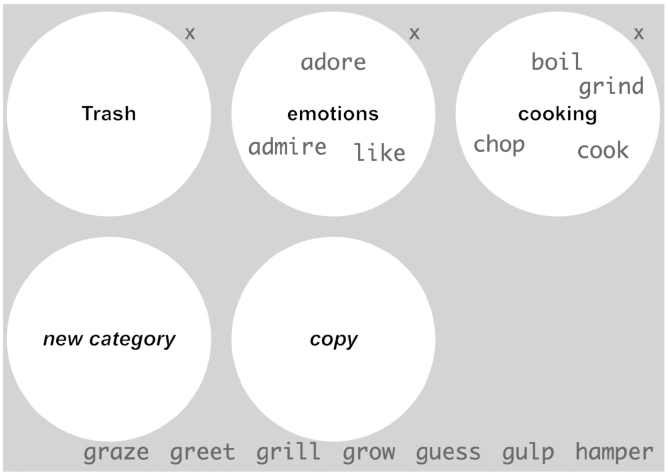


Figure 2
The rough clustering task layout (zoomed in). Verbs can be dragged onto the "new category" circle to create a new grouping, onto "copy" to create a duplicate label, or "Trash" to dispose of the unwanted duplicate.

"copy" and "trash" (Figure 2). They are instructed to drag and drop the verbs from the list one by one into the circles, creating new ones as they work through the sample. Each circle represents a semantic cluster created by the participant and serves as a container for a single grouping of similar and related verbs. If a single verb fits in more than one group, the guidelines instructed to copy the verb label (as many times as required, by dropping it onto the "copy" circle) and put each in a different circle of related verbs. This was illustrated in the annotation guidelines with the verb *draw*, which could be clustered with art-related verbs (e.g., *paint, design*) or verbs such as *pull* and *drag*. The copying functionality allowed handling of both polysemous and vague verbs.

4.1 Participants

Two native English speakers first participated in a test round of the rough clustering task. The clusters they produced showed an encouraging degree of overlap, calculated based on the B-Cubed metric (Bagga and Baldwin 1998) extended by Amigó et al. (2009) to overlapping clusters and by Jurgens and Klapaftis (2013) to fuzzy clusters, as used in related work (Jurgens and Klapaftis 2013; Majewska et al. 2018a). The B-Cubed metric, based on precision and recall, estimates the overlap between two clusterings X and Y at the item level. Let U represent the collection of items, X_i the set of clusters containing item i in clustering X, Y_i the set of clusters containing i in clustering Y. Let $j \in X_i$ and $j \in Y_i$ be an item, including i, from the set of clusters containing i in clustering X and Y, respectively. B-Cubed precision P and recall R are defined as:

$$P = \frac{1}{|U|} \sum_{i \in U} \frac{1}{|j \in X_i|} \sum_{j \in X_i} \frac{min(|X_i \cap X_j|, |Y_i \cap Y_j|)}{|X_i \cap X_j|} \quad (1)$$

$$R = \frac{1}{|U|} \sum_{i \in U} \frac{1}{|j \in Y_i|} \sum_{j \in Y_i} \frac{min(|X_i \cap X_j|, |Y_i \cap Y_j|)}{|Y_i \cap Y_j|} \quad (2)$$

Precision and Recall are combined into F-measure as follows, defined as their harmonic mean where $\alpha = 0.5$:

$$F_\alpha(P, R) = \frac{1}{\alpha(\frac{1}{P}) + (1 - \alpha)(\frac{1}{R})} \quad (3)$$

The obtained B-Cubed inter-annotator agreement (IAA) score (0.400) compares favorably to previous work on verb clustering (in Majewska et al. [2018a], B-Cubed IAA scores ranged between 0.172 and 0.338). It is also promising compared to results obtained in SemEval (Jurgens and Klapaftis 2013), where scores ranged between 0.201 and 0.483, given that cluster labels in that task were selected from a small number of fixed classes per item based on WordNet (Miller 1995).

Subsequently, a group of 10 English native speakers from the UK and the US, with a minimum undergraduate level of education, participated in the task, spending 2.4 hours on average to complete it. The number of the produced clusters ranged between 10 and 67 (27.5 on average), each with an average of 12.3–82.5 verb members.

4.2 Cluster Analysis

Before unifying the clusterings from individual annotators for our second phase (see Section 4.3), we applied network analysis to manually scrutinize the rough clustering data. The ultimate goal is to obtain an average classification where membership and size of each class is determined by the intersection of the classes from all annotators (the core), extended by additional valid member verbs on which there was partial agreement. From the entire set of all clusters (G) produced by the 10 annotators, we extracted all pairs of verbs put together in a cluster g by an annotator, that is, $v_1v_2 \in P_g$ where P_g is the set of all verb pairings in cluster g. Let P_G be the multiset of all such pairs from $\{v_1v_2 \in P_g : g \in G\}$. Each verb in the pair represents a node in the network, linked by an edge weighted according to the number of annotators clustering them together, that is, the number of occurrences in the multiset P_G, denoted as $N(P_G, v_1v_2)$. Thus, the edge weight is calculated as $w(v_1v_2) = N(P_G, v_1v_2)$.

We applied a weight-based threshold to eliminate weak ties (where $w(v_1, v_2) < 6$, i.e., there is no majority from the 10 annotators[5]) and reduce computational burden for network processing, given that the full graph had approximately 285,000 edges. We then used Cytoscape open source software (Shannon et al. 2003; Li et al. 2017) for analysis and visualization (see Figure 3).

To identify higher density areas, corresponding to groupings of similar verbs, we performed cluster analysis with a selection of graph clustering algorithms designed for detecting overlapping and non-overlapping clusters (Li et al. 2008; Wang et al. 2011,

5 We experimented with different thresholds and settled for 6 as the value representing the actual majority of annotators and a good compromise between comprehensiveness and computational efficiency: At this point, we wanted to include as much variation as possible in the graph (to also see edges weaker than those representing perfect agreement), while discarding the pairings on which annotator consensus was below the minimum majority threshold.

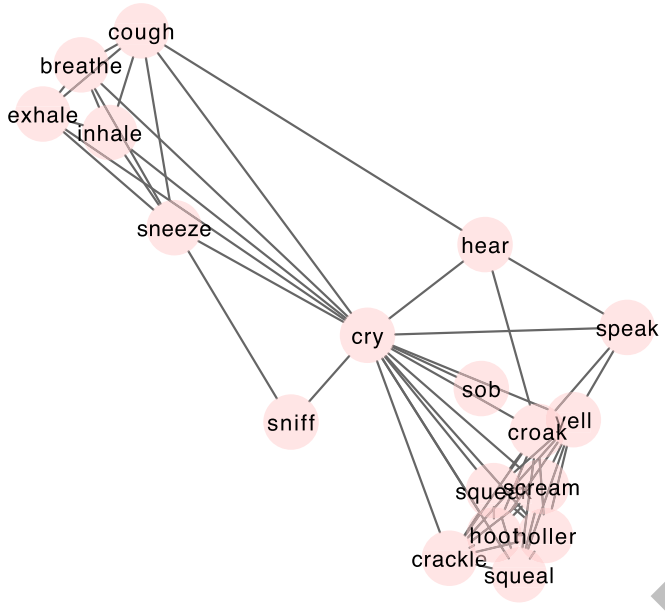


Figure 3
Visualization of a fragment of the network with the verb *cry* acting as a connector node.

2012; Nepusz, Yu, and Paccanaro 2012). Table 1 presents the results of this analysis. The labels are given for descriptive purposes alone. All four approaches identified the same largest area of high density of links, formed by the "movement" verbs (e.g., *move, fly, swim, walk*). Other large areas of interconnected nodes include, for example, "communication" verbs, verbs related to crime and violence, "negative emotions" and "cognitive" verbs (Table 1).[6] We explored the clusters with two network analysis metrics as follows: closeness centrality (Newman 2005):

$$C_c(n) = \frac{1}{avg(L(n,m))} \tag{4}$$

and betweenness centrality (Brandes 2001):

$$C_b(n) = \sum_{s \neq n \neq t} \frac{\delta_{st}(n)}{\delta_{st}} \tag{5}$$

$L(n,m)$ is the length of the shortest path between two nodes n and m, and $C_c(n)$ of n is the reciprocal of the average shortest path length. s and t are nodes different from n, δ_{st} is the number of shortest paths from s to t, and $\delta_{st}(n)$ is the number of shortest paths from s to t on which lies n (i.e., the number of paths equal to the shortest length overall).

6 We manually analyzed the clusters output by the four algorithms to identify the main areas of agreement without imposing strict overlapping membership criteria.

Table 1
Main clusters identified by *N* graph clustering algorithms in the network created from the 825-verb manual clustering data and example member verbs. Cluster labels are given for descriptive purposes.

Cluster label	Example verbs	*N*
movement	*wander, swing, fly, glide, roam*	4
communication	*persuade, command, tell, ask, say*	4
crime & violence	*beat, abduct, abuse, shoot, kill*	4
negative emotions	*offend, aggravate, enrage, disgust*	4
positive emotions	*admire, respect, adore, like, approve*	4
cognitive processes	*suppose, assume, realize, know*	4
cooking	*cook, slice, grind, stew, boil*	4
possession	*belong, accumulate, obtain, acquire*	4
physiological processes	*perspire, sweat, vomit, inhale*	4
perception	*glance, observe, stare, look*	4
destruction	*perish, demolish, decompose*	4
accomplishment	*accomplish, succeed, excel*	4
construction	*repair, fasten, mend, fit, fix*	2
sound	*hoot, roar, crackle, rattle, hum*	2
rate of change	*boost, raise, accelerate*	3

The closeness centrality measure identifies the nodes with the shortest total distance to all other nodes, that is, the prototypical member of a class. The verbs with the highest $C_c(n)$ values can be seen as representing the underlying common "theme" of the cluster. For example, verbs with the highest $C_c(n)$ score are *speak* for communication verbs, *annoy* for the "negative emotions," and *destroy* for "destruction" verbs. Betweenness centrality, on the other hand, quantifies the number of times a node acts as a bridge along the shortest path between two other nodes. Thus, it can be used to identify the verbs that act as connectors between different clusters, such as ambiguous verbs whose different senses belong to different groups. One example is *cry*, which connects the "sound" cluster comprising verbs such as *scream, holler, squeal*, and the "physiological processes" verbs, like *breathe, cough, sneeze* (Figure 3). Identifying prototypical members and verbs acting as inter-class links can be especially useful for creation of a comprehensive verb resource.

In the next section (4.3), we outline the protocol for selection of classes for Phase 2. Although we used cluster analysis primarily as an exploratory means, allowing us to examine annotator decision patterns on the rough clustering task and the emerging semantic categories, it also served as a preliminary step that informed our decisions relative to Phase 2 class selection. We kept the minimum majority threshold of 6 annotators, as high enough to ensure semantically coherent classes and discard noise, but also comprehensive enough to leave room for some degree of variation in clustering decisions, reflecting their inherent flexibility (i.e., there is no single perfect clustering solution). The chosen threshold also guaranteed the desired granularity and nature of resultant classes: While higher thresholds produced many narrow clusters of synonyms or close-synonyms (e.g., *join, connect, associate*, or *forbid, deny, disallow, refuse*) lowering the cutoff value yielded broader semantic classes including the less prototypical members (on which there was partial agreement), which was the intended output of Phase 1.

4.3 Class Selection for Phase 2

The class selection protocol that determined the classes to be used in Phase 2 was the following. Clusters obtained from the verb pairings on which any 6+ participants (the majority) agreed were used as a starting point and determined the broad semantics of the classes (e.g., "perception," "movement," "communication"). Post-processing was limited to (1) merging smaller semantically related clusters to produce large, all-encompassing classes based on semantic relatedness of class members, and (2) populating the thus created sets with the verbs missing from the majority classes based on their relatedness to the already-classified members. These lower-agreement verbs were reviewed and manually added to related classes by one of the authors. Clusterability of Phase 1 verbs (i.e., the SimVerb sample) was guaranteed by balanced sampling from across different VerbNet classes (Gerz et al. 2016). Ambiguous verbs could be placed in several classes with semantically related members, by means of the copying functionality described above (Section 4). Six out of 10 annotators used label copying to capture ambiguity and 234 different verbs (out of 825) were assigned to more than one class. The average pairwise percent agreement on ambiguity decisions (i.e., a binary choice whether a verb is ambiguous or not) was 91.1%. The final number of produced classes was 17.

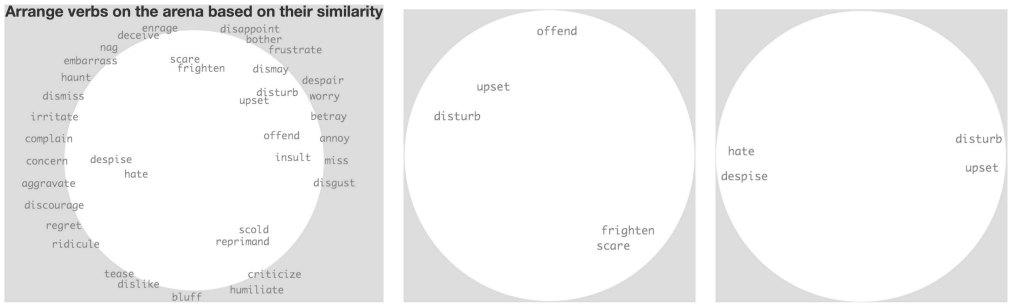
The main clusters identified by the clustering algorithms in Section 4.2 overlap very closely with the final classes used for Phase 2. All the semantic areas (see descriptive labels in Table 1) recognized through network clustering are represented in Phase 2. The only discrepancies lie in the granularities, for instance, while the clustering algorithms unify all verbs related to motion, our class selection protocol produced two different classes split along a line mirroring the intransitive/transitive distinction, that is, movement verbs where the intransitive sense is predominant (*crawl, dash, fly*), and transitive verbs describing causing something to move (*drag, tow, fling*). Similarly, the broad cluster related to "crime and violence" is split into two Phase 2 classes: verbs of physical contact (*beat, kill*) and verbs describing criminal acts and legal terms (*kidnap, abuse*). Among the smaller areas of higher density identified by cluster analysis that are not represented as separate Phase 2 classes were narrow semantic groupings, usually of synonyms or close-synonyms, such as (*imitate, mimic, impersonate*), (*crave, yearn, want*), or (*help, assist, aid, rescue, protect*), as well as few examples of small clusters based on association (e.g., *embarrass, worry, weep, regret, sprain*, or *stop, withdraw, unload*).

5. Phase 2: Multi-Arrangement

In Phase 2, participants performed the spatial multi-arrangement task. Each of the 17 verb classes output by Phase 1 was individually displayed on the computer screen, in random order, around a *circular arena* (Figure 4). The participants were instructed to arrange verbs based on similarity of their meaning, dragging and dropping the labels one by one onto the circle, so that similar words ended up closer together and less similar ones further apart, while the relative positions and distances between them reflected the degree of similarity.

5.1 Participants

The minimum number of annotators to work on each class was set to 10. We asked each annotator to arrange at least 3 classes, presented in random order, and permitted rest breaks between classes. Annotator recruitment continued until the minimum number



(a) trial 1 (b) trial 2 (c) trial 3

Figure 4
Consecutive Phase 2 trials on a single class (zoomed in). In the first trial (a), the whole class is presented around the arena and words are dragged and dropped one by one, with their relative distances representing the degree of similarity. Words put closer together in the first trial are subsampled in the subsequent trials (b and c), and arranged again in a less crowded space, which ensures a higher signal-to-noise ratio (i.e., since annotators use the whole space available in each trial, the items are more spread out and placement error is a smaller proportion of the dissimilarity signal). The RDM estimate is updated after each trial and the evidence from consecutive 2D arrangements is combined to produce the final pairwise dissimilarities for the entire word set.

of annotators per class was satisfied. Overall, 40 native English speakers from the UK and the US, with a minimum undergraduate level of education, took part in the multi-arrangement task, producing ultimately a total of 314,137 individual pairwise scores. For each class and annotator, we recorded the time spent on each individual trial (i.e., each consecutive arrangement of subsets of a single class). The average total time spent completing the task for all 17 classes was 735 minutes, with the average time spent on a single task (equivalent to arranging one class) ranging from 15.5 minutes (for the smallest class) to 60 minutes (for the largest class).

5.2 Post-Processing

We applied the following steps to ensure high quality of the resultant data. First, we discarded annotations where word placements were executed too quickly in the first arrangement of each class (i.e., where the average time spent on dragging and dropping a single verb label was less than 1 second). This heuristic allowed us to quickly identify and eliminate rogue annotators: Our trial experiments showed that users spend much longer on the first arrangement than the consecutive ones for that class, given that it is the first time they see a given word sample and extra time is needed to familiarize oneself with the set. Extremely short times spent on word placements in the first trial were therefore a clear indicator of low-effort responses. Second, for each class we excluded outlier annotators for whom the average pairwise Spearman correlation of arena distances with distances from all other annotators was more than one standard deviation below the mean of all such averages. The same criterion was adopted as the acceptability threshold in the creation of SimLex (Hill, Reichart, and Korhonen 2015).

For each class, we computed the average of the Euclidean distances from all accepted annotators for each verb pair and obtained an average RDM (as shown in Figure 5). The averaged pairwise distances (= dissimilarity scores) in each class were

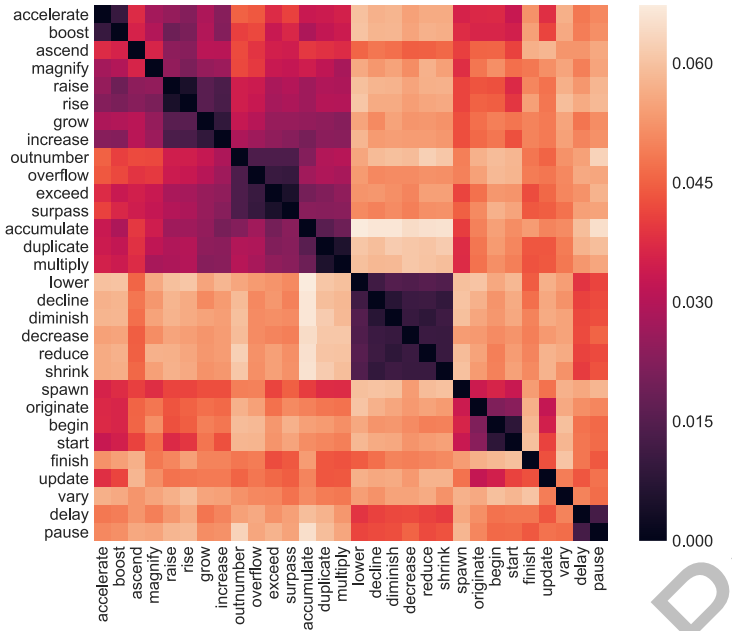


Figure 5
Average ordered dissimilarity matrix for one of the verb classes (dark-to-light color scale for small-to-large dissimilarities), with dark areas corresponding to clusters of similar verbs (e.g., *lower, decline, diminish, decrease, reduce, shrink*).

then scaled to have a RMS equal to 1, as done in previous work using inverse MDS (Kriegeskorte and Mur 2012; Mur et al. 2013), to ensure inter-class consistency. For each class, the scaled distances $d'_i, \ldots, d'_N$ were thus obtained for N pairs by dividing each pairwise distance d_i by the square root of the mean of N distances squared (d_i^2):

$$d'_i = \frac{d_i}{\sqrt{\frac{1}{N}\sum_{i=1}^{N} d_i^2}} \tag{6}$$

The final data set, SpA-Verb, collates the thus obtained scaled averaged pairwise distances for each of the 17 verb classes, comprising (dis)similarity scores for the total of 29,721 unique verb pairs.

6. Inter-Annotator Agreement

We measure inter-annotator agreement in Phase 2 based on Spearman's rank correlation coefficient (ρ): For each class, we calculate the average correlation of an individual annotator with the average of all other annotators (*mean* Spearman's ρ) (Hill, Reichart,

Table 2
IAA (mean Spearman's ρ) by verb class (ρ^A) of N verbs and N^A unique verb pairs and set of N^{SV} verb pairs shared with SimVerb in that class (ρ^{SV}), and examples of verbs in each class.

#	Example verbs	N	N^A	ρ^A	ρ^{SV}	N^{SV}
1	*beat, punch, smash, slap*	48	1128	0.53	0.50	92
2	*accuse, condemn, forbid, blame*	80	3160	0.27	0.61	134
3	*accelerate, decrease, shrink, increase*	30	435	0.64	0.71	38
4	*achieve, aim, tackle, accomplish*	57	1596	0.34	0.41	98
5	*acquire, have, keep, borrow*	47	1081	0.40	0.50	102
6	*dismay, frustrate, upset, irritate*	38	703	0.24	0.35	73
7	*ask, confess, discuss, inquire*	85	3570	0.27	0.30	194
8	*approve, desire, prefer, respect*	23	253	0.41	0.33	31
9	*calculate, analyze, predict, guess*	75	2775	0.31	0.51	159
10	*climb, jump, roam, slide*	100	4950	0.26	0.48	253
11	*bake, grate, slice, broil*	53	1378	0.52	0.66	85
12	*cough, gulp, inhale, sniff*	56	1540	0.29	0.69	52
13	*chirp, hoot, roar, whistle*	34	561	0.53	0.65	51
14	*build, fasten, mend, restore*	62	1891	0.24	0.46	89
15	*drag, fling, haul, toss*	87	3741	0.19	0.36	129
16	*demolish, erode, wreck, disintegrate*	27	351	0.46	0.62	51
17	*glance, observe, perceive, look*	41	820	0.43	0.71	76

and Korhonen 2015; Gerz et al. 2016) (see Table 2).[7] We do not calculate IAA over the entire data set as different groups of annotators worked on different classes.

The characteristic flexibility offered by our drag-and-drop interface, where similarity judgments expressed through word placements produce fine-grained pairwise similarity scores differing by fractions, based on the words' relative positions in the circular space, leaves a lot of room for divergence in scores across annotators compared with discrete ordinal rating scales. Nonetheless, the resultant IAA scores (ρ^A) are promising. In particular, they compare favorably with inter-subject correlations reported in cognitive neuroscience research for spatial multiple arrangements of concrete visual stimuli (real-world objects like in Figure 1): for example, Mur et al. (2013) report an average total pairwise inter-subject Spearman's ρ correlation of 0.32, Cichy et al. (2019) report scores in the range of approximately 0.12–0.21 ($p < 0.001$).

Effect of Class Size and "Clusterability". The main factor affecting the difficulty of the task was class size, as reflected in the differences in agreement scores reported in Table 2: We observe negative correlation between inter-annotator agreement and the number of verbs in a class (Spearman's $\rho = -0.67$).

7 Rank correlation metrics like Spearman's ρ, which measure the correlation between rankings of (dis)similarity scores, rather than the absolute scores, are recommended for comparing RDMs (Nili et al. 2014). Given the free nature of the arrangement task, some degree of inter-subject variability in the usage of the arrangement space and the raw inter-item distances in each trial is expected, regardless of the degree of consensus on the relative similarities of word pairs in the arena, which is of interest in this study; Therefore, comparing rank orders of the dissimilarities, rather than the variance of their raw values, provides an informative measure of agreement on similarity judgments. Note also that there is no fixed relationship between screen distance and dissimilarity that holds across trials: Because participants "zoom in" on items previously clustered together by spreading them out upon successive trials (Figure 4), it is the relative screen distances (i.e., screen distance ratios) that reflect the relative dissimilarities on each trial.

However, the semantics of the classes seem to play a role as well: For instance, the agreement on the largest 100-verb class of movement verbs (*#10*) is higher than could be expected based on its size alone ($\rho = 0.26$), compared to the smaller Class 15, where the agreement is the lowest. We observe class "clusterability" to be an important factor, namely, the availability of underlying structure within a bigger class, where words cluster into balanced sub-groups with clearly defined shared semantics.[8] For instance, many movement verbs have well-defined, concrete meanings, clusterable into smaller groupings, for instance, based on the medium in which the movement takes place (on land [*walk, crawl*], in water [*swim, dive*], in the air [*glide, fly*]). The lowest-agreement Class 15, comprising verbs of motion undergone by the verb's *object*, such as *add, dip, flush, spread*, is more heterogeneous, that is, there is more variety in verbs' semantic properties and the dimensions along which the class members differ are less clearly defined, which means there are many equally valid arrangements possible. As indicated in annotator feedback, this characteristic made it harder to identify the potential groupings and sub-categories into which words could be classified; Consequently, their relative positions varied by participant.

In order to examine the impact of sample size on IAA, we carried out a follow-up experiment on the lowest-IAA class (*#15*). Our goal was to verify if higher IAA scores can be obtained on the same verb pairs split into smaller samples. Five new annotators subsequently arranged three equal 29-word subsets randomly sampled from the entire 87-word class (*#15*), each working on the three subsets one by one, with breaks in between. The IAA computed for the smaller sets proved lower than in the full-class (87-word) setting, producing an average across the three subsets of $\rho = 0.098$, compared with $\rho = 0.19$ on the full class. This analysis suggests that although big samples are generally more challenging, the task's difficulty very much lies in the verbs included in the sample, and this class proves particularly difficult due to its heterogeneity. While annotators consistently place similar verbs close together (e.g., *smear - smudge*, *seize - snatch*), there is greater variability in the distances between the less similar words. In the follow-up study, this issue was further aggravated by randomly splitting the coherent big set and potentially separating verbs naturally clusterable together. These findings also indicate that simply reducing the number of words to be arranged in the arena does not guarantee higher agreement: Being presented with a semantically clusterable bigger set of words (like those produced in Phase 1) may be preferable to imposing an arbitrary limit on class size. The greater difficulty of some verb sets resulted in inter-annotator agreement scores for some classes showing low positive correlation. Therefore, we recommend that evaluation of representation models best be focused on classes with higher inter-annotator agreement and consequently clearer semantics.

SpAM vs. Pairwise Ratings. Because our verb sample is the same as SimVerb's, we can directly compare IAA recorded for each class with the IAA on the verb pairs in that class also occurring in SimVerb. The results of this analysis are shown in ρ^{SV} of Table 2. In what follows, we use this comparison to highlight the main similarities and differences between the output of our Phase 2 method and the pairwise rating approach used with

8 In the preliminary trial experiments, annotators reported that the availability of words that naturally group together within a bigger class based on some criterion (e.g., animal sounds, human sounds) significantly facilitated the spatial arrangement task, in contrast to having small but randomly sampled sets of words to arrange, with many semantic "isolates," that is, words which were dissimilar from all others.

SimVerb, which due to its scale and sole focus on verbs is the most similar resource currently available.

Even though the two resources share the starting verb sample, the number of overlap pairs in each class (as shown in column N^{SV} of Table 2) is reduced due to the differences between the annotation paradigms used in SimVerb and SpA-Verb. In SimVerb, pairs that end up in the final data set were selected to cover different degrees of relatedness, including completely unassociated pairs, whereas our rough clustering phase (Phase 1) divides the sample into classes based on relatedness, therefore the possible pairwise combinations of verbs are limited to related verbs. These discrepancies are reflected in the different score distributions in both data sets, as illustrated in Figure 6. SimVerb scores show a peak at the 0–1 unrelated end of the distribution: The most numerous are the easy to annotate unrelated verb pairs, which are filtered in Phase 1 of our approach (see Section 8.6 for a discussion of the implications of these differences for intrinsic evaluation).

The sets of shared pairs are on average over one order of magnitude smaller than our respective complete classes. What is more, the overlap pairs are more spread out in terms of degree of similarity compared to the complete classes, which comprise very many nearly equidistant verb pairs. Crucially, for each cluster of similar verbs in a Phase 2 arena, our data set includes all the possible unique pairwise combinations; consequently, many scores differ by small amounts. Only some of those pairs appear in SimVerb—for example, out of Class 9 pairs *decide-choose*, *decide-select*, *decide-elect*, *decide-pick*, only the first one is present. These highlighted differences explain the lower correlation scores obtained on most of the entire classes compared to overlap pairs (ρ^{A} vs. ρ^{SV}), which, in turn, reflect the greater difficulty in making subtle distinctions between very many semantically related words appearing in the same arena in our spatial arrangement task.[9] While many concurrent decisions make judgments in the arena harder, the resultant scores are more thorough, offering a more comprehensive coverage of a given semantic domain (i.e., a complete pairwise similarity matrix for each arena).

Even though SimVerb and SpA-Verb are produced by different paradigms, there is a reasonable level of correlation between the two resources on all the 1,682 shared pairs: $\rho = 0.62$. Crucially, by eliciting simultaneous judgments on multiple lexical items our approach significantly speeds up the data collection process. As an example, with our SpAM-based method 60 minutes of work of a single annotator produces pairwise similarity scores for 4,950 unique verb pairs. In the pairwise rating approach used for SimVerb, the same number of similarity judgments would take a single rater over 8 hours to record (requiring approximately 8 minutes to complete 79 questions by a single participant [Gerz et al. 2016]. Our two-phase design and the modular nature of the annotation pipeline make it particularly suitable for crowd-sourcing. In the following section, we explore in detail the properties of each of the two phases, highlighting their benefits beyond what is offered by pairwise rating data sets.

7. Phase 1 vs. Phase 2 Analysis

As a consequence of our annotation design, the nature of semantic information captured in the two phases changes from broad thematic similarity in Phase 1, to similarity of

9 The ρ^{SV} scores computed for the overlap pairs are promising compared to the $\rho = 0.612$ correlation reported for SimVerb (Pilehvar et al. 2018), especially in light of the fact that the easy cases of pairs of very disparate verbs (split into different classes in our Phase 1) are not included in our results.

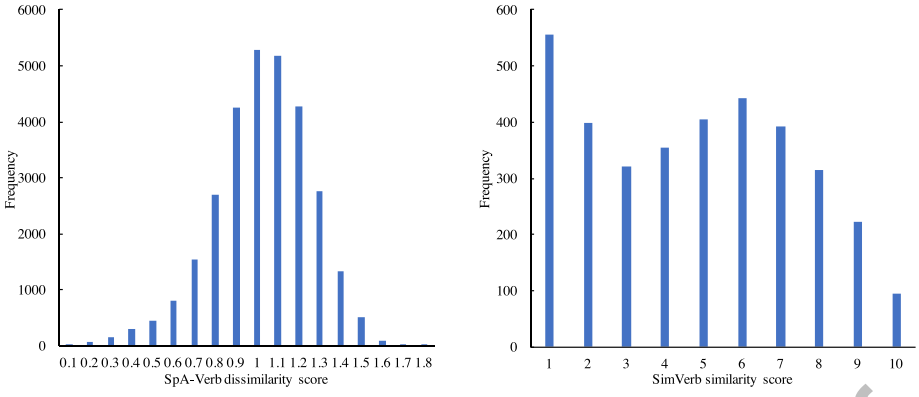


Figure 6
Score distribution for SpA-Verb (dissimilarities are scaled Euclidean distances (Equation 6)) and SimVerb (ratings on a 0–10 interval) in terms of frequency of each score interval (i.e., the number of individual ratings belonging to a given score interval in each data set). Each score interval label gives the upper bound.

meaning (understood in terms of shared semantic properties) in Phase 2, represented as varying distances between related words. In order to examine how our collected human judgments reflect these different assumptions, we carry out a comparative analysis of our data with two lexical resources, FrameNet (Baker, Fillmore, and Lowe 1998) and VerbNet (Kipper Schuler 2005). Next, we investigate whether the perceived semantic distances between related words recorded in the spatial arrangement task reflect finer-grained lexical relations like synonymy or hypernymy by comparing average similarity scores collected for verb pairs grouped according to the WordNet relations in which they participate.

Comparison with FrameNet and VerbNet. Based on Frame Semantics theory (Fillmore 1976, 1977, 1982), FrameNet includes over 1220 semantic frames, that is, descriptions of types of events, relations, or entities, and their participants. Each frame records the semantic type of a given predicate, usually a verb, and the semantic roles and syntactic realizations of its arguments. The lexical units associated with it share similar semantics and argument structures. On the other hand, VerbNet extends Levin (1993)'s taxonomy and groups verbs into classes based on shared semantic and syntactic properties. Each class is described by thematic roles, selectional restrictions on the arguments, and frames including a syntactic description and a semantic representation.

For each phase of our design, we compute the overlap between the resultant human classes/clusters and classes/frames extracted from both of these resources. For Phase 1, we compute the B-Cubed metric directly between our 17 classes and FrameNet parent frames (i.e., one level up in the hierarchy from fine-grained FrameNet frames) and top-level VerbNet classes, extracted for the 825 verbs in our sample.[10] For Phase 2, for three selected verb classes (the largest (#*10*), the lowest IAA (#*15*), and highest

10 We selected the hierarchy levels in FrameNet and VerbNet for comparison with our Phase 1 classes and Phase 2 clusters aiming to compare similar granularity levels, comparing broader Phase 1 classes with higher levels of each hierarchy. However, there is still a major difference in the number of classes in our Phase 1 (17) and FrameNet parent frames (128) and VerbNet top-level classes (101) (for the shared verbs).

Table 3
Comparison of Phase 1 (P1) classes and clusters extracted from Phase 2 (P2) distance matrices for classes 10, 15, and 3 against FrameNet fine-grained frames (**frame**) and parent frames (**parent**), and VerbNet top-level (**top**) and first-level (**1st**) classes. All scores are B-Cubed F-scores, measuring the overlap between P1 classes/P2 clusters and the FrameNet and VerbNet classes for the shared verbs. For Phase 2, *optimal* columns show scores obtained for the optimal clustering solution in terms of F1 score, determined iteratively over values of $k = \{1, ..., N\}$, where N is the size of each class (#10–100, #15–87, #3–30); *gold* columns show scores for clustering solutions with $k = K_{gold}$, where K_{gold} is the number of classes in FrameNet or VerbNet in which the shared verbs participate. We do not report *gold* values where the number of gold classes was larger than the number of verbs in a given P2 sample ($K_{gold} > N$), due to multiple class membership of individual verbs in *gold* resources.

	FrameNet				**VerbNet**			
	parent		frame		top		1st	
P1	0.247		–		0.302		–	
P2 $k =$	*optimal*	*gold*	*optimal*	*gold*	*optimal*	*gold*	*optimal*	*gold*
#10	0.527	0.247	0.407	–	0.666	0.259	0.481	0.337
#15	0.470	0.289	0.448	–	0.407	0.324	0.449	0.388
#3	0.546	0.501	0.578	–	0.642	0.566	0.618	0.616

IAA class (#3)) we first extract K_{Gold} FrameNet frames and parent frames, and K_{Gold} top and first level VerbNet classes (e.g., 17.1). K_{Gold} is the number of frames or classes in which the verbs in a given Phase 2 sample (#10, #15, #3) participate in these resources (see Table 2 for examples of verbs in each). We then apply hierarchical agglomerative clustering with average linkage (Day and Edelsbrunner 1984) on top of the distance matrices produced by Phase 2. We calculate B-Cubed for the optimal clustering solution, defined in terms of highest value of F1 score, and for $k = K_{Gold}$, reported in Table 3.

The limited degree of overlap between Phase 1 and the two resources is understandable due to the different granularity: Phase 1 includes only 17 classes, compared to 128 FrameNet parent frames[11] and 101 VerbNet top-level classes in which the 825 verbs in our sample participate. For Phase 2, the fact that the two resources include overlapping classes, while the clusters extracted from our distance matrices are exclusive, negatively affects the overlap scores. However, the encouraging B-Cubed results against VerbNet classes (> 0.6 for classes #10 and #3) suggest our arena-based approach allows annotators to intuitively differentiate between degrees of overlap in verbs' properties and create, by deliberate word placements, clusters of similar verbs within a broader related set that reflect some of the fine-grained class divisions in the expert-created lexicon. This is also observable in the differences in pairwise scores (and the growing distances as we move up the hierarchy) collected for verbs (a) belonging to the same low-level VerbNet subclass (17.1-1-1: *throw-toss*, dissimilarity $d = 0.273$), (b) verbs in a class-subclass relation (17.1-1 *toss* - 17.1-1-1 *fling*, $d = 0.350$), (c) verbs sharing the same first-level class (11.4: *tow* - *haul*, $d = 0.421$), or (d) the same top-level class (11.4 *tow* - 11.1 *take*, $d = 0.584$), or (e) belonging to different top-level classes (11 *tow* - 17 *chuck*, $d = 0.943$).

11 Further analysis could also explore indirect inheritance.

Table 4
Average SpA-Verb dissimilarities, SimVerb similarity ratings and USF free association scores across shared pairs representing four semantic relations (extracted from WordNet): synonymy, hyper/hyponymy, cohyponymy, antonymy. Score ranges represent the actual interval of scores in each source (*SpA-Verb scores are based on Euclidean distances scaled to have an RMS of 1 (Equation 6) to guarantee inter-class consistency, as detailed in Section 3.1. **SimVerb scores were originally collected as 0–6 ratings and scaled linearly to the 0–10 interval by Gerz et al. (2016)).

	SpA-Verb	**SimVerb**	**USF**
Synonymy	0.482	6.79	0.190
Hyper/hyponymy	0.593	4.00	0.120
Cohyponymy	0.686	2.79	0.060
Antonymy	1.019	0.54	0.154
Score range	0.065–1.720*	0–10**	0–1

Spatial Similarity vs. WordNet Relations. Phase 1 produces classes encompassing a range of finer-grained lexical relations within related words, including synonymy (*try-attempt*), hyper/hyponymy (*think-rationalize*), cohyponymy (*suck-sip*), or antonymy (*appear-disappear*). In Phase 2, annotators differentiate between these relations, deciding on relative semantic distances between words participating in them. The collected distance matrices include pairwise distances between all possible pairings of items within a class, and hence encode all pairwise relations within that set. This allows us to zoom in on a particular relation type and see how it is reflected in the pairwise scores for word pairs which exemplify it.

To illustrate this, we compute average dissimilarity scores for pairs exhibiting the four relations (extracted from WordNet) in our data set and compare them to average SimVerb similarity scores and USF association scores for the same pairs (Table 4). We see smallest dissimilarity scores for synonyms, which increase through hyper/hyponyms and cohyponyms as the degree of association decreases. Antonyms are furthest apart, despite their relatively high (compared to synonyms) USF association score. This supports our hypothesis that the Phase 2 multi-arrangement task set up allows annotators to differentiate between similarity and association, as well as a range of fine-grained lexical-semantic relations.

These findings are noteworthy in light of the different strategies used by the two paradigms that produced SimVerb and SpA-Verb (each characterized by a different means of expressing similarity judgments) to handle the relatedness/similarity distinction and antonymy. In SimVerb, the guidelines instruct annotators to assign low numerical scores both to antonyms (*stay - leave*) and to related but non-similar words, for example, *walk - crawl* (Gerz et al. 2016). In our approach, this is handled by means of the two-phase design: First, similar and related verbs are grouped together to form broad semantic classes. Then, fine-grained similarity judgments are made among already related verbs. As emphasized in Section 3.2, this avoids conflating scores for unrelated and antonymous pairs: The former are split into distinct Phase 1 classes, which reserves low similarity scores (= large distances in the arena) for the latter. This is confirmed by manual inspection of the outputs of both phases: In Phase 1, we find antonymous words in the same broad groups, based on their relatedness (e.g., antonymous pairs

stay and *leave*, and *lose* and *gain* end up clustered together),[12] while in Phase 2, antonyms are placed far apart in the arena: Out of 67 antonymy pairs shared with SimVerb (i.e., labeled ANTONYMS in SimVerb), only 2 are placed closer in the arena (*inhale - exhale* and *sink - swim*). This is also illustrated by the RDM in Figure 5, where separate clusters (dark areas) are formed by verbs such as *raise, rise, grow* and *diminish, decline, lower*, whereas *finish* is kept separate from *begin* and *start*.

SpAM for Graded Multi-Way Lexical Relations. Despite the parallels in the treatment of semantic relations in SpA-Verb and SimVerb, comparative analyses shed light on some important differences between judgments yielded by our SpAM method and pairwise rating-based methods, revealing potential benefits not offered by pairwise data sets. As all verbs are simultaneously judged in the context of all other related verbs, not only pairwise but also multi-way relations can be captured, reminiscent of lexical taxonomies. Degrees of similarity can be recorded in a meaningful way and adjusted in the presence of another word, distinguishing between dissimilar unrelated words and words which, despite their lack of similarity to the target word, nonetheless stand in some lexical-semantic relation to it. Such relations are exemplified, for instance, by lexical triplets (e.g., *try-succeed-fail*), where the first element expresses the necessary presupposition for the pair of complementaries (i.e., words which divide some conceptual domain into two mutually exclusive parts) (Cruse 1986). According to Cruse (1986), the binary relation between satisfactives *try* and *succeed* (an attempt vs. successful performance) is a weak form of oppositeness, while *succeed-fail* present a strong oppositeness.

Our design allows capturing these three-way relations simultaneously, grading similarity and oppositeness: Synonyms *try-attempt* receive a 0.283 score, satisfactives *try-succeed* and *attempt-succeed* 0.891 and 0.861, respectively, *try-fail* 0.960 and antonyms *succeed-fail* 1.063. To compare, in SimVerb, where pairs are judged independently, *attempt-succeed* receive a 2.16 score on the 0–10 scale. This score is lower than those for dissimilar and unrelated pairs such as *perish-sob* and *blur-rush* (2.49), so the information about *attempt-succeed* standing in some meaningful relation as opposed to the unrelated pairs is not captured. There is no consensus on what the best treatment of such cases is, but distinguishing between weak oppositeness, strong oppositeness and unrelatedness and capturing meaningful degrees of dissimilarity (e.g., "*attempt-succeed* is less dissimilar than *perish-sob*") may be beneficial for reconstruction of complex semantic hierarchies and bottom-up creation of lexical taxonomies which go beyond pairwise similarity.

Similar comparisons can be made about capturing troponymy. In WordNet, *jump* forms a synonym set (synset) with *leap, bound, spring*, and their troponyms include *hop, skip*, and *bounce*, which describe a specific manner of jumping. The troponymy relation is reflected in small distances in the arena: for example, *spring-bounce* 0.373 and *jump-skip* 0.277. In SimVerb, *spring-bounce* have a high 8.80 rating, but *jump-skip* receive a much lower 5.48 score, despite holding an analogous troponymy relation. This is a score equal to the similarity rating of *embarrass-blush*, which are strongly associated, but dissimilar. Meanwhile *jump* and *skip* display a high degree of semantic overlap (i.e., describe a similar kind of motion). The availability of scores for all possible pairings allows us to trace and reconstruct semantic and taxonomic links like those in WordNet: For

12 However, there are exceptions: Positive (e.g., *love*) and negative (e.g., *hate*) emotion verbs form two different classes. There are also separate classes of "construction" and "destruction" verbs. See Table 2.

Table 5
Examples of fine-grained DBSCAN clusters extracted from dissimilarity matrices of two classes, #7 (left) and #17 (right).

{ *accept, agree, approve, concur* }	{ *blur, disappear, fade, vanish* }
{ *ask, inquire, request* }	{ *differ, vary* }
{ *advise, clarify, educate, explain, inform, teach* }	{*glance, glare, look, observe, perceive, see, squint, stare, watch*}
{ *comfort, console, protect, soothe* }	{ *flash, glow, shine* }
{ *collaborate, cooperate* }	{ *discover, find* }
{ *depend, rely* }	{ *hunt, search, seek* }
{ *debate, disagree, protest* }	{ *hear, listen* }
{ *say, speak, talk* }	{ *happen, result* }
{ *reply, respond* }	{*disguise, imitate, impersonate, mimic, portray, pretend*}

the "choose, select" WordNet synset, synonymous *choose-select* are very close together (0.121), close but slightly further away from their direct hypernym *decide* (*choose-decide* 0.283, *select-decide* 0.216), and still further away from their troponym *elect* (*elect-choose* 0.544, *elect-select* 0.512). The distance grows slightly with inherited hypernymy across three levels (*elect-decide* 0.592) and co-hyponymy (*elect-pick* 0.556). In SimVerb, *elect-choose* receive score 8.47, but *elect-select* 5.15, despite standing in analogous relations. Such discrepancies in scores for similar relations may be a consequence of judging pairs in isolation in SimVerb: When simultaneously presented with all verbs belonging to the "jumping" or "choosing" domain (in a given sample), it should be easier to record consistent similarity judgments across relations of the same kind and degree (e.g., troponymy), which is a considerable benefit of the proposed SpAM approach.

SpAM for In-depth Exploration of Semantic Domains. The availability of complete distance matrices for each Phase 1 class enables clustering analyses aimed at producing fine-grained classes within each semantic domain. Table 5 shows examples of narrow semantic clusters output by DBSCAN algorithm (Ester et al. 1996)[13] run on top of two of the Phase 2 distance matrices. Such automatically extracted clusters, after manual review, can be used in future work to produce evaluation classes allowing for testing the models' capacity to create fine-grained semantic classifications and taxonomies automatically.

The complete RDMs from Phase 2 permit in-depth analyses of the resultant nets of semantic relations. In order to understand better what information is being captured and what underlying features and dimensions inform human similarity judgments, we applied Principal Coordinates Analysis (PCoA) (Gower 1966) to each distance matrix to examine the dimensions and meaning components characterizing the semantic space in question (Gärdenfors 2004). Figure 7 provides an example visualization of the main axes (PC1, PC2) for Class 3. The first dimension (PC1), which explains 50% of the variance, roughly corresponds to word polarity: We note positive rate of change verbs clustered on the negative side of the PC1 axis (e.g., *accelerate, increase, raise* and *exceed, surpass, overflow*), less strongly polarized verbs closer to the middle (e.g., *update, vary*), and the

13 We selected DBSCAN as it does not require specifying the value of k upfront and thus avoids explicitly imposing a predetermined cluster granularity.

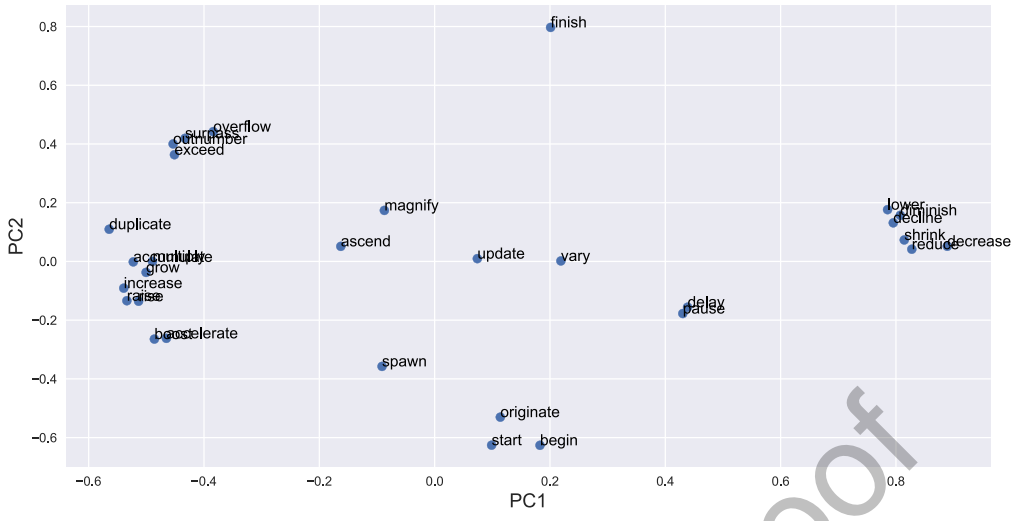


Figure 7
Visualization of PCoA applied to the Phase 2 distance matrix for rate of change verbs (Class 3).

negative rate of change verbs on the far right (e.g., *lower, decline, shrink*). Whereas the second axis (PC2) constitutes the dimension of difference for verbs expressing inception and termination, with *start, begin, originate*, and *finish* occupying its opposing poles.

Similar analyses can be used on all Phase 2 matrices to identify the most salient semantic features for each class and to gain a deeper understanding of the implicit meaning dimensions underlying human similarity judgments, and by extension, the organization of human lexical semantics and the representations constituting a conceptual space (Gärdenfors 2004; Hollis and Westbury 2016).

8. Evaluation with Representation Learning Architectures

In order to fully analyze the properties of our two-phase evaluation data set creation method, and its potential as an evaluation resource, we evaluate a representative selection of state-of-the-art representation models on two tasks, corresponding to the two phases of our design: (1) clustering, using Phase 1 classes as gold truth, and (2) word similarity, using pairwise scores from the entire SpA-Verb (29,721 pairs) and the thresholded subset (SpA-Verb-THR), including 10,371 pairs from the classes with Spearman's IAA ≥ 0.3, as well as chosen subsets with different semantic characteristics. Several different reference scales have been proposed for the interpretation of the Spearman's correlation coefficient in terms of descriptors such as "strong," "moderate," or "weak" (Chan 2003; Akoglu 2018; Schober, Boer, and Schwarte 2018), and it has been noted (Schober, Boer, and Schwarte 2018) that the range of values being assessed should be considered in the interpretation (i.e., a wider range of values tends to show a higher correlation than a smaller range, as is the case for our similarity data, see Figure 6). We choose $\rho = 0.3$ as a confidence threshold in light of these considerations and exclude the classes where IAA results show low positive correlation from the thresholded data set. This subset comprises data from 10 of the 17 classes (see Table 2). In our analyses, we also compare model performance on the subset of pairs from SimVerb-3500 within our data set (1,682 pairs) and the original SimVerb-3500 ratings. The selected architectures represent different modeling assumptions, data requirements, and underlying

methodologies, which we briefly describe below. For all models, d refers to the embedding dimensionality, and ws is the window size in case of bag-of-words (BOW) contexts.

8.1 Representation Models

Included in our selection is an unsupervised model that learns solely from distributional information in large text corpora, the skip-gram with negative sampling (SGNS) (Mikolov et al. 2013b) with BOW contexts trained on the English preprocessed Polyglot Wikipedia (Al-Rfou, Perozzi, and Skiena 2013) by Levy and Goldberg (2014) (SGNS-BOW2; $d = 300$ and $ws = 2$ as in prior work).[14] We also include models using subword-level information. The first is an extension of the original CBOW model (Mikolov et al. 2013b) (CBOW-CC) with position weights and subword information, introduced by Grave et al. (2018). Before taking the sum of context words, each word vector is element-wise multiplied by a position dependent vector as done in Mnih and Kavukcuoglu (2013). Word vectors are sums of their constituent n-grams as in Bojanowski et al. (2017) and Mikolov et al. (2018). The method is trained on deduplicated and tokenized English Common Crawl corpus.[15]

We also experiment with a more recent representation model that computes dynamic word representations conditioned on the surrounding word context (Peters et al. 2018) (ELMo-Static, $d = 300$). The model is based on a deep character-level language model implemented as a bidirectional LSTM. To be comparable to other static word embeddings, we use the static context-insensitive fully character-based type layer to obtain static ELMo vectors. The same technique was used by Peters et al. (2018), and we refer the reader to the original paper for further details. We use the ELMo variant pre-trained on the 1 Billion Word Benchmark.[16] We also evaluate two approaches to extracting word-level representations from pre-trained Transformer-based BERT models (Devlin et al. 2019), whereby words are fed into the model (i) *in isolation* or (ii) *in context*. To obtain lexical-level representations with the first method (i), we follow prior work (Liu et al. 2019b; Vulić et al. 2020) and compute each verb representation by (1) feeding it to a pre-trained BERT model *in isolation*; and then (2) averaging the H hidden representations (bottom-to-top) for each of the verb's constituent subwords. We then (3) average the resulting subword representations to produce the final d-dim vector (ISO). This approach does not require any additional external corpora for the induction of such BERT-based embeddings. We experimented with different values of $H = \{4, 6, 8\}$, as well as an alternative approach where only the representation from the input embedding layer is used, without layer-wise averaging, as done in prior work (Wang et al. 2019; Conneau et al. 2020). We also examined two approaches to subword representation averaging, one where special tokens (`[CLS]` and `[SEP]`) are included and one where they are excluded from the averaging step. We found that exclusion of special tokens results in consistently stronger performance across models and values of H. We report results for the strongest performing configuration across tasks, averaging representations from the first $H = 6$ layers and excluding special tokens from subword averaging.[17]

14 `https://levyomer.wordpress.com/2014/04/25/dependency-based-word-embeddings/`.

15 `https://fasttext.cc/docs/en/crawl-vectors.html`.

16 `https://allennlp.org/elmo`.

17 Although exclusion of special tokens produced consistently stronger embeddings, we observed more variation in scores for different values of H, suggesting careful tuning of this parameter is necessary to achieve optimum performance. Supplementary results for the different word embedding extraction configurations from BERT models are included in the Appendix.

The second method (ii) allows us to encode word meaning *in context*, using external text corpora, by first learning N token-level representations for each word and subsequently aggregating them into a static type-level representation as before. We chose English Europarl (Koehn 2005) as the external corpus, from which we (1) randomly sampled N sentences containing each word in the sample; then, (2) we computed each word's representation in N sentential contexts (averaging over constituent subword representations and hidden layers as in steps (2)–(3) of method (i) above), and finally (3) averaged over the N representations to obtain the final representation for each word. We report results for three values of N: 10, 100, and 500 (CONTEXT-10, CONTEXT-100, CONTEXT-500). We probe three different variants of English BERT: BERT-BASE ($d = 768$), BERT-LARGE ($d = 1,024$), and BERT-LARGE with whole word masking (BERT-LARGE-WWM, $d = 1,024$), available in the Transformers repository (Wolf et al. 2019).[18] For further technical details, we refer the reader to the original work.

Furthermore, we test architectures that leverage linguistic information available in external semantic resources. We include sparse binary high-dimensional vectors ($d = 172,418$) proposed by Faruqui and Dyer (2015) (NON-DISTRIBUTIONAL vectors). The vectors are based on a wide variety of hand-crafted linguistic resources such as WordNet, Supersenses, FrameNet, Emotion and Sentiment lexicons, Connotation lexicon, among others.[19] Moreover, we evaluate a retrofitting method that generalizes the model of Wieting et al. (2015) and counter-fitting of Mrkšić et al. (2016) (SGNS+ATTRACT REPEL; AR). It fine-tunes any word vector space by pulling words standing in desirable (i.e., ATTRACT) relations closer together, while simultaneously pushing words in undesirable relations (i.e., REPEL) away from each other (Mrkšić et al. 2017). We evaluate best-performing AR-specialized vectors, reaching human performance on SimLex and SimVerb, introduced by Vulić and Korhonen (2018): They use SGNS-BOW2 as the starting space ($d = 300$), and WordNet and Roget's Thesaurus (Kipfer 2009) as the source of external knowledge. For more details on the exact AR procedure, we refer the reader to the referenced papers. Additionally, we evaluate two collections of BOW2 distributional vectors specialized by Attract-Repel (as in Vulić, Mrkšić, and Korhonen [2017]) using constraints drawn from VerbNet (Kipper et al. 2006) (BOW2-VN) and FrameNet (Baker, Fillmore, and Lowe 1998) (BOW2-FN) to reflect the semantic (shared FrameNet frames) and syntactic-semantic (membership in VerbNet classes) relationships between verbs encoded in those resources.

8.2 Clustering

For each collection of distributional or specialized vectors, we apply a choice of three off-the-shelf clustering algorithms to group the 825-verb sample (used in Phase 1) into classes based on their similarity: the MNCut spectral clustering algorithm (Meila and Shi 2001), as in prior work (Brew and Schulte im Walde 2002; Sun and Korhonen 2009; Sun et al. 2010), the K-means clustering (Brew and Schulte im Walde 2002; Sun et al. 2010), and agglomerative clustering with average linkage.

We apply standard evaluation metrics from previous work on verb clustering (Ó Séaghdha and Copestake 2008; Sun and Korhonen 2009; Sun et al. 2010; Falk, Gardent, and Lamirel 2012; Vulić, Mrkšić, and Korhonen). Modified purity (MPUR), that is, the mean precision of automatically induced verb clusters, is calculated as:

18 github.com/huggingface/transformers.

19 https://github.com/mfaruqui/non-distributional.

$$\text{MPUR} = \frac{\sum_{C \in \textbf{Clust}, n_{prev(C)} > 1} n_{prev(C)}}{\#test\_verbs} \quad (7)$$

where each cluster C from the set of all K_{Clust} induced clusters **Clust** is associated with its prevalent gold class (from Phase 1), and $n_{prev(C)}$ is the number of verbs in an induced cluster C taking that prevalent class, with all other verbs considered errors. *#test_verbs* is the total number of test verb instances.[20] Weighted class accuracy (WACC), which targets recall, is computed as:

$$\text{WACC} = \frac{\sum_{C \in \textbf{Gold}} n_{dom(C)}}{\#test\_verbs} \quad (8)$$

where for each class C from the set of gold standard classes **Gold** (Phase 1 classes, $K_{Gold} = 17$) we identify the dominant cluster from the set of induced clusters having most verbs in common with C ($n_{dom(C)}$). We combine the two metrics into an F1 score, calculated as the balanced harmonic mean of MPUR and WACC.

Results and Discussion. We obtain strongest results from spectral clustering (as previously in Scarton et al. 2014; Vulić, Mrkšić, and Korhonen 2017), 0.01 points on average ahead of K-means and 0.02 in favor of agglomerative clustering. Table 6 summarizes the F1 spectral clustering scores for our chosen vector collections, for the optimal value of k (optimal) and for $k = K_{Gold}$ (gold). We note strongest results from the FrameNet-specialized vectors (BOW2-FN), which is an outcome attributable to the nature of the Phase 1 classes, characterized by thematic similarity, in line with the rationale underlying FrameNet frames. While the absolute scores are not high (F1$<$ 0.5 for all vector collections), the relative scores are informative. We note progressive improvement in performance across the different types of external knowledge used for vector space fine-tuning, from the WordNet- and Roget's Thesaurus-specialized SGNS+ATTRACT-REPEL vectors, through VerbNet-specialized BOW2-VN embeddings, up to the top-performing BOW2-FN. FrameNet is a fine-grained resource including 1,224 semantic frames, some of which describing very specific semantic scenarios. It is therefore quite different in structure from our broad Phase 1 classes. However, the fact that FrameNet knowledge boosts clustering performance suggests the rationale behind human judgments in our rough clustering task aligns somewhat with the hypothesis underlying the organization of verbs into FrameNet frames.

Overall, we observe stronger performance from the static distributional models compared to the Transformer-based BERT architectures. Manual inspection of clusters output by the latter systems reveals groupings biased by subword information (e.g., BERT-BASE clusters words *nag*, *wag*; *thaw, yawn*, and *soar, soak* together), rather than reflecting semantic overlap (as in the "sound" cluster [*cry, squeal, squeak, roar, rattle, hoot, scream*, etc.] produced by BOW2-FN). Our extracted BERT word-level representations capture substantial surface-level information that impacts cluster assignments; However, the embeddings computed *in context* offer improvements over their *in isolation* counterparts. We also note that the number of contextual representations (N) aggregated into the final word-level embeddings which yield strongest results varies

20 As in prior work, we discard clusters with $n_{prev(C)} = 1$ from the count to avoid bias from singleton clusters (Vulić, Mrkšić, and Korhonen 2017; Sun and Korhonen 2009; Sun et al. 2010).

Table 6
F1 scores obtained by representation models on the clustering task, for the optimal value of k (F1 optimal) and for $k = K_{Gold}$ (F1 gold), evaluated against Phase 1 classes. For BERT-BASE and BERT-LARGE models, we evaluate both the embeddings computed *in isolation* (ISO) and *in context*, for three values of N (10, 100, 500), corresponding to the number of contextualized representations aggregated into the final word-level embedding. Numbers in brackets refer to vector dimensionality.

Model (Dimensionality)	F1 optimal	F1 gold
SGNS-BOW2 (300)	0.355	0.326
CBOW-CC (300)	0.426	0.383
ELMo-Static (300)	0.394	0.387
NON-DISTRIBUTIONAL (172,418)	0.391	0.360
SGNS+ATTRACT-REPEL (300)	0.392	0.354
BOW2-VN (300)	0.416	0.404
BOW2-FN (300)	**0.444**	**0.429**
BERT-BASE (768) (ISO)	0.338	0.310
CONTEXT-10	0.338	0.312
CONTEXT-100	0.340	0.322
CONTEXT-500	0.332	0.309
BERT-LARGE (1,024) (ISO)	0.297	0.269
CONTEXT-10	0.339	0.325
CONTEXT-100	0.334	0.304
CONTEXT-500	0.350	0.323
BERT-LARGE-WWM (1,024) (ISO)	0.323	0.308

between models and the values of k. Contextualized BERT-LARGE embeddings achieve the highest scores across BERT models, producing clusters characterized by greater semantic coherence (e.g., "possession" verbs including *gather, buy, collect, possess, obtain, steal, borrow, earn, get*, and "cognitive" verbs like *assume, realize, examine, compute, analyze, doubt, guess, understand*). In the next section, we further examine the impact of computing representations in context rather than in isolation on the quality of the semantic information captured in the word similarity task.

8.3 Word Similarity

Table 7 reports the results of evaluation of the chosen models on SpA-Verb (29,721 pairs) and the thresholded subset (SpA-Verb-THR), and the subset of pairs shared with SimVerb-3500 (1,682 pairs), using both the original SimVerb scores and those obtained via our arena-based method. The reported scores are Spearman's ρ coefficients of the correlation between the ranks derived from models' similarity scores (i.e., cosine distances in the embedding space) and from human similarity judgments in Phase 2.

Results and Discussion. A number of interesting observations can be drawn from the evaluation. First, we note that the highest scores are obtained by linguistically informed models, drawing from diverse rich lexical resources to better capture a range of semantic relations and phenomena (e.g., synonymy and antonymy [Vulić and Korhonen 2018], sentiment polarity and connotation [Faruqui and Dyer 2015]).

Table 7
Evaluation of selected state-of-the-art representation learning models on the full SimVerb-3500 data set (**SV-3500**), the subset of pairs shared by SimVerb and our data set, using both the original SimVerb scores (**SV∩SpA_SVs**) and scores obtained via our arena-based method (**SV∩SpA_SpAs**), as well as our full similarity data set (**SpA-Verb**) and the thresholded subset (**SpA-Verb-THR**) of the whole data set (10,371 pairs from the classes with IAA $\geq$ 0.3). All scores are Spearman's ρ correlations. Numbers in brackets refer to vector dimensionality.

Model (Dimensionality)	SV-3500	SV∩ SpA_SVs	SV∩ SpA_SpAs	SpA-Verb	SpA-Verb-THR
SGNS-BOW2 (300)	0.275	0.197	0.136	0.179	0.158
CBOW-CC (300)	0.365	0.264	0.242	0.271	0.305
ELMo-Static (300)	0.414	0.327	0.310	0.230	0.227
NON-DISTRIBUTIONAL (172,418)	0.606	0.543	0.479	0.295	0.310
SGNS+ATTRACT-REPEL (300)	**0.766**	**0.730**	**0.567**	**0.385**	**0.394**
BERT-BASE (768) (ISO)	0.338	0.224	0.207	0.235	0.240
CONTEXT-10	0.436	0.326	0.228	0.262	0.266
CONTEXT-100	0.438	0.326	0.231	0.265	0.271
CONTEXT-500	0.439	0.327	0.231	0.265	0.270
BERT-LARGE (1,024) (ISO)	0.319	0.240	0.188	0.224	0.215
CONTEXT-10	0.403	0.305	0.226	0.255	0.269
CONTEXT-100	0.402	0.304	0.225	0.256	0.270
CONTEXT-500	0.403	0.304	0.225	0.256	0.270
BERT-LARGE-WWM (1,024) (ISO)	0.396	0.307	0.257	0.237	0.246

The fact that these representations score the highest on SpA-Verb reveals the potential of our spatial arrangement-based method to capture fine-grained semantic properties. It also indicates that non-expert native speakers without formal linguistic training reflect on the components of word meaning and perform some form of linguistic analysis intuitively. This suggests that the spatial method may lend itself to the creation of rich lexical resources, and not only simple pairwise similarity data sets. We observe that the performance of pre-trained encoders on SpA-Verb (and SimVerb) lags behind that of the top-performing static representations. However, they consistently outperform the SGNS-BOW2 model and, at their strongest, are competitive with the CBOW-CC and ELMo-Static vectors. Within the three pre-training models, we observe that BERT-BASE mostly outperforms the larger BERT-LARGE model, whose performance improves, however, when using word-level masking (BERT-LARGE-WWM). We again note a clear advantage of computing word-level representations *in context*, which better leverages the ability of the Transformer models to learn dynamic representations of meaning: We see noticeable improvements over the *in isolation* (ISO) variant across the board, even if the number of contextualized (token-level) representations aggregated into the final word-level (type-level) representation has little to no impact on performance.

SpA-Verb vs. SimVerb. Interesting observations can be drawn from an analysis of correlation between model rankings on the shared subset of pairs in SimVerb-3500 (SV∩SpA_SVs) and scores obtained via our arena-based method (SV∩SpA_SpAs). The two sets of results show very strong correlation (Spearman's $\rho = 0.86$), with near-perfect

correlation ($\rho = 0.98$) between results obtained by static embeddings. This supports the hypothesis that it is indeed possible for humans to capture semantic similarity in their spatial arrangements, and SpA-Verb can be reliably used for comparing representation models, while offering a more comprehensive and challenging evaluation benchmark.

The analysis of correlation between model rankings on full SimVerb and SpA-Verb again produces a high correlation score ($\rho = 0.77$). These figures are enlightening when compared with similar analyses in previous work (Vulić, Kiela, and Korhonen 2017). While Vulić, Kiela, and Korhonen (2017) report very high correlations between model rankings on SimLex and SimVerb (>0.95), both of which measure semantic similarity, the scores are much lower between model rankings on SimLex or SimVerb and MEN (Bruni, Tran, and Baroni 2014) (0.342 and 0.448, respectively), a data set which captures broader conceptual relatedness. This suggests that SpA-Verb aligns with SimLex and SimVerb in its treatment of semantic similarity and relatedness, and that our spatial interface combined with instructions to arrange words based on the similarity of their meaning allow the annotators to capture word similarity as distinct from relatedness and association.

As results in Table 7 indicate, SpA-Verb is particularly challenging for models learning from co-occurrence information (however, incorporation of subword information helps performance, as seen in the scores obtained by the CBOW-CC model). Completely unrelated word pairings, which are easy to capture based on distributional data, are removed in the first phase, leaving only fine-grained semantic distinctions between related concepts.

8.4 Evaluation on Highly Associated Pairs and on High-IAA Classes

One of the main challenges for distributional models is to tease apart associated *and* similar words from those that are highly associated and frequently co-occurring but dissimilar (e.g., *cook - bake* vs. *cook - eat*). As our Phase 2 focuses on similarity of meaning disregarding association, we can subsample all the highly associated pairs—both similar and dissimilar—to create an evaluation sample specifically testing the models' capacity to recognize this distinction. Following the evaluation on the highly associated set, we examine whether, in turn, the semantic classes that proved easier for annotators to reason about are also less challenging for models by focusing the evaluation on classes with the strongest annotator consensus.

Highly Associated Pairs. We evaluate our selection of models on the top most associated quartile of pairs (according to their USF association score, as in Hill, Reichart, and Korhonen [2015]), in comparison with their performance on the entirety of SpA-Verb. This also allows us to further investigate the nature of the semantic relations captured by our arena-based judgments and our design's capacity to produce similarity ratings unbiased by association, despite not using an explicitly defined rating scale. As has been observed for data sets capturing similarity as distinct from association (e.g., SimLex), we expected that performance of models learning solely from distributional information may be negatively affected, as strong co-occurrence evidence for the highly associated pairs has been shown to cause systems to overestimate word similarity (Hill, Reichart, and Korhonen 2015). Table 8 presents the results of this analysis. As predicted, we see a performance drop for the SGNS-BOW2 model, the subword-informed CBOW-CC vectors, as well as the three BERT models, both the *in isolation* and *in context* variants (the latter again proving more robust than the former). The remaining systems improve, with linguistically informed NON-DISTRIBUTIONAL and SGNS+ATTRACT-

Table 8
Evaluation on the top quartile of most associated pairs in SpA-Verb (**Top Q**), compared against Spearman's correlation scores on the whole data set (**SpA-Verb**), and on the top 3 classes with the highest IAA ($\rho > 0.50$) (**#3, #1, #13**).

Model	SpA-Verb	Top Q	#3	#1	#13
SGNS-BOW2	0.179	0.069	0.082	0.045	0.108
CBOW-CC	0.271	0.216	0.084	0.322	0.378
ELMo-Static	0.230	0.237	0.098	0.114	0.238
NON-DISTRIBUTIONAL	0.295	0.450	0.386	**0.359**	0.480
SGNS+ATTRACT-REPEL	**0.385**	**0.526**	**0.718**	0.338	**0.534**
BOW2-VN	0.176	0.205	−0.037	0.198	0.212
BOW2-FN	0.210	0.287	0.095	0.242	0.156
BERT-BASE (ISO)	0.235	0.181	0.123	0.201	0.330
CONTEXT-10	0.262	0.181	0.154	0.177	0.366
CONTEXT-100	0.265	0.183	0.149	0.193	0.380
CONTEXT-500	0.265	0.185	0.148	0.190	0.379
BERT-LARGE (ISO)	0.224	0.152	0.164	0.245	0.246
CONTEXT-10	0.255	0.163	0.180	0.242	0.318
CONTEXT-100	0.256	0.167	0.172	0.245	0.326
CONTEXT-500	0.256	0.167	0.171	0.245	0.326
BERT-LARGE-WWM (ISO)	0.237	0.195	0.170	0.215	0.276

REPEL models performing noticeably better on these difficult cases than on the entire data set. Notably, the consistently strong performance of representations drawing on lexicon information shows that human judgments collected in our arena-based task correlate with the expert knowledge coded in manually crafted linguistic resources.

High-IAA Classes. Having evaluated how well the different representation models cope with the difficult subset of highly associated pairs, we were interested to see whether high human agreement on certain verb classes correlates with model performance, that is, to what extent the classes which were easier for human annotators to judge prove to be an easier benchmark for distributional models. Table 8 presents the results of the evaluation on the three Phase 1 classes with highest IAA ($\rho > 0.50$).

While most models improve on class #13 (verbs of sound) compared with the entire data set, the results do not show consistent performance gains for most models. Only the NON-DISTRIBUTIONAL embeddings improve across the three classes, whereas the SGNS+ATTRACT-REPEL model records the largest gain scoring a $\rho = 0.718$ on the highest IAA class #3. Notably, all models except these two record a performance drop on the same class. This is interesting considering the nature of the class, which, as illustrated in Figure 5, contains many verbs with opposite polarity (i.e., negative and positive rate of change), forming pairs of synonyms and antonyms. Vulić and Korhonen's (2018) word vector space specialization model is designed precisely to allow fine-tuning of distributional vector spaces to distinguish between synonymy and antonymy, making use of linguistic constraints derived from external resources that specify the exact lexical relation between a pair of words. This also explains the very low correlation scores achieved by FrameNet and VerbNet-specialized models: Both of these resources group antonymous rate of change verbs together, due to their shared syntactic behavior and high semantic overlap along all meaning

dimensions but one, that of polarity or direction of change. Making this distinction is perennially difficult for statistical models learning purely from distributional information, since antonyms and synonyms have similar co-occurrence patterns in corpora; BERT embeddings prove the strongest in this category, outperforming the BOW models and static ELMo vectors on this difficult class. Crucially, the high correlation between the SGNS+ATTRACT-REPEL model and the human judgments suggests that our approach enables capturing these important, cognitively salient semantic relations between the otherwise related items, and holds promise for more fine-grained linguistic analyses.

8.5 Evaluation on Semantically Focused Subsets

Typological study of the regularities in the way conceptual components are encoded in lexical items, that is, lexicalization patterns, groups languages into types based on the lexicalization strategies they permit. As far as verbs are concerned, cross-linguistic differences regard, for instance, the elements that are encoded in or outside the verb. The strategy characteristic for English directed motion verbs is to conflate the semantic elements of "Motion" and "Manner" inside the verb, and express "Path" outside (e.g., *The tennis ball rolled down the slide*) (as opposed to, for example, Italian, where "Motion" and "Path" are encoded together in the verbal root, and "Manner" may be expressed as a gerundive adjunct) (Talmy 1985; Folli and Ramchand 2005). The preference for a certain lexicalization pattern impacts the verb inventory of a given language: English and other languages where the first pattern is typical tend to have large repertoires of verbs expressing motion occurring in various manners. This is reflected in the most numerous class in our data set, which includes 100 motion verbs, many of which make subtle distinctions regarding the way in which an action is performed. However, this phenomenon is not restricted to verbs of motion: Languages that display a preference for lexicalizing manner of motion often possess an extensive inventory of verbs expressing manner in general, for example, manner of speaking, manner of looking (Majewska et al. 2018b). The vast coverage of our resource allows us to zoom in to these densely populated meaning domains to examine the capacity of representation models to capture the subtle meaning distinctions in the manner in which an action described by a verb is performed.

We evaluate our selection of models on five verb pair sets, including verbs belonging to specific meaning domains. Included are motion verbs (Class 10, 4950 pairs), as well as four subsets of pairs defined in terms of participation in FrameNet frames: verbs related to heat (*Absorb_heat*, *Apply_heat*; 46 pairs), experiencing emotions (*Experiencer_obj*, *Experiencer_focus*, *Cause_to_experience*, *Feeling*; 237 pairs), producing sound (*Cause_to_make_noise*, *Communication_noise*, *Make_noise*, *Motion_noise*, *Sound_movement*; 235 pairs), and causing or experiencing pain (*Cause_harm*, *Experience_bodily_harm*, *Perception_body*, *Cause_bodily_experience*; 219 pairs). Table 9 summarizes the results. Across the domains, we see best performance from the linguistically informed representations and consistently strong performance from SGNS+ATTRACT-REPEL. Moreover, the patterns of correlation scores obtained by the two sets of vectors specialized for VerbNet (BOW2-VN) and FrameNet (BOW2-FN) provide some more evidence regarding where these two lexical resources and our data set align and diverge in terms of the organization of the same concepts. The BOW2-VN model achieves by far the highest result on the Heat subset (0.787), but performs very poorly on the Pain subset (0.075), where, in turn, the FrameNet-specialized model leads (0.465). In the meaning domain related to causing and experiencing pain, the annotators' judgments align more closely with

Table 9
Evaluation of representation models on subsets of SpA-Verb verb pairs focused on particular semantic domains. All scores are Spearman's ρ.

Model	Motion	Heat	Sound	Emotion	Pain
SGNS-BOW2	0.247	0.078	0.186	0.160	0.023
CBOW-CC	0.275	0.534	0.416	0.265	0.277
ELMo-Static	0.300	0.232	0.301	0.317	0.003
NON-DISTRIBUTIONAL	0.341	0.631	0.549	0.232	0.224
SGNS+ATTRACT-REPEL	**0.410**	0.374	**0.588**	**0.359**	0.445
BOW2-VN	0.327	**0.787**	0.201	0.294	0.075
BOW2-FN	0.368	0.393	0.419	0.264	**0.465**
BERT-BASE (ISO)	0.262	0.138	0.230	0.201	0.185
CONTEXT-10	0.239	−.125	0.283	0.336	0.243
CONTEXT-100	0.239	−.081	0.290	0.330	0.253
CONTEXT-500	0.238	−.081	0.289	0.331	0.255
BERT-LARGE (ISO)	0.254	0.011	0.176	0.248	0.280
CONTEXT-10	0.258	0.179	0.272	0.333	0.337
CONTEXT-100	0.256	0.188	0.279	0.329	0.347
CONTEXT-500	0.256	0.188	0.280	0.332	0.348
BERT-LARGE-WWM (ISO)	0.283	0.339	0.197	0.242	0.260

the distinctions captured by the different FrameNet frames (i.e., causing harm vs. experiencing harm vs. experiencing a [non-harmful] bodily sensation vs. causing a non-harmful bodily experience) than they do with the semantic-syntactic classes containing the same verbs in VerbNet. However, where FrameNet frames are broader, as it is the case in the Heat domain, we see the VerbNet-specialized vectors achieving better performance: For example, while the FrameNet *Apply_heat* frame groups verbs such as *cook*, *boil*, and *melt* together, VerbNet divides them into two classes, *cooking-45.3* and *other_cos (change of state)-45.4*, which aligns more closely with our annotators' judgments (*cook-boil* have a small score 0.238, while *melt-cook* 0.797 and *melt-boil* 0.885, reflecting the greater distances between these concepts).

Examination of the scores recorded for the three BERT model variants *in isolation* reveals the large model with whole word masking (BERT-LARGE-WWM) to be the most robust across the different semantic areas, and this advantage is particularly visible on the smallest set of Heat verbs. Notwithstanding the subpar performance of contextualized BERT-BASE on this semantic domain (where fluctuations in scores are more likely given the very small size of this set), the benefits of computing BERT embeddings *in context* are again clear: For instance, contextualized BERT-BASE and BERT-LARGE embeddings are especially competitive on Emotion verbs, coming a close second behind SGNS+ATTRACT-REPEL.

The scale of our data set allows for zooming in on word subsets with desired characteristics and creating smaller data sets controlling for some specific feature, showing potential for more focused analyses of representation models. Similar analyses could shed light on particular strengths and weaknesses of representation architectures and help identify meaning domains and semantic properties requiring systems to be more specialized, or different modeling strategies altogether.

8.6 Further Discussion

In the preceding sections, we examined how well the lexical representations derived from the selected models correlate with human judgments collected through spatial arrangements. We observed that the semantic distinctions that are easier for humans to make often elude representation models, and that discriminating between similar and highly associated but dissimilar words remains a challenge for most systems. Moreover, we saw that model performance varies across different semantic classes, revealing inconsistencies in representation quality for verbs belonging to different domains. The results have also revealed interesting patterns that open up further questions concerning the implications of evaluating representation architectures on SpA-Verb (and lexical semantic similarity data sets in general), which we address below.

SpA-Verb vs. SimVerb. In Sections 6 and 7, we examined how the differences in two data collection methodologies, pairwise ratings and spatial arrangements, affect the characteristics of the produced data sets. In turn, the evaluation results in Table 7 showed that these differences translate to the data sets' respective difficulty: The consistently lower performance on SpA-Verb with respect to SimVerb suggests the former is a more challenging benchmark. In light of these differences, an important question that arises is: Are the insights into the intrinsic quality of lexical representations which each of these data sets provides fundamentally different or rather complementary? And if SpA-Verb offers larger, more comprehensive coverage and higher granularity with respect to SimVerb, is there some important signal that SimVerb provides that is missing from SpA-Verb?

As discussed in Section 6, an important difference between the two data sets concerns unrelated verb pairs (e.g., *broil - respect*, *bounce - prohibit*). Our annotation design filters out the completely unrelated pairs in Phase 1, where verbs are grouped based on shared semantics and relatedness. The main motivation behind this choice was to elicit similarity judgments on comparable concepts (the importance of which has been emphasized in psychology [Turner et al. 1987]), while avoiding the conflation of scores for unrelated words and antonyms, which are related but dissimilar (a phenomenon characteristic for SimVerb). However, a potential disadvantage of such a solution is the complete exclusion of unrelated pairs from SpA-Verb, which precludes direct evaluation of the capacity of a model to tackle such examples. In other words, a system could be overestimating the similarity of unrelated words and still score highly on SpA-Verb.

First of all, it is important to note that such a hypothetical scenario seems rather unlikely. Because distributional models learn about word meaning from co-occurrence patterns, the notion of (associative) relatedness, characterizing words frequently appearing together in text, is the knowledge easiest to glean from raw data. SpA-Verb requires models to exhibit an understanding of lexical semantics that is much more advanced: Systems need to be sensitive to fine-grained meaning distinctions and degrees of similarity between related words. Crucially, it includes antonymous pairs, which pose a significant challenge to models learning from distributional information (Mrkšić et al. 2016; Vulić and Korhonen 2018). Further, the high number of close similarity scores makes the task difficult, compared to the sparse, discrete ratings in SimVerb. Viewed this way, the exclusion of unrelated pairs makes the data set more challenging: The unrelated pairs are easiest to judge both for humans (i.e., pairs of words which have nothing in common are given 0 scores in SimVerb) and for distributional semantic models, based on the negative signal (i.e., no co-occurrence in text) that is easily derived

from corpora. Because SpA-Verb does not reward models on these easy cases, the absolute scores are unsurprisingly lower than on SimVerb. However, they may provide a more realistic measure of representation quality and the capacity of models to reason about lexical semantics. Conversely, the fact that the unrelated pairs are most numerous in SimVerb (i.e., the 0–1 score interval in Figure 6) may result in an undesirable inflation of the estimate of representation quality. To sum up, while both SimVerb and SpA-Verb reward models capable of distinguishing between similar and related-but-dissimilar concepts, SpA-Verb requires models to make many nuanced, fine-grained distinctions between the *degrees* of similarity and dissimilarity of related, semantically proximate words, in a densely populated semantic space.

Notwithstanding the potential benefits of not including unrelated pairs in the data set, researchers may be explicitly interested in examining the models' effectiveness in dealing with such cases. While using the subset of unrelated pairs from SimVerb is the obvious choice, unrelated examples are easily obtainable from SpA-Verb as well, based on the output of the semantic clustering in Phase 1. Because annotators group verbs into theme classes based on their semantics, the unrelated words are separated into different clusters. Unrelated pairs can therefore be easily generated by randomly sampling pairs from different Phase 1 classes.

Static Word Embeddings vs. Pre-trained Encoders. Another noteworthy pattern that emerges from the evaluation experiments is the relatively weaker performance of the BERT models compared with the stronger static representations, despite the proven superiority of Transformer-based architectures across diverse NLP tasks. Because intrinsic tasks such as word similarity prediction serve as a proxy for estimating model performance in downstream applications, the under-performing BERT embeddings raise questions: What makes the contextualized representations less successful on SpA-Verb, and what is it telling us about the lexical semantic signal they encode?

The first important factor responsible for this phenomenon lies in the fundamental difference between static word embeddings and the representations produced by BERT. Whereas the former assign a single, fixed vector to a given word, the latter encode meaning dynamically, in context. In order to derive comparable lexical representations from BERT, we need to abstract away from individual word occurrences and aggregate token embeddings into a single, static type-level representation. A number of solutions have been proposed to this end (Liu et al. 2019b; Conneau et al. 2020; Cao, Kitaev, and Klein 2020, inter alia). In this article, we experimented with two different approaches, one where the target word is fed into the model in isolation, and one where it appears in N full sentences and the type-level embedding is derived by averaging over each such token-level representation. Each of these variants, however, requires careful tuning of the configuration of the following parameters: (i) the choice of hidden representations to average over; (ii) the selection of special tokens (i.e., `[SEP]` and `[CLS]` tokens in BERT) to include in the subword representation averaging step. While the results presented in this article are achieved by BERT representations yielded by the strongest such parameter configuration across the reported tasks and word pair sets, further investigations into alternative approaches to deriving lexical representations from pre-trained models are needed to maximize their competitiveness.

What is worth noting is that intrinsic word similarity data sets do not directly reward the capacity of Transformer-based models to capture word meaning in context. Indeed, the results in Table 7 show analogous patterns of model scores on SimVerb and on SpA-Verb, and for both the static embeddings drawing on external linguistic information outperform the embeddings derived from BERT by a significant margin.

Notably, the scores achieved by all models on the shared pairs from both resources show strong correlation ($\rho = 0.86$), which indicates that the observed phenomenon is not an idiosyncrasy of the spatial similarity data set, but is common to both resources. However, this characteristic does not preclude SpA-Verb's applicability as a discriminator of the quality of lexical representations yielded by pre-trained encoders such as BERT. Many recent efforts focused on investigating *why* state-of-the-art pre-trained models perform as well as they do in downstream applications, probing the linguistic knowledge captured by those architectures (Liu et al. 2019a; Tenney, Das, and Pavlick 2019; Jawahar, Sagot, and Seddah 2019; Hewitt and Manning 2019). Importantly, it has been noted that their success may be due to learning shortcuts in NLP tasks, rather than being a direct product of the quality and richness of the encoded linguistic knowledge (Rogers, Kovaleva, and Rumshisky 2020). Indeed, a number of works have drawn attention to BERT's subpar verbal reasoning abilities and its reliance on shallow heuristics in natural language inference and reading comprehension (McCoy, Pavlick, and Linzen 2019; Zellers et al. 2019; Si et al. 2019; Rogers et al. 2020; Sugawara et al. 2020).

Because BERT takes subword tokens as input in pre-training, the question of whether and how it captures *lexical* signal merits investigation. SpA-Verb meets this need as a lexical semantic probing tool, enabling direct evaluation of the quality of the lexical knowledge stored in the parameters of BERT. Our experiments on SpA-Verb and specific subsets of the data set have already provided some insights: We observed that the word-level embeddings obtained by averaging over N occurrences in context encode richer lexical semantic knowledge than those derived by feeding words into the pre-trained model in isolation. Further, our experimentation with different configurations of lexical representation extraction parameters, such as the choice of hidden layers from which to derive the ultimate representation or the inclusion of special tokens, revealed that it is advantageous to draw type-level verbal lexical knowledge from the first 6 layers, while the inclusion of special tokens degrades representation quality. Future experiments using the spatial similarity data may help scrutinize the nature and location of the lexical semantic signal encoded in these representations, and its contribution to downstream performance. Moreover, probing analyses using subsets of SpA-Verb targeting particular semantic domains may help uncover the areas where BERT's lexical representation quality is still insufficient and aid development of systems with a stronger grasp of verbs' lexical-semantic properties.

9. Conclusion and Future Work

We presented and thoroughly analyzed a new method for large-scale collection of semantic similarity data based on clustering and spatial arrangements of lexical items. The method adapts the spatial approach, previously used only with visual stimuli, to polysemous lexical items in a large-scale setting, that is, a word sample seven times as numerous as the biggest stimuli sets used in SpAM-based research to date. Our two-phase approach, consisting of rough clustering of a large verb sample into classes of similar and related verbs and subsequent spatial arrangements of these classes in a 2D arena, crucially produces both semantic clusters and word pair scores within an integrated framework, and can be readily applied to other parts of speech and types of stimuli.

Our methodology offers several benefits compared with the traditional approaches using discrete pairwise ratings. First, the two-phase design enabled us to handle lexical ambiguity as a natural consequence of overlap in class membership in the first rough

clustering phase. Cluster analysis performed on the output from that first phase demonstrated that the approach could be very useful in future to aid building more comprehensive lexical resources. Furthermore, our spatial arrangement approach (the second phase) captures non-expert intuitions about word meaning, allowing annotators to make nuanced decisions by considering the semantics of multiple lemmas together that elude simple pairwise similarity judgments. Moreover, the method is easily portable to other languages, demonstrating potential for faster creation of human evaluation data sets to support multilingual NLP. Our approach yielded SpA-Verb, a data set of fine-grained similarity scores for 29,721 unique verb pairs, together with 17 thematic verb classes.

The comparative analyses against FrameNet, VerbNet, and WordNet showed that our two-phase design allows humans to differentiate between a range of semantic relations and intuitively capture fine-grained linguistic distinctions pertaining to verb semantics through subtle relative judgments. The encouraging overlap with VerbNet classes suggests the method could aid incorporating new verbs into the existing network, or producing similar resources from scratch for other languages. The automatic clustering experiments run on top of the distance matrices from Phase 2 also demonstrated the potential of our design to yield semantic clusters within each broad class, which can be used in future work to evaluate the capacity of models to create semantic classifications and taxonomies automatically. What is more, by yielding complete distance matrices for each class, our design allows in-depth exploration of the dimensions underlying the organization of the semantic space in question, holding promise to support cognitive linguistics research.

We examined the properties of our data set and its potential as an evaluation resource by evaluating a selection of state-of-the-art representation models on two tasks, corresponding to the two phases of our design: (1) clustering, using Phase 1 classes as gold truth, and (2) word similarity, using pairwise scores from the entire SpA-Verb (29,721 pairs) and the thresholded set (10,371 pairs), as well as selected subsets with different semantic characteristics. While overall our experiments showed the stronger static word vectors, especially those drawing on external linguistic knowledge, to surpass the embeddings extracted from pre-trained Transformer-based BERT models on both word-level semantic similarity and clustering, we found that the potential of these architectures to capture word-level semantics can be better leveraged by aggregating a number of contextualized token-level representations into the final type-level embeddings. Further, the experiments focused on chosen semantic subsets allowed us to contrast the performance of embeddings incorporating external linguistic information from either VerbNet or FrameNet on specific semantic domains, while providing additional evidence for the importance of models' capacity to distinguish between relations of synonymy and antonymy. Importantly, the low performance of many state-of-the-art models on our data set suggests that the complex, many-to-many judgments recorded in the arena are still very difficult to model. The fact that our data set contains nuanced similarity judgments between semantically close verbs means that the resource provides a challenging benchmark for state-of-the-art systems, which will be useful in research aimed at improving the capacity of NLP models to represent the complex meaning of verbs and events they describe. Moreover, the large size of the data set offers vast possibilities for robust analyses on different word subsets and semantically related classes, allowing for better informed tuning and comparison of the adequacy and potential of various representation learning architectures to capture fine-grained semantic distinctions present in the mental lexicon, while helping achieve greater model interpretability.

The resource and the analyses reported in this article open up several avenues for future work. First, we will expand the evaluation of representation models on SpA-Verb data further and test them in a verb classification task on the narrow semantic clusters extracted from Phase 2 distance matrices. This will allow for assessing models' capacity to create fine-grained verb classes automatically, which could support creation of lexical resources in languages and domains where those are still lacking. Furthermore, to investigate the method's portability, we will carry out data collection for other parts of speech and typologically diverse languages to analyze cross-linguistic similarities and variation.

Appendix A: Supplementary Results

Table A.1 presents additional results of the evaluation of BERT embeddings computed *in isolation* using different lexical representation extraction configurations. To identify

Table A.1
Evaluation results across different lexical representation extraction configurations on SimVerb -3500 and SpA-Verb data sets and the subset of shared pairs (cf. Table 7). L_0 refers to the input embedding layer; $\leq L_n$ refers to embeddings computed by averaging representations over all Transformer layers up to and inclusive of the nth layer. For each layer averaging configuration we consider two configurations of special tokens (column SPEC): one where special tokens [CLS] and [SEP] are included (+) and one where they are excluded (−) from the subword embedding averaging step. All scores are Spearman's ρ correlations.

Configuration		SV-3500	SV∩SpA_SVs	SV∩SpA_SpAs	SpA-Verb	SpA-Verb-THR
L	**SPEC**					
BERT-BASE						
L_0	+	0.219	0.142	0.096	0.167	0.195
	−	0.318	0.214	0.169	0.212	0.220
$\leq L_4$	+	0.222	0.139	0.107	0.175	0.204
	−	0.338	0.221	0.195	0.239	0.246
$\leq L_6$	+	0.204	0.130	0.093	0.158	0.187
	−	0.338	0.224	0.207	0.235	0.240
$\leq L_8$	+	0.181	0.107	0.079	0.149	0.177
	−	0.315	0.202	0.198	0.219	0.224
BERT-LARGE						
L_0	+	0.214	0.141	0.087	0.156	0.175
	−	0.314	0.221	0.165	0.214	0.225
$\leq L_4$	+	0.190	0.133	0.053	0.158	0.191
	−	0.331	0.248	0.191	0.229	0.222
$\leq L_6$	+	0.184	0.127	0.038	0.145	0.180
	−	0.319	0.240	0.188	0.224	0.215
$\leq L_8$	+	0.184	0.127	0.037	0.141	0.174
	−	0.319	0.240	0.193	0.221	0.214
BERT-LARGE-WWM						
L_0	+	0.211	0.128	0.090	0.167	0.191
	−	0.347	0.245	0.198	0.230	0.242
$\leq L_4$	+	0.226	0.142	0.110	0.177	0.202
	−	0.390	0.295	0.249	0.248	0.258
$\leq L_6$	+	0.210	0.131	0.104	0.160	0.189
	−	0.396	0.307	0.257	0.237	0.246
$\leq L_8$	+	0.211	0.133	0.107	0.156	0.186
	−	0.395	0.312	0.258	0.224	0.234

the strongest performing configuration (the results reported in Section 8) we fine-tuned the following parameters: (1) the choice of hidden layers to average over (L_0 refers to the input embedding layer, $\leq L_n$ refers to the Transformer layers included in layer-wise averaging, inclusive of L_n), (2) the choice of special tokens ([CLS] and [SEP]) included in the subword averaging step. We report the subset of results on SpA-Verb and SimVerb-3500 data (including the subset of shared pairs and the thresholded set of SpA-Verb) to illustrate the impact of the different configurations of lexical representation extraction parameters on performance on both data sets.

Acknowledgments

This work is supported by the ERC Consolidator grant LEXICAL [grant number 648909] and the ESRC Doctoral Fellowship [grant number ES/J500033/1] (to O.M.).

References

Agirre, Eneko, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Pasca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and WordNet-based approaches. In *Proceedings of NAACL-HLT*, 19–27, Boulder, CO. DOI: https://doi.org/10.3115/1620754.1620758

Akoglu, Haldun. 2018. User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine*, 18(3):91–93. DOI: https://doi.org/10.1016/j.tjem.2018.08.001, PMID: 30191186, PMCID: PMC6107969

Al-Rfou, Rami, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual NLP. In *Proceedings of CoNLL*, pages 183–192, Sofia.

Altmann, Gerry T. M. and Yuki Kamide. 1999. Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, 73(3):247–264. DOI: https://doi.org/10.1016/S0010-0277(99)00059-1

Amigó, Enrique, Julio Gonzalo, Javier Artiles, and Felisa Verdejo. 2009. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval*, 12(4):461–486. DOI: https://doi.org/10.1007/s10791-008-9066-8

Armendariz, Carlos Santos, Matthew Purver, Matej Ulčar, Senja Pollak, Nikola Ljubešić, and Mark Granroth-Wilding. 2020. CoSimLex: A resource for evaluating graded word similarity in context. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5878–5886, European Language Resources Association, Marseille.

Asaadi, Shima, Saif Mohammad, and Svetlana Kiritchenko. 2019. Big BiRD: A large, fine-grained, bigram relatedness dataset for examining semantic composition. In *Proceedings of NAACL-HLT*, pages 505–516, Minneapolis, MN. DOI: https://doi.org/10.18653/v1/N19-1050

Avraham, Oded and Yoav Goldberg. 2016. Improving reliability of word similarity evaluation by redesigning annotation task and performance measure. In *Proceedings of REPEVAL*, pages 106–110, Berlin. DOI: https://doi.org/10.18653/v1/W16-2519-2519

Bagga, Amit and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *Proceedings of LREC*, 563–566, Granada.

Baker, Collin F., Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of COLING*, pages 86–90, Montreal. DOI: https://doi.org/10.3115/980845.980860

Baker, Simon, Roi Reichart, and Anna Korhonen. 2014. An unsupervised model for instance level subcategorization acquisition. In *Proceedings of EMNLP*, pages 278–289, Doha. DOI: https://doi.org/10.3115/v1/D14-1034, PMID: 25178667

Baroni, Marco, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of ACL*, pages 238–247, Baltimore, MD. DOI: https://doi.org/10.3115/v1/P14-1023

Baroni, Marco and Alessandro Lenci. 2011. How we BLESSed distributional semantic evaluation. In *Proceedings of the GEMS 2011 Workshop on Geometrical Models of Natural Language Semantics, EMNLP 2011*, pages 1–10, Edinburgh.

Batchkarov, Miroslav, Thomas Kober, Jeremy Reffin, Julie Weeds, and David Weir. 2016.

A critique of word similarity as a method for evaluating distributional semantic models. In *Proceedings of REPEVAL*, pages 7–12, Berlin. DOI: https://doi.org/10.18653/v1/W16-2502

Blair, Philip, Yuval Merhav, and Joel Barry. 2017. Automated generation of multilingual clusters for the evaluation of distributed representations. In *Proceedings of ICLR Workshop Papers*, volume abs/1611.01547, Toulon.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the ACL*. 5:135–146. DOI: https://doi.org/10.1162/tacl_a_00051

Brandes, Ulrik. 2001. A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, 25(2):163–177. DOI: https://doi.org/10.1080/0022250X.2001.9990249

Brew, Chris and Sabine Schulte im Walde. 2002. Spectral clustering for German verbs. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 117–124, Philadelphia, PA. DOI: https://doi.org/10.3115/1118693.1118709

Bruni, Elia, Gemma Boleda, Marco Baroni, and Nam Khanh Tran. 2012. Distributional semantics in technicolor. In *Proceedings of ACL*, pages 136–145, Jeju Island.

Bruni, Elia, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49:1–47. DOI: https://doi.org/10.1613/jair.4135

Budanitsky, Alexander and Graeme Hirst. 2001. Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In *Proceedings of the Workshop on WordNet and Other Lexical Resources*, pages 29–34, Pittsburgh, PA.

Budanitsky, Alexander and Graeme Hirst. 2006. Evaluating WordNet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47. DOI: https://doi.org/10.1162/coli.2006.32.1.13

Cao, Steven, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. In *International Conference on Learning Representations*, Online, https://iclr.cc/virtual/poster_r1xCMyBtPS.html

Casasanto, Daniel. 2008. Similarity and proximity: When does close in space mean close in mind? *Memory & Cognition*, 36(6):1047–1056. DOI: https://doi.org/10.3758/MC.36.6.1047, PMID: 18927023

Chan, Y. H. 2003. Biostatistics 104: correlational analysis. *Singapore Medical Journal*, 44(12):614–9.

Charest, Ian, Rogier A. Kievit, Taylor W. Schmitz, Diana Deca, and Nikolaus Kriegeskorte. 2014. Unique semantic space in the brain of each beholder predicts perceived similarity. *Proceedings of the National Academy of Sciences*, 111(40):14565–14570. DOI: https://doi.org/10.1073/pnas.1402594111, PMID: 25246586, PMCID: PMC4209976

Chiarello, Christine, Curt Burgess, Lorie Richards, and Alma Pollock. 1990. Semantic and associative priming in the cerebral hemispheres: Some words do, some words don't... sometimes, some places. *Brain and Language*, 38(1):75–104. DOI: https://doi.org/10.1016/0093-934X(90)90103-N

Cichy, Radoslaw M., Nikolaus Kriegeskorte, Kamila M. Jozwik, Jasper J. F. van den Bosch, and Ian Charest. 2019. The spatiotemporal neural dynamics underlying perceived similarity for real-world objects. *NeuroImage*, 194:12–24. DOI: https://doi.org/10.1016/j.neuroimage.2019.03.031, PMID: 30894333, PMCID: PMC6547050

Conneau, Alexis, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034. Online. DOI: https://doi.org/10.18653/v1/2020.acl-main.536

Cruse, David A. 1986. *Lexical Semantics*. Cambridge University Press.

Dalitz, Christoph and Katrin E. Bednarek. 2016. Sentiment lexica from paired comparisons. In *Proceedings of ICDM*, pages 924–930, Barcelona. DOI: https://doi.org/10.1109/ICDMW.2016.0135

Day, William H. E. and Herbert Edelsbrunner. 1984. Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of Classification*, 1(1):7–24. DOI: https://doi.org/10.1007/BF01890115, https://doi.org/10.1007/BF01908061

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT:

Pre-training of deep bidirectional transformers for language understanding. 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)*, pages 4171–4186. Minneapolis, MN.

Ester, Martin, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 226–231, Portland, OR.

Estes, Zachary, Sabrina Golonka, and Lara L. Jones. 2011. Thematic thinking: The apprehension and consequences of thematic relations. In *Psychology of Learning and Motivation*, volume 54. Elsevier, pages 249–294. DOI: https://doi.org/10.1016/B978-0-12-385527-5.00008-5

Falk, Ingrid, Claire Gardent, and Jean-Charles Lamirel. 2012. Classifying French verbs using French and English lexical resources. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 854–863, Jeju Island, Korea.

Faruqui, Manaal and Chris Dyer. 2015. Non-distributional word vector representations. In *Proceedings of NAACL-HLT*, pages 464–469. China.

Faruqui, Manaal, Yulia Tsvetkov, Pushpendre Rastogi, and Chris Dyer. 2016. Problems with evaluation of word embeddings using word similarity tasks. In *Proceedings of REPEVAL*, pages 30–35, Berlin. DOI: https://doi.org/10.18653/v1/W16-2506

Fellbaum, Christiane, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press. DOI: https://doi.org/10.7551/mitpress/7287.001.0001

Fillmore, Charles J. 1976. Frame semantics and the nature of language. In *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*, volume 280, pages 20–32, New York. DOI: https://doi.org/10.1111/j.1749-6632.1976.tb25467.x

Fillmore, Charles J. 1977. The need for a frame semantics in linguistics. In *Statistical Methods in Linguistics*. Ed. Hans Karlgren. Scriptor, pages 5–29.

Fillmore, Charles J. 1982. Frame semantics, In *Linguistics in the Morning Calm*. The Linguistic Society of Korea. Hanshin Publishing Co., pages 111–137.

Finkelstein, Lev, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2002. Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, 20(1):116–131. DOI: https://doi.org/10.1145/503104.503110

Folli, Raffaella and Gillian Ramchand. 2005. Prepositions and results in Italian and English: An analysis from event decomposition. In *Perspectives on Aspect*, Springer, pages 81–105. DOI: https://doi.org/10.1007/1-4020-3232-3_5

Frenck-Mestre, Cheryl and Steve Bueno. 1999. Semantic features and semantic categories: Differences in rapid activation of the lexicon. *Brain and Language*, 68(1–2):199–204. DOI: https://doi.org/10.1006/brln.1999.2079, PMID: 10433759

Gärdenfors, Peter. 2004. *Conceptual Spaces: The Geometry of Thought*. MIT Press.

Gentner, Dedre. 1983. Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2):155–170. DOI: https://doi.org/10.1207/s15516709cog0702_3

Gerz, Daniela, Ivan Vulić, Felix Hill, Roi Reichart, and Anna Korhonen. 2016. SimVerb-3500: A large-scale evaluation set of verb similarity. In *Proceedings of EMNLP*, pages 2173–2182, Austin, TX. DOI: https://doi.org/10.18653/v1/D16-1235

Gladkova, Anna and Aleksandr Drozd. 2016. Intrinsic evaluations of word embeddings: What can we do better? In *Proceedings of REPEVAL*, 36–42, Berlin. DOI: https://doi.org/10/W16-2507

Gladkova, Anna, Aleksandr Drozd, and Satoshi Matsuoka. 2016. Analogy-based detection of morphological and semantic relations with word embeddings: What works and what doesn't. In *Proceedings of the NAACL Student Research Workshop*, pages 8–15, San Diego, CA. DOI: https://doi.org/10.18653/v1/N16-2002

Goldstone, Robert. 1994. An efficient method for obtaining similarity data. *Behavior Research Methods, Instruments, & Computers*, 26(4):381–386. DOI: https://doi.org/10.3758/BF03204653

Gower, John C. 1966. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53(3–4):325–338. DOI:

https://doi.org/10.1093/biomet/53.3-4.325

Grave, Edouard, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, pages 3483–3487, Miyazaki.

Hewitt, John and Christopher D. Manning. 2019. A structural probe for finding syntax in word 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)*, pages 4129–4138, Minneapolis, MN.

Hill, Felix, Roi Reichart, and Anna Korhonen. 2015. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695. DOI: https://doi.org/10.1162/COLI_a_00237

Hollis, Geoff and Chris Westbury. 2016. The principals of meaning: Extracting semantic dimensions from co-occurrence models of semantics. *Psychonomic Bulletin & Review*. 23(6):1744–1756. DOI: https://doi.org/10.3758/s13423-016-1053-2, PMID: 27138012

Hout, Michael C., Stephen D. Goldinger, and Ryan W. Ferguson. 2013. The versatility of SpAM: A fast, efficient, spatial method of data collection for multidimensional scaling. *Journal of Experimental Psychology: General*, 142(1):256. **DOI:** https://doi.org/10.1037/a0028860, PMID: 22746700, PMCID: PMC3465534

Huang, Eric H., Richard Socher, Christopher D. Manning, and Andrew Y. Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 873–882, Jeju Island.

Jackendoff, Ray. 1972. *Semantic Interpretation in Generative Grammar*. MIT Press.

Jarmasz, Mario and Stan Szpakowicz. 2003. Roget's thesaurus and semantic similarity. In *Recent Advances in Natural Language Processing III, Selected Papers from RANLP 2003*, pages 111–120. Borovets. DOI: https://doi.org/10.1075/cilt.260.12jar

Jawahar, Ganesh, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657. Florence. DOI: https://doi.org/10.18653/v1/P19-1356

Jurgens, David and Ioannis Klapaftis. 2013. SemEval-2013 Task 13: Word sense induction for graded and non-graded senses. In *Proceedings of SEMEVAL*, pages 290–299, Atlanta, GA.

Kacmajor, Magdalena and John D. Kelleher. 2020. Capturing and measuring thematic relatedness. *Language Resources and Evaluation*, 54:645–682. DOI: https://doi.org/10.1007/s10579-019-09452-w

Kipfer, Barbara Ann. 2009. *Roget's 21st Century Thesaurus (3rd Edition)*. Philip Lief Group.

Kipper, Karin, Anna Korhonen, Neville Ryant, and Martha Palmer. 2006. Extending VerbNet with novel verb classes. In *Proceedings of LREC*, pages 1027–1032, Genoa.

Kipper Schuler, Karin. 2005. em VerbNet: A Broad-coverage, Comprehensive Verb Lexicon. Ph.D. thesis, University of Pennsylvania.

Kiritchenko, Svetlana and Saif Mohammad. 2017. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *Proceedings of ACL*, pages 465–470, Vancouver. https://doi.org/10.18653/v1/P17-2074

Kiritchenko, Svetlana and Saif M. Mohammad. 2016. Capturing reliable fine-grained sentiment associations by crowdsourcing and best–worst scaling. In *Proceedings of NAACL-HLT*, pages 811–817, San Diego, CA. DOI: https://doi.org/10.18653/v1/N16-1095

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT Summit*, volume 5, pages 79–86, Citeseer.

Kriegeskorte, Nikolaus and Marieke Mur. 2012. Inverse MDS: Inferring dissimilarity structure from multiple item arrangements. *Frontiers in Psychology*, 3:245. DOI: https://doi.org/10.3389/fpsyg.2012.00245, PMID: 22848204, PMCID: PMC3404552

Kriegeskorte, Nikolaus, Marieke Mur, and Peter A. Bandettini. 2008. Representational similarity analysis—Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2(4):1–28. DOI: https://doi.org/10.3389/neuro.06.004.2008, PMID: 19104670, PMCID: PMC2605405

Kuznetsova, Alina, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi

Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, and Vittorio Ferrari. 2020. The Open Images Data set V4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*. 128(7):1956–1981. DOI: https://doi.org/10.1007/s11263-020-01316-z

Lakoff, George and Mark Johnson. 1999. *Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought*, volume 4. University of Chicago Press.

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

Lemaire, Benoit and Guy Denhiere. 2006. Effects of high-order co-occurrences on word semantic similarity. *Current Psychology Letters. Behaviour, Brain & Cognition*, 1(18). DOI: https://doi.org/10.4000/cpl.471

Levin, Beth. 1993. *English Verb Classes and Alternations: Preliminary Investigation*. University of Chicago Press.

Levine, Gary M., Jamin B. Halberstadt, and Robert L. Goldstone. 1996. Reasoning and the weighting of attributes in attitude judgments. *Journal of Personality and Social Psychology*, 70(2):230. DOI: https://doi.org/10.1037/0022-3514.70.2.230

Levy, Omer and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of ACL*, pages 302–308, Baltimore, MD. DOI: https://doi.org/10.3115/v1/P14-2050, PMID: 25270273

Li, Min, Jian-er Chen, Jian-xin Wang, Bin Hu, and Gang Chen. 2008. Modifying the DPClus algorithm for identifying protein complexes based on new topological structures. *BMC Bioinformatics*, 9(1):398. DOI: https://doi.org/10.1186/1471-2105-9-398, PMID: 18816408 PMCID: PMC2570695

Li, Min, Dongyan Li, Yu Tang, Fangxiang Wu, and Jianxin Wang. 2017. CytoCluster: A cytoscape plugin for cluster analysis and visualization of biological networks. *International Journal of Molecular Sciences*, 18(9):1880. DOI: https://doi.org/10.3390/ijms18091880, PMID: 28858211, PMCID: PMC5618529

Lin, Emilie L. and Gregory L. Murphy. 2001. Thematic relations in adults' concepts. *Journal of Experimental Psychology: General*, 130(1):3. DOI: https://doi.org/10.1037/0096-3445.130.1.3, PMID: 11293459

Liu, Nelson F., Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019a. 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)*, pages 1073–1094, Minneapolis, MN.

Liu, Qianchu, Diana McCarthy, Ivan Vulić, and Anna Korhonen. 2019b. Investigating cross-lingual alignment methods for contextualized embeddings with token-level evaluation. In *Proceedings of CoNLL*, pages 33–43, Hong Kong. DOI: https://doi.or/10.18653/v1/K19-1004

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019c. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Louviere, Jordan J., Terry N. Flynn, and Anthony Alfred John Marley. 2015. *Best-Worst Scaling: Theory, Methods and Applications*. Cambridge University Press. DOI: https://doi.org/10.1017/CBO9781107337855

Louviere, Jordan J. and George G. Woodworth. 1991. Best-worst scaling: A model for the largest difference judgments. Technical report, University of Alberta.

Lupker, Stephen J. 1984. Semantic priming without association: A second look. *Journal of Verbal Learning and Verbal Behavior*, 23(6):709–733. DOI: https://doi.org/10.1016/S0022-5371(84)90434-1

Majewska, Olga, Diana McCarthy, Jasper van den Bosch, Nikolaus Kriegeskorte, Ivan Vulić, and Anna Korhonen. 2020. Spatial multi-arrangement for clustering and multi-way similarity data set construction. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5749–5758, Marseille.

Majewska, Olga, Diana McCarthy, Ivan Vulić, and Anna Korhonen. 2018a. Acquiring verb classes through bottom-up semantic verb clustering. In *Proceedings of LREC*, pages 952–958, Miyazaki, Japan.

Majewska, Olga, Ivan Vulić, Diana McCarthy, Yan Huang, Akira Murakami,

Veronika Laippala, and Anna Korhonen. 2018b. Investigating the cross-lingual translatability of VerbNet-style classification. *Language Resources and Evaluation*, 52(3):771–799. DOI: https://doi.org/10.1007/s10579-017-9403-x, PMID: 30956632, PMCID: PMC6428229

McCoy, Tom, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages, 3428–3448, Florence DOI: https://doi.org/10.18653/v1/P19-1334

McRae, Ken, Todd R. Ferretti, and Liane Amyote. 1997. Thematic roles as verb-specific concepts. *Language and Cognitive Processes*, 12(2–3):137–176. DOI: https://doi.org/10.1080/016909697386835

McRae, Ken, Saman Khalkhali, and Mary Hare. 2012, Semantic and associative relations in adolescents and young adults: Examining a tenuous dichotomy. In V. F. Reyna, S. B. Chapman, M. R. Dougherty, and J. Confrey, editors, *The Adolescent Brain: Learning, Reasoning, and Decision Making*, American Psychological Association, pages 39–66.

Meila, Marina and Jianbo Shi. 2001. A random walks view of spectral segmentation. In *Proceedings of AI and STATISTICS (AISTATS) 2001*, pages 177–182, Key West, FL.

Mikolov, Tomas, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Mikolov, Tomas, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of LREC*, pages 52–55, Miyazaki.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119 Lake Tahoe, NV.

Milajevs, Dmitrijs and Sascha Griffiths. 2016. A proposal for linguistic similarity data sets based on commonality lists. In *Proceedings of REPEVAL*, pages 127–133, Berlink. DOI: https://doi.org/10.18653/v1/W16-2523

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

Mnih, Andriy and Koray Kavukcuoglu. 2013. Learning word embeddings efficiently with noise-contrastive estimation. In *Proceedings of NIPS*, pages 2265–2273, Brussels.

Mrkšić, Nikola, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. Counter-fitting word vectors to linguistic constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 142–148, San Diego, CA. DOI: https://doi.org/10.18653/v1/N16-1018

Mrkšić, Nikola, Ivan Vulić, Diarmuid Ó Séaghdha, Ira Leviant, Roi Reichart, Milica Gašić, Anna Korhonen, and Steve Young. 2017. Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints. *Transactions of the Association for Computational Linguistics*, 5:309–324. DOI: https://doi.org/10.1162/tacl_a_00063

Mur, Marieke, Mirjam Meys, Jerzy Bodurka, Rainer Goebel, Peter A. Bandettini, and Nikolaus Kriegeskorte. 2013. Human object-similarity judgments reflect and transcend the primate-IT object representation. *Frontiers in Psychology*, 4:128. DOI: https://doi.org/10.3389/fpsyg.2013.00128, PMID: 23525516, PMCID: PMC3605517

Nepusz, Tamás, Haiyuan Yu, and Alberto Paccanaro. 2012. Detecting overlapping protein complexes in protein-protein interaction networks. *Nature Methods*, 9(5):471. DOI: https://doi.org/10.1038/nmeth.1938, PMID: 22426491, PMCID: PMC3543700

Newman, Mark E. J. 2005. A measure of betweenness centrality based on random walks. *Social Networks*, 27(1):39–54. DOI: https://doi.org/10.1016/j.socnet.2004.11.009

Nili, Hamed, Cai Wingfield, Alexander Walther, Li Su, William Marslen-Wilson, and Nikolaus Kriegeskorte. 2014. A toolbox for representational similarity analysis. *PLoS Computational Biology*, 10(4):e1003553. DOI: https://doi.org/10.1371/journal.pcbi

.1003553, PMID: 24743308, PMCID: PMC3990488

Ó Séaghdha, Diarmuid and Ann Copestake. 2008. Semantic classification with distributional kernels. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 649–656, Manchester. DOI: https://doi.org/10.3115/1599081.1599163

Pennington, Jeffrey, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha. DOI: https://doi.org/10.3115/v1/D14-1162

Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL-HLT*, pages 2227–2237, New Orleans, LA. DOI: https://doi.org/10.18653/v1/N18-1202

Pilehvar, Mohammad Taher and Jose Camacho-Collados. 2019. WiC: The word-in-context data set for evaluating context-sensitive meaning 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)*, pages 1267–1273, Minneapolis, MN.

Pilehvar, Mohammad Taher, Dimitri Kartsaklis, Victor Prokhorov, and Nigel Collier. 2018. Card-660: Cambridge Rare Word Data set—a reliable benchmark for infrequent word representation models. In *Proceedings of EMNLP*, pages 1391–1401, Brussels. DOI: https://doi.org/10.18653/v1/D18-1169

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).

Resnik, Philip. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of IJCAI*, pages 448–453, Montreal.

Resnik, Philip and Mona T. Diab. 2000. Measuring verb similarity. In *Proceedings of the 22nd Annual Meeting of the Cognitive Science Society (CogSci 2000)*, pages 399–404, Philadelphia, PA.

Rogers, Anna, Olga Kovaleva, Matthew Downey, and Anna Rumshisky. 2020. Getting closer to AI complete question answering: A set of prerequisite real tasks. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020*, pages 8722–8731, New York. DOI: https://doi.org/10.1609/aaai.v34i05.6398

Rogers, Anna, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *arXiv preprint arXiv:2002.12327*. DOI: https://doi.org/10.1162/tacl_a_00349

Sauppe, Sebastian. 2016. Verbal semantics drives early anticipatory eye movements during the comprehension of verb-initial sentences. *Frontiers in Psychology*, 7:95. DOI: https://doi.org/10.3389/fpsyg.2016.00095

Scarton, Carolina, Lin Sun, Karin Kipper-Schuler, Magali Sanches Duran, Martha Palmer, and Anna Korhonen. 2014. Verb clustering for Brazilian Portuguese. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages, 25–39, Kathmandu. DOI: https://doi.org/10.1007/978-3-642-54906-9_3

Schober, Patrick, Christa Boer, and Lothar A. Schwarte. 2018. Correlation coefficients: Appropriate use and interpretation. *Anesthesia & Analgesia*, 126(5):1763–1768. DOI: https://doi.org/10.1213/ANE.0000000000002864, PMID: 29481436

Schwartz, Roy, Roi Reichart, and Ari Rappoport. 2015. Symmetric pattern based word embeddings for improved word similarity prediction. In *Proceedings of CoNLL*, pages 258–267, Beijing, China. DOI: https://doi.org/10.18653/v1/K15-1026

Shannon, Paul, Andrew Markiel, Owen Ozier, Nitin S. Baliga, Jonathan T. Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. 2003. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11):2498–2504. DOI: https://doi.org/10.1101/gr.1239303, PMID: 14597658, PMCID: PMC403769

Si, Chenglei, Shuohang Wang, Min-Yen Kan, and Jing Jiang. 2019. What does BERT learn from multiple-choice reading comprehension data sets? *arXiv preprint arXiv:1910.12391*.

Sugawara, Saku, Pontus Stenetorp, Kentaro Inui, and Akiko Aizawa. 2020. Assessing the benchmarking capacity of machine reading comprehension data sets. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020*, pages 8918–8927, AAAI Press, New York, New York. DOI: https://doi.org/10.1609/aaai.v34i05.6422

Sun, Lin and Anna Korhonen. 2009. Improving verb clustering with automatically acquired selectional preferences. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, pages 638–647, Singapore. DOI: https://doi.org/10.3115/1699571.1699596

Sun, Lin, Anna Korhonen, Thierry Poibeau, and Cédric Messiant. 2010. Investigating the cross-linguistic potential of VerbNet-style classification. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1056–1064, Beijing, China.

Talmy, Leonard. 1985. Lexicalization patterns: Semantic structure in lexical forms. *Language Typology and Syntactic Description*, 3(99):36–149.

Tenney, Ian, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. DOI: https://doi.org/10.18653/v1/P19-1452

Turner, John C., Michael A. Hogg, Penelope J. Oakes, Stephen D. Reicher, and Margaret S. Wetherell. 1987. *Rediscovering the Social Group: A Self-Categorization Theory.* Basil Blackwell.

Turney, Peter D. 2001. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of ECML*, pages 491–502, Freiburg, Germany. DOI: https://doi.org/10.1007/3-540-44795-4_42

Turney, Peter D. 2006. Similarity of semantic relations. *Computational Linguistics*, 32(3):379–416. DOI: https://doi.org/10.1162/coli.2006.32.3.379

Turney, Peter D. and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188. DOI: https://doi.org/10.1613/jair.2934

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

Vulić, Ivan, Douwe Kiela, and Anna Korhonen. 2017. Evaluation by association: A systematic study of quantitative word association evaluation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 163–175, Valencia. DOI: https://doi.org/10.18653/v1/E17-1016

Vulić, Ivan and Anna Korhonen. 2018. Injecting lexical contrast into word vectors by guiding vector space specialisation. In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 137–143, Melbourne, Australia. DOI: https://doi.org/10.18653/v1/W18-3018

Vulić, Ivan, Nikola Mrkšić, and Anna Korhonen. 2017. Cross-lingual induction and transfer of verb classes based on word vector space specialisation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2546–2558, Copenhagen, Denmark. DOI: https://doi.org/10.18653/v1/D17-1270

Vulić, Ivan, Simon Baker, Edoardo Maria Ponti, Ulla Petti, Ira Leviant, Kelly Wing, Olga Majewska, Eden Bar, Matt Malone, Thierry Poibeau, Roi Reichart, and Anna Korhonen. 2020. Multi-SimLex: A large-scale evaluation of multilingual and cross-lingual lexical semantic similarity. *CoRR*, abs/2003.04866. DOI: https://doi.org/10.1162/coli_a_00391

Wang, Jianxin, Min Li, Jianer Chen, and Yi Pan. 2011. A fast hierarchical clustering algorithm for functional modules discovery in protein interaction networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(3):607–620. DOI: https://doi.org/10.1109/TCBB.2010.75, PMID: 20733244

Wang, Jianxin, Jun Ren, Min Li, and Fang-Xiang Wu. 2012. Identification of hierarchical and overlapping functional modules in PPI networks. *IEEE Transactions on NanoBioscience*, 11(4):386–393. DOI: https://doi.org/10.1109/TNB.2012.2210907, PMID: 22955967

Wang, Yuxuan, Wanxiang Che, Jiang Guo, Yijia Liu, and Ting Liu. 2019. Cross-lingual BERT transformation for zero-shot dependency parsing. In *Proceedings of the*

2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 5721–5727, Hong Kong. DOI: https://doi.org/10.18653/v1/D19-1575, PMCID: PMC6915745

Wieting, John, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. From paraphrase database to compositional paraphrase model and back. *Transactions of the ACL*, 3:345–358. DOI: https://doi.org/10.1162/tacl_a_00143, https://doi.org/10.1162/tacl_a_00246

Wieting, John, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2016. CHARAGRAM: Embedding words and sentences via character n-grams. In *Proceedings of EMNLP*, pages 1504–1515, Austin, Texas. DOI: https://doi.org/10.18653/v1/D16-1157

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv*, abs/1910.03771. DOI: https://doi.org/10.18653/v1/2020.emnlp-demos.6, PMCID: PMC7365998

Yang, Dongqiang and David M. W. Powers. 2006. Verb similarity on the taxonomy of WordNet. *Proceedings of the 3rd International WordNet Conference (GWC-06)*, pages 121–128, Jeju Island.

Yang, Zhilin, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R. Salakhutdinov, and Quoc V. Le. 2019, XLNet: Generalized autoregressive pretraining for language understanding, In Wallach, H., H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, Curran Associates, Inc., pages 5753–5763.

Zellers, Rowan, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence. DOI: https://doi.org/10.18653/v1/P19-1472