Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 5749–5758
Marseille, 11–16 May 2020

Spatial Multi-Arrangement for Clustering and Multi-way Similarity Dataset Construction

Olga Majewska[1], Diana McCarthy[1], Jasper van den Bosch[2], Nikolaus Kriegeskorte[3], Ivan Vulić[1], Anna Korhonen[1]

[1] University of Cambridge, Dept. of Theoretical and Applied Linguistics, Cambridge, CB3 9DA, United Kingdom
[2] University of Birmingham, School of Psychology, Edgbaston, Birmingham, B15 2TT, United Kingdom
[3] Columbia University, Jerome L. Greene Science Center, 3227 Broadway, New York, NY 10027, United States
{om304, iv250, alk23}@cam.ac.uk, diana@dianamccarthy.co.uk, vandejjf@bham.ac.uk, nk2765@columbia.edu

Abstract

We present a novel methodology for fast bottom-up creation of large-scale semantic similarity resources to support development and evaluation of NLP systems. Our work targets verb similarity, but the methodology is equally applicable to other parts of speech. Our approach circumvents the bottleneck of slow and expensive manual development of lexical resources by leveraging semantic intuitions of native speakers and adapting a spatial multi-arrangement approach from cognitive neuroscience, used before only with visual stimuli, to lexical stimuli. Our approach critically obtains judgments of word similarity in the context of a set of related words, rather than of word pairs in isolation. We also handle lexical ambiguity as a natural consequence of a two-phase process where verbs are placed in broad semantic classes prior to the fine-grained spatial similarity judgments. Our proposed design produces a large-scale verb resource comprising 17 relatedness-based classes and a verb similarity dataset containing similarity scores for 29,721 unique verb pairs and 825 target verbs, which we release with this paper.

Keywords: Lexicon, Lexical Database, Semantics, Crowdsourcing

1. Introduction

Verbs pose a particular challenge to machine interpretation of sentence meaning due to their complex linguistic properties. The role as sentence pivots, encoding crucial information about the structural and semantic relationships between the elements of the clause, assigns a lot of weight to verbs as carriers of information. This is why accurate, nuanced analysis and representation of their meaning is especially important for NLP systems to get closer to human levels of language understanding (Jackendoff, 1972; Levin, 1993; McRae et al., 1997; Altmann and Kamide, 1999; Resnik and Diab, 2000; Ferretti et al., 2001; Sauppe, 2016, *inter alia*).

The demand for verb-specific resources to support NLP has been recognised in recent years, as reflected in the publication of a large verb similarity dataset for English, SimVerb-3500 (hereafter SimVerb) (Gerz et al., 2016). However, the need for high-quality, wide-coverage lexical resources targeting verb semantics has by no means been satisfied. Rich lexical resources encoding information about verbs' semantic properties such as FrameNet (Baker et al., 1998) or VerbNet (Kipper Schuler, 2005; Kipper et al., 2006) are still unavailable for most languages, and evaluation datasets dedicated to or dominated by nouns are by far predominant (Finkelstein et al., 2002; Agirre et al., 2009; Bruni et al., 2012; Hill et al., 2015). Therefore, we propose methodology aimed at alleviating the evaluation data scarcity problem and overcoming the bottleneck of manual gold standard creation.

We present a novel approach to obtaining semantic similarity data by means of a two-phase design consisting in (1) *bottom-up semantic clustering* of verbs into relatedness-based classes and (2) *spatial similarity judgments* obtained via a multi-arrangement method so far employed only in psychology and cognitive neuroscience research and with visual stimuli (Kriegeskorte and Mur, 2012; Mur et al., 2013; Charest et al., 2014). We show how it can be adapted for the purposes of a large-scale linguistic task with *polysemous lexical stimuli* and used to obtain verb similarity data. The promise of this method lies in the intuitive nature of the task (i.e., relative similarities between items are signaled by the geometric distances within a 2D arena) and a user-friendly drag-and-drop interface. This significantly facilitates and speeds up the task, as many similarity judgments are expressed with a single mouse drag. Moreover, no structure or criteria are pre-imposed on the annotators, and similarities between individual verbs are judged *in the context of other verbs* appearing in the arena, rather than in isolated pairs. Crucially, the method allows for clustering of verbs and pairwise semantic similarity ratings at the same time, which can be of great benefit in NLP as a means of rapid creation of evaluation data.

We make available 17 relatedness-based classes and *SpA-Verb*, a large intrinsic evaluation dataset including 29,721 unique pairwise verb (dis)similarity scores for 825 target verbs.[1] The number of pairwise scores surpasses the largest existing verb-specific evaluation resource, SimVerb with 3,500 pairwise scores, by a significant margin. We release the data at the following link: `https://github.com/om304/SpA-Verb`.

2. Related Work

Recent years have seen word representation learning take center stage in NLP, with novel architectures pushing performance to new heights. Further advances rely on the

[1]The scores are Euclidean distances corresponding to dissimilarities between words, with smaller scores for similar verbs and larger scores for dissimilar verbs. The total number of pairwise scores results from aggregating the scores recorded for all possible unique pairings of verbs in each of the 17 classes (see Table 1).

availability of high-quality evaluation resources, which are still limited, and few and far between. Rich expert-created resources such as WordNet (Miller, 1995; Fellbaum, 1998), VerbNet (Kipper Schuler, 2005; Kipper et al., 2006), or FrameNet (Baker et al., 1998) encode a wealth of semantic, syntactic and predicate-argument information for English words, but are expensive and time-consuming to create. Crowd-sourcing with non-expert annotators has been adopted as a quicker alternative to produce evaluation benchmarks. Semantic models have been predominantly evaluated on datasets consisting of human similarity ratings collected for sets of word pairs (Baroni et al., 2014; Levy and Goldberg, 2014; Pennington et al., 2014; Dhillon et al., 2015; Schwartz et al., 2015; Wieting et al., 2016; Bojanowski et al., 2017; Mrkšić et al., 2017).

Various views of what constitutes 'semantic similarity' between words have been adopted, and it is undecided what kind of meaning relationship word embeddings should capture. The term *semantic relatedness* has been employed to refer to words linked by any kind of semantic relation (Budanitsky and Hirst, 2001; Budanitsky and Hirst, 2006; Turney and Pantel, 2010), including synonyms (*baffle-perplex*), meronyms and holonyms (*finger-hand*) or antonyms (*soft-hard*). Similarity defined as association, i.e., the mental activation of a term when another is presented (Chiarello et al., 1990; Lemaire and Denhiere, 2006), e.g., *knife-murder*, has been estimated in terms of frequency of co-occurrence of words in language (and the physical world) (Turney, 2001; Turney and Pantel, 2010; McRae et al., 2012; Bruni et al., 2012). In contrast to *associative relatedness*, a concept of semantic similarity defined in terms of shared superordinate category (Lupker, 1984; Resnik, 1995) (*taxonomical similarity* (Turney and Pantel, 2010)) or shared semantic features (Tversky, 1977; Frenck-Mestre and Bueno, 1999; Turney, 2006) has been proposed. Here, similarity is quantified in terms of degree of overlap in semantic properties, e.g., shared function or physical features, with synonyms occupying the top region of the similarity scale (e.g. *fiddle-violin* (Cruse, 1986)). In this work, we reserve the term (semantic) similarity for this latter definition of closeness of meaning, and distinguish it from the more general relatedness, which also includes association, as in previous work (Resnik, 1995; Resnik and Diab, 2000; Agirre et al., 2009; Hill et al., 2015; Gerz et al., 2016). 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.

Despite their wide usefulness, most available datasets used for intrinsic evaluation in distributional semantics are restricted in size and coverage, many do not distinguish similarity and relatedness, and only a few target verbs in particular. The prominent word pair datasets include WordSim-353 (Finkelstein et al., 2002; Agirre et al., 2009), comprising 353 noun pairs, and SimLex-999 (Hill et al., 2015), comprising 999 word pairs out of which 222 are verb pairs. Verb-only datasets include YP-130 (Yang and Powers, 2006) (130 verb pairs) and the dataset of Baker et al. (2014) (143 verb pairs). A resource aimed at addressing the problem of insufficient verb-specific evaluation data is SimVerb (Gerz et al., 2016), providing pairwise similarity scores for 3,500 verb pairs.

Although pairwise rating datasets have been ubiquitous in intrinsic evaluation, alternative annotation approaches and dataset types have been proposed to address some of their limitations. These include *best-worst scaling* (Louviere and Woodworth, 1991; Louviere et al., 2015; Avraham and Goldberg, 2016; Kiritchenko and Mohammad, 2016; Kiritchenko and Mohammad, 2017; Asaadi et al., 2019), which relies on 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 annotators choose which of the two items has more of a given property). Models have also been evaluated on synonym detection datasets gathered via English as foreign or second language tests (Landauer and Dumais, 1997; Turney, 2001) and word games (Jarmasz and Szpakowicz, 2003), composed of 5-word tuples (one target word and 4 potential synonyms, only one correct), and on analogy (Mikolov et al., 2013; Gladkova et al., 2016) and semantic relation datasets (Baroni and Lenci, 2011).

The largest verb-focused dataset currently available, SimVerb, is a result of a crowd-sourcing effort involving over 800 raters, each completing the pairwise similarity rating task for 79 verb pairs. In this paper, we present an alternative novel approach which allows an annotator to implicitly express multiple pairwise similarity judgments by a single mouse drag, instead of having to consider each word pair independently. This lets us scale up the data collection and, starting from the same set of verbs as those used in SimVerb, generate similarity ratings for over 8 times as many verb pairs. This approach is coupled with a precursor method for creating relatedness-based item classes 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 used before to collect similarity judgments through geometric arrangements of visual stimuli in psychology and cognitive neuroscience (Goldstone, 1994; Levine et al., 1996; Kriegeskorte and Mur, 2012; Hout et al., 2013; Mur et al., 2013; Charest et al., 2014). However, its applicability to semantic similarity of lexical stimuli has not yet been explored. To the best of our knowledge, this is the first adaptation of SpAM to lexical stimuli.

In the pairwise rating method (e.g. used with 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. The inter-stimulus Euclidean distances represent pairwise dissimilarities. This set-up ensures that all stimuli are considered in the context of the entire sample. Each placement simultaneously communicates similarity relationship of the item to *all other items* in the set.

SpAM taps into the spatial nature of humans' mental representation of concept similarity (Lakoff and Johnson, 1999;

Gärdenfors, 2004; Casasanto, 2008). It allows for a freer, intuitive expression of similarity judgments as continuous distances, rather than requiring assignment of discrete numerical ratings. The latter, although ubiquitous in intrinsic evaluation of distributional semantic models, have a number of limitations (Batchkarov et al., 2016; Faruqui et al., 2016; Gladkova and Drozd, 2016; Kiritchenko and Mohammad, 2017). Human ratings of isolated pairs of words are likely to be biased by word frequency, prototypicality, order of presentation and speed of association, rather than reflecting semantic factors. At the same time, degrees of similarity and subtle meaning distinctions between words are very difficult to quantify and translate onto a discrete scale without context or points of reference. This leads to inconsistencies in annotations by the same rater or across raters. By allowing repeated multi-wise, relative continuous similarity judgments, SpAM addresses shortcomings of the absolute pairwise ratings to produce evaluation data capturing the complexity of lexical relations in continuous semantic space.

In this work, we adapt the multi-arrangement method proposed by Kriegeskorte and Mur (2012). It uses inverse multidimensional scaling to obtain a distance matrix from multiple spatial arrangements of subsets of items within a 2D space. The subsets of items displayed are designed by an adaptive algorithm aimed at providing optimal evidence for the dissimilarity estimates. The subject drags and drops the stimuli within a circular *arena* on the computer screen, placing items perceived as similar close together and those dissimilar further apart.

The first arrangement of all items within a sample provides an initial estimate of the representational dissimilarity matrix (RDM). The subject then continues work on subsets sampled from the entire stimuli set. The process can be terminated at any time after the first arrangement onward. The adaptive subset selection algorithm elicits repeated judgments on items placed close together in the previous trial to ensure enough evidence is gathered for the relative distances between the similar items and for each possible pairing. Therefore, an earlier termination entails a potentially noisier final RDM. For each arrangement the subject is instructed to use the entire space available. This spreads out items previously clustered together thus reducing bias from placement error. It is the relative inter-item distances, rather than the absolute screen distances, that represent dissimilarities between the items from trial to trial.

Adapting the underlying multi-arrangement approach for our purposes concerns two key challenges, previously unsolved by SpAM-based methods: *scalability* and *semantic ambiguity*. In cognitive science the approach has been used on fairly small stimuli sets ($<$100 items). In preliminary tests, we found that handling larger samples is technically and cognitively difficult for human subjects. First, the dimensions of the arena within which the items are arranged are restricted by the size and resolution of the computer screen (Figure 1). With 100+ items, the arena becomes overcrowded, which makes it difficult to distribute the items as required. With longer sessions, participant fatigue affects the quality and consistency of the judgments. Second, lexical stimuli exhibit semantic ambiguity. Without multiple sense labels, annotators consider different word senses and hence, differ

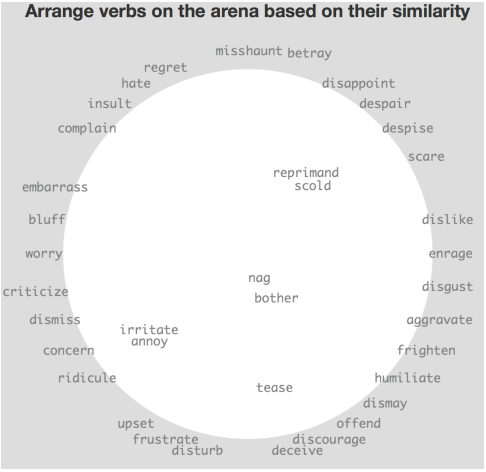

Figure 1: The arena layout (the first trial) with the complete class to be arranged displayed around the circle.

in their similarity judgments. In what follows, we propose a new SpAM-inspired framework that resolves both issues with scalability and semantic ambiguity. We show how these key challenges are addressed by our two-phase study design.

3.2. Two-Phase Design

First, in a *rough clustering phase* (Phase 1), the large sample is split into smaller, broad classes of semantically similar and related verbs. Second, in a *spatial multi-arrangement phase* (Phase 2) the verbs in the classes created in the previous phase are repeatedly arranged within the 2D space.

This solution allows us to overcome the challenges of ambiguity and scale mentioned in §3.1. It divides the large sample into manageable relatedness-based classes, which can be accommodated by most computer screens without a decrease in legibility. Furthermore, the two-phase set-up handles ambiguity by permitting copying verb labels to capture different senses in Phase 1. The rough clustering phase guarantees that each verb label 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).[2] The actual sense is implied by the surrounding words: this helps avoid mismatches in similarity judgments between participants for ambiguous verbs. What is more, this avoids the common problem of ambiguous low similarity scores (Milajevs and Griffiths, 2016) that conflate similarity judgments on antonyms (*vanish - appear*) and completely unrelated notions (*fry - appear*), and focuses on judgments between comparable concepts.

3.3. Data

To test the scaling-up potential of our approach and to enable direct comparisons with the standard pairwise similarity rating methods, we select the 827 verbs from SimVerb (Gerz et al., 2016) as our item sample.[3] The sample presents a challenge due to its size (i.e., it is almost nine times as

[2]Following Turner et al. (1987), '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.'

[3]Two verbs, *tote* and *pup*, were removed from the final sample due to their very low frequency, resulting in a 825-verb sample.

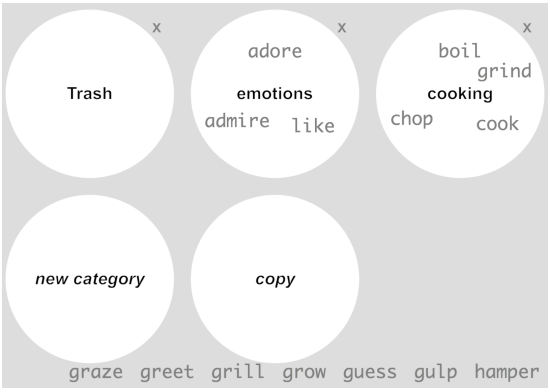

Figure 2: The rough clustering task layout (zoomed in). Verbs can be dragged onto the 'new category' circle to create a new grouping, onto 'copy' to create a duplicate label, or 'Trash' to dispose of the unwanted duplicate.

numerous as the biggest stimuli sets used in SpAM to date), and covers a wide range of verb meaning: each top-level VerbNet class is represented by at least 3 member verbs.

3.4. Interface and Task Structure

Our study was set up on an online platform which allows users to log in and out to save and resume tasks as required.[4] Phase 1 and Phase 2 were set up consecutively as separate experiments and participants were recruited for each individually. Both experiments had guidelines embedded in the task structure, available at any point, and included a short qualification task of 7 verbs (averaging 1.5 mins for Phase 1 and 7 mins for Phase 2) simulating the full experiment.

4. Phase 1: Rough Clustering

In Phase 1, 825 English verbs were classified into groups based on their meaning. The annotators were instructed to group similar and related words together. The exact number and size of the classes were left unspecified, but the guidelines instructed the participants to aim for broad categories of similar and related verbs, so as to end up with groups of roughly 30-50 words. Smaller or larger groupings were permitted if unavoidable (i.e., a smaller or larger number of verbs representing a coherent semantic category or 'theme' was identified).

The online interface contains a scrollable alphabetic list of 825 verbs at the bottom of the screen (Figure 2). The task consists of placing the verbs from the list one by one into empty circles representing clusters created by the user. Each circle acts as a container for a single grouping of similar and related verbs. If a single verb could be put in more than one group, the annotators were instructed to copy the verb label (as many times as needed) and put each in a different circle.[5] This allows handling of both polysemous and vague verbs.

[4] www.meadows-research.com

[5] This was illustrated with the verb *draw*, clusterable with art-related verbs like *paint* or verbs such as *pull* and *drag*.

4.1. Participants

The rough clustering task was first tested by two native English speakers. They produced clusters with an encouraging degree of overlap. It was computed using 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., 2018). B-Cubed, 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_i be the set of items in X_i but excluding i and K_i be the set of items in Y_i but excluding i. B-Cubed precision P and recall R are defined as:

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

$$R = \frac{1}{|U|} \sum_{i \in U} \frac{1}{|K_i|} \sum_{k \in K_i} \frac{min(|X_i \cap X_k|, |Y_i \cap Y_k|)}{|Y_i \cap Y_k|}$$

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

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

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

Next, 10 native speakers from the UK and the US, with a minimum undergraduate level of education, completed the task. It took 2.4 hours on average to complete the task across annotators. Between 10-67 clusters (27.5 on average) were produced, with an average of 12.3-82.5 verbs each.

4.2. Class Selection

The goal of Phase 1 was to obtain an average classification where membership and size of each class is determined by the intersection of the classes from individual annotators (the core), extended by additional valid member verbs on which there was partial agreement. These classes, subsequently used in Phase 2, were determined as follows. Clusters obtained from the verb pairings on which any 6+ participants (majority) agreed were used as a starting point and determined the semantics of the classes (e.g., 'movement', 'communication'). Post-processing was limited to (1) merging smaller semantically related clusters to obtain large, all-encompassing classes based on semantic relatedness, and (2) populating the thus created sets with the verbs missing from the majority classes based on their relatedness. These low-agreement verbs were reviewed and added manually to related classes by one of the authors.[6] The final number of

[6] Clusterability of Phase 1 verbs was guaranteed by balanced sampling from across different VerbNet classes (Gerz et al., 2016).

classes was 17.

5. Phase 2: Multi-Arrangement

In Phase 2, the spatial multi-arrangement task, each of the 17 verb classes was individually presented to the participant on the screen, in random order, around a *circular arena* (Figure 1). The guidelines instructed the participants to arrange verbs based on similarity of their meaning, dragging and dropping the verbs one by one onto the circle, putting similar words closer together and less similar ones further apart, with the relative positions and distances between the words reflecting the degree of similarity.

5.1. Participants

The minimum number of annotators set for each class was 10. Each annotator was asked to arrange at least 3 classes, presented in a random order. Participant recruitment was ongoing until the minimum number of annotators per class was satisfied. Overall, 40 native English speakers from the UK and the US, with minimum undergraduate education level, participated in the multi-arrangement task, producing a total of 314,137 individual pairwise judgments. For each class and annotator, we recorded 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

To ensure high quality of the resultant data, we discarded annotations where word placements were executed too fast in the first arrangement of each class (i.e., where the average time spent on moving a single verb was less than 1 second). Subsequently, for each arena 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. This criterion was the acceptability threshold used in the creation of SimLex (Hill et al., 2015).

For each class, we calculated the average of the Euclidean distances from all accepted annotators for each verb pair and obtained an average RDM, as shown in Figure 3. The averaged pairwise distances (= dissimilarity scores) in each class were then scaled to have a root mean square of 1, as 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 $d'_i, ..., d'_N$ 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):

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

In case of ambiguity, a verb could be added to several classes with semantically related members. Six out of 10 annotators used the copying functionality 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%.

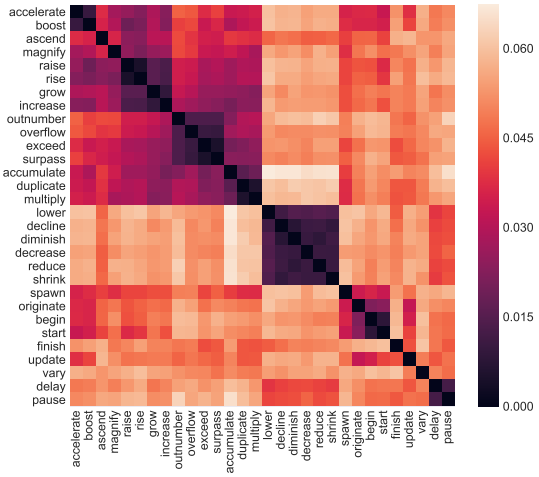

Figure 3: Average ordered dissimilarity matrix for one of the verb classes (dark-to-light color scale for small-to-large dissimilarities), with dark areas corresponding to clusters of similar verbs (e.g., *lower, decline, diminish, decrease, reduce, shrink*).

The final dataset collates the 17 classes and contains scaled averaged dissimilarity scores for 29,721 unique verb pairs.

6. Inter-Annotator Agreement

IAA after Phase 2 is measured based on Spearman's rank correlation coefficient (ρ): for each class, we compute the average correlation of an individual annotator with the average of all other annotators (*mean*) (Hill et al., 2015; Gerz et al., 2016), see Table 1.[7]

The flexibility of similarity judgments expressed by mouse drags in our setup, where pairwise similarity scores from an annotator differ by fractions based on the verbs' relative positions within the circular space, leaves much more room for divergence in scores across annotators than ordinal rating scales, nevertheless the obtained IAA scores (ρ^A) are

[7] We do not calculate IAA over the entire dataset as different annotators worked on different classes.

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

Table 1: 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.

promising.[8] Class size was the main factor affecting the difficulty of the task, as reflected in the agreement scores: we observe negative correlation between agreement and the number of verbs in a class (Spearman's $\rho = -0.67$).

The lowest IAA is observed for the 2nd (*#15*) largest class (87 verbs); higher IAA scores are reported with smaller classes (*#3,#1,#13*). However, we also note that the agreement on the largest class of 'movement verbs' (*#10*, 100 members) is higher than could be expected based on its size alone ($\rho = 0.26$). This is likely due to the fact that many 'movement verbs' have well-defined, concrete meanings, clusterable into smaller groupings (e.g. based on the medium, i.e., movement on land, in water, in the air). The lowest-agreement Class 15 was more heterogeneous, including such members as *add, dip, flush, spread*. The potential subcategories to which words belong or dimensions along which they differ were harder to define (as indicated in annotator feedback): their relative positions varied by participant.

Given the strong negative correlation between sample size and IAA, we examined whether higher IAA scores can be obtained on the same verbs pairs split into smaller samples in a follow-up study with 5 annotators. We randomly split the 87-word lowest-IAA class (15) into three equal 29-word subsets. Each annotator worked on the three subsets one by one, with breaks in between. The IAA on the smaller sets was lower than in the full-class (87-word) setting (average across the three subsets: $\rho = 0.098$, compared to $\rho = 0.19$ on the full class). This suggests that while big arenas are generally harder, the difficulty of the task has much to do with the verbs in the sample, and this class is especially challenging due to its heterogeneity (an issue further aggravated by randomly splitting the big set and potentially separating verbs clusterable together).

While similar verbs do end up placed together (e.g., *seize - snatch*, *smear - smudge*), there is greater variability in the distances between the less similar words. These results also indicate that decreasing the number of words to be arranged does not guarantee higher agreement, and being presented with a semantically clusterable bigger set of words (like the ones produced in Phase 1) may be preferable to imposing an arbitrary size limit on the classes. As a consequence of the difficulty of some verb sets, the inter-annotator agreement scores for some classes show low positive correlation. Therefore, evaluation of representation models would best be focused on classes with higher inter-annotator agreement and consequently clearer semantics.

7. SpAM vs Pairwise Ratings

We chose the SimVerb-3500 dataset for comparison of the pairwise rating-based approach with our spatial arrangement-based method as it is the most similar resource currently available due to its scale and sole focus on verbs. Since our verb sample is the same as SimVerb, we can compare our IAA per class with the IAA that we obtain on the verb pairs in that class also occurring in SimVerb. This is shown in ρ^{SV} of Table 1. Also, we computed Spearman's correlation between SimVerb similarity scores and our average pairwise distances on all shared verb pairs (1,682).

Despite the shared sample of verbs, the number of overlap pairs is reduced due to the differences between SimVerb and our design. In SimVerb, pairs are chosen to cover different degrees of relatedness, including completely unassociated pairs. Our Phase 1 separates the sample into classes based on relatedness, therefore the possible pairwise combinations of verbs are limited to related verbs. These differences are highlighted in Figure 4 which shows score distributions in both datasets. SimVerb has a peak at the 0-1 unrelated end of the distribution. These are the easy to annotate unrelated verb pairs which are filtered at Phase 1 in our approach.

The overlap sets are on average over one order of magnitude smaller than our respective complete classes. Moreover, the overlap pairs are more spread out in terms of degree of similarity compared to the complete classes. For each grouping of similar verbs within an arena, our dataset includes all the possible pairwise combinations, which results in many scores differing by small amounts. Only some of those pairs are included in SimVerb (e.g., out of Class 9 pairs *decide-choose*, *decide-select*, *decide-elect*, *decide-pick*, only the first one is present in SimVerb). These differences explain the lower correlation scores on most of the entire classes with respect 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 task.[9] While these datasets are produced by different paradigms, there is still a reasonable level of correlation between the two resources on shared pairs: $\rho = 0.62$.

We did not give explicit guidance on how to treat antonyms and how to distinguish between related and semantically similar words. We observe that our Phase 1 set-up tends to encourage placing 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).[10] However, in Phase 2 antonyms are predominantly kept apart: out of 67 antonymy pairs shared with SimVerb (i.e., pairs labeled ANTONYMS in SimVerb), only 2 are placed closer in the arena (*inhale - exhale* and *sink - swim*). This tendency is also illustrated by the RDM in Figure 3: separate clusters are formed by verbs such as *raise, rise, grow* and *diminish, decline, lower*, and *finish* is kept separate from *begin* and *start*.

Crucially, our spatial approach records simultaneous judgments on multiple related words, which helps improve judgment consistency (e.g., word pairs holding analogous relations have similar scores) and allows making subtle distinctions based on varying degrees of similarity by means of

[8] Notably, the reported scores compare favourably with inter-subject correlations reported for spatial multiple arrangements of concrete visual stimuli (real-world objects) in cognitive neuroscience research (Cichy et al., 2019), where Spearman's correlation scores are in the range of approx. 0.12-0.21 and are considered high ($p < 0.001$)).

[9] The ρ^{SV} scores are promising compared to the $\rho = 0.612$ SimVerb IAA (Pilehvar et al., 2018), despite the fact that the easy cases of verb pairs involving very disparate verbs (in different classes) are not included in our results.

[10] There are exceptions: positive (e.g. *love*) and negative (e.g. *hate*) emotion verbs form two different classes; there are also separate groupings with 'construction' and 'destruction' verbs. See Table 1.

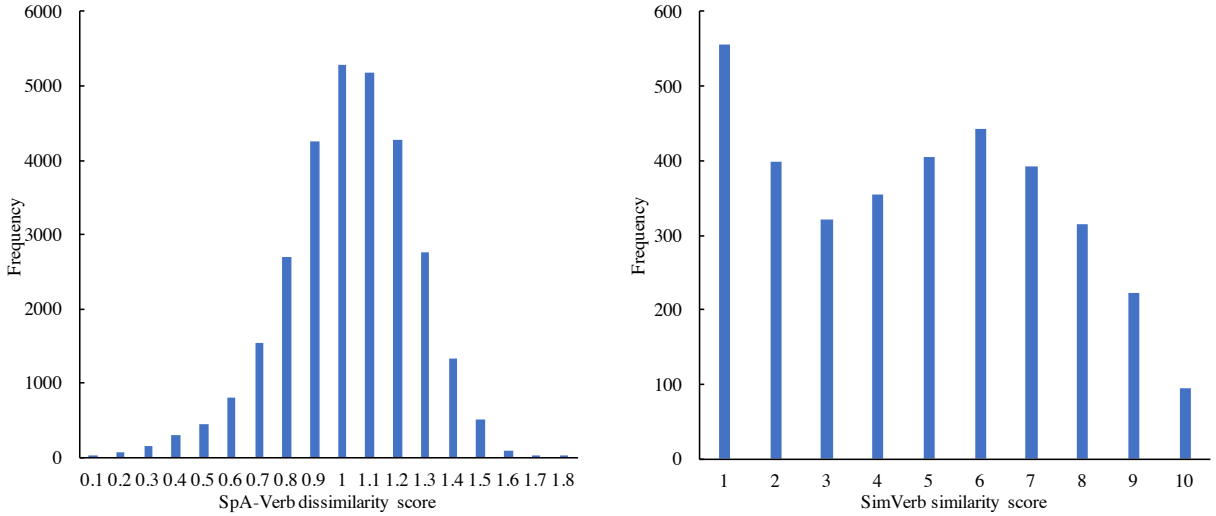

Figure 4: Score distribution for SpA-Verb (dissimilarities are Euclidean distances in the arena) and SimVerb (ratings on a 0-10 interval) in terms of frequency of each score interval (i.e., the number of pairwise scores falling within a given score interval in each dataset). Each score interval label gives the upper bound.

small adjustments of relative distances between all members of a class. For example, word pairs linked by the troponymy relation in WordNet are placed at similar distances in the arena (e.g., *bounce - jump* receive score 0.375, *bounce - spring* 0.373). In SimVerb, where pairs are rated in isolation, *bounce - spring* have a high 8.80 rating, but *bounce - jump* receive a lower score of 6.81. Similarly, pairs of WordNet synonyms *jump - leap* and *jump - skip*, which are placed close together in the arena (scores 0.343 and 0.277), receive quite divergent scores in SimVerb (9.63 and 5.48, respectively). Notably, the score of 5.48 is the similarity rating of *embarrass - blush*, which are strongly associated but dissimilar, while *jump - skip* display a high degree of semantic overlap (i.e., describe a similar kind of motion). Arranging groups of words simultaneously in the same space helps adjust relative similarities based on other words present, without the need to refer back to a previously given score, which is not necessarily possible when word pairs are judged in small batches.

Moreover, by eliciting simultaneous judgments on multiple lexical items we can significantly speed up the data collection process. As an example, with our SpAM-based approach 60 minutes of work of a single annotator produces pairwise similarity scores for 4,950 unique verb pairs. In the pairwise approach used for SimVerb, it would take over 8 hours for a single rater to record the same number of similarity judgments (approx. 8 minutes to complete 79 questions by a single participant (Gerz et al., 2016)). Our two-phase design and the modular nature of the task make it particularly appropriate for crowd-sourcing.

8. Conclusion and Future Work

We presented a new method for bottom-up, large-scale collection of semantic similarity data based on spatial arrangements of lexical items. We adapted the spatial approach, previously used only with visual stimuli, to polysemous lexical items in a large-scale setting. We applied this approach to a word sample almost nine times as numerous as the biggest stimuli sets used in SpAM-based research to date. The 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, can be readily applied to other parts of speech and types of stimuli. Crucially, the method produces both semantic clusters and word pair scores within an integrated framework. Moreover, the two-phase design enabled us to handle lexical ambiguity as a natural consequence of overlap in the first rough clustering phase. Our approach captures non-expert intuitions about word meaning, allowing fine-grained linguistic distinctions by considering the semantics of multiple lemmas together that elude simple pairwise similarity judgments. Furthermore, the method is easily portable to other languages, demonstrating potential for faster creation of evaluation datasets to support multilingual NLP.

Our method yielded SpA-Verb, a dataset of fine-grained similarity scores for 29,721 unique verb pairs, together with 17 relatedness-based verb classes, released online along with the data collection guidelines.

The scale of SpA-Verb offers many possibilities for robust analyses on semantically related classes allowing for better informed tuning and comparison of the adequacy and potential of representation learning architectures to capture semantic distinctions present in the mental lexicon, while helping achieve greater model interpretability. In future work, in order to examine the properties of our dataset and its potential as an evaluation resource, we will evaluate 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 Phase 2 and selected subsets with different semantic characteristics. Moreover, we will carry out in-depth qualitative and quantitative analyses of the information captured by each stage of our design, in comparison with existing lexical-semantic resources. Future work will also involve model evaluation on a verb classification task on clusters extracted from Phase 2 distance matrices, to assess 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. To investigate the method's portability, we will carry out data collection for other parts of speech and conduct experiments in other typologically diverse languages to analyse cross-linguistic similarities and variation. The data and annotation guidelines are available at the following link: `https://github.com/om304/SpA-Verb`.

9. Acknowledgements

We gratefully acknowledge the funding support of the Economic and Social Research Council [PhD Award Number ES/J500033/1] and the European Research Council (ERC) [Consolidator Grant 648909)].

10. Bibliographical References

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Pasca, M., and Soroa, A. (2009). A study on similarity and relatedness using distributional and WordNet-based approaches. In *Proceedings of NAACL-HLT*, pages 19–27.

Altmann, G. T. and Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, 73(3):247–264.

Amigó, E., Gonzalo, J., Artiles, J., and Verdejo, F. (2009). A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval*, 12(4):461–486.

Asaadi, S., Mohammad, S., and Kiritchenko, S. (2019). Big BiRD: A large, fine-grained, bigram relatedness dataset for examining semantic composition. In *Proceedings of NAACL-HLT*, pages 505–516.

Avraham, O. and Goldberg, Y. (2016). Improving reliability of word similarity evaluation by redesigning annotation task and performance measure. In *Proceedings of REPEVAL*, pages 106–110.

Bagga, A. and Baldwin, B. (1998). Algorithms for scoring coreference chains. In *Proceedings of LREC*, pages 563–566.

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The Berkeley FrameNet project. In *Proceedings of COLING*.

Baker, S., Reichart, R., and Korhonen, A. (2014). An unsupervised model for instance level subcategorization acquisition. In *Proceedings of EMNLP*, pages 278–289.

Baroni, M. and Lenci, A. (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.

Baroni, M., Dinu, G., and Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of ACL*, pages 238–247.

Batchkarov, M., Kober, T., Reffin, J., Weeds, J., and Weir, D. (2016). A critique of word similarity as a method for evaluating distributional semantic models. In *Proceedings of REPEVAL*, pages 7–12.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the ACL*, 5:135–146.

Bruni, E., Boleda, G., Baroni, M., and Tran, N. K. (2012). Distributional semantics in technicolor. In *Proceedings of ACL*, pages 136–145.

Budanitsky, A. and Hirst, G. (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.

Budanitsky, A. and Hirst, G. (2006). Evaluating WordNet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47.

Casasanto, D. (2008). Similarity and proximity: When does close in space mean close in mind? *Memory & Cognition*, 36(6):1047–1056.

Charest, I., Kievit, R. A., Schmitz, T. W., Deca, D., and Kriegeskorte, N. (2014). Unique semantic space in the brain of each beholder predicts perceived similarity. *Proceedings of the National Academy of Sciences*, 111(40):14565–14570.

Chiarello, C., Burgess, C., Richards, L., and Pollock, A. (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.

Cichy, R. M., Kriegeskorte, N., Jozwik, K. M., van den Bosch, J. J., and Charest, I. (2019). The spatiotemporal neural dynamics underlying perceived similarity for real-world objects. *NeuroImage*, 194:12–24.

Cruse, D. A. (1986). *Lexical Semantics*. Cambridge University Press.

Dalitz, C. and Bednarek, K. E. (2016). Sentiment lexica from paired comparisons. In *Proceedings of ICDM*, pages 924–930.

Dhillon, P. S., Foster, D. P., and Ungar, L. H. (2015). Eigenwords: Spectral word embeddings. *Journal of Machine Learning Research*, 16:3035–3078.

Faruqui, M., Tsvetkov, Y., Rastogi, P., and Dyer, C. (2016). Problems with evaluation of word embeddings using word similarity tasks. In *Proceedings of REPEVAL*, pages 30–35.

Christiane Fellbaum, editor. (1998). *WordNet: An Electronic Lexical Database*. MIT Press.

Ferretti, T. R., McRae, K., and Hatherell, A. (2001). Integrating verbs, situation schemas, and thematic role concepts. *Journal of Memory and Language*, 44(4):516–547.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., and Ruppin, E. (2002). Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, 20(1):116–131.

Frenck-Mestre, C. and Bueno, S. (1999). Semantic features and semantic categories: Differences in rapid activation of the lexicon. *Brain and Language*, 68(1-2):199–204.

Gärdenfors, P. (2004). *Conceptual Spaces: The Geometry of Thought*. MIT press.

Gerz, D., Vulić, I., Hill, F., Reichart, R., and Korhonen, A. (2016). SimVerb-3500: A large-scale evaluation set of verb similarity. In *Proceedings of EMNLP*, pages 2173–2182.

Gladkova, A. and Drozd, A. (2016). Intrinsic evaluations of word embeddings: What can we do better? In *Proceedings of REPEVAL*, pages 36–42.

Gladkova, A., Drozd, A., and Matsuoka, S. (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.

Goldstone, R. (1994). An efficient method for obtaining similarity data. *Behavior Research Methods, Instruments, & Computers*, 26(4):381–386.

Hill, F., Reichart, R., and Korhonen, A. (2015). SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Hout, M. C., Goldinger, S. D., and Ferguson, R. W. (2013). The versatility of SpAM: A fast, efficient, spatial method of data collection for multidimensional scaling. *Journal of Experimental Psychology: General*, 142(1):256.

Jackendoff, R. (1972). *Semantic Interpretation in Generative Grammar*. MIT Press.

Jarmasz, M. and Szpakowicz, S. (2003). Roget's thesaurus and semantic similarity. In *Recent Advances in Natural Language Processing III, Selected Papers from RANLP 2003, Borovets, Bulgaria*, pages 111–120.

Jurgens, D. and Klapaftis, I. (2013). SemEval-2013 Task 13: Word sense induction for graded and non-graded senses. In *Proceedings of SEMEVAL*, pages 290–299.

Kipper, K., Korhonen, A., Ryant, N., and Palmer, M. (2006). Extending VerbNet with novel verb classes. In *Proceedings of LREC*, pages 1027–1032.

Kipper Schuler, K. (2005). *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph.D. thesis, University of Pennsylvania.

Kiritchenko, S. and Mohammad, S. M. (2016). Capturing reliable fine-grained sentiment associations by crowdsourcing and best–worst scaling. In *Proceedings of NAACL-HLT*, pages 811–817.

Kiritchenko, S. and Mohammad, S. (2017). Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *Proceedings of ACL*, pages 465–470, July.

Kriegeskorte, N. and Mur, M. (2012). Inverse MDS: Inferring dissimilarity structure from multiple item arrangements. *Frontiers in Psychology*, 3:245.

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

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

Lemaire, B. and Denhiere, G. (2006). Effects of high-order co-occurrences on word semantic similarity. *Current Psychology Letters. Behaviour, Brain & Cognition*, 1(18).

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

Levine, G. M., Halberstadt, J. B., and Goldstone, R. L. (1996). Reasoning and the weighting of attributes in attitude judgments. *Journal of Personality and Social Psychology*, 70(2):230.

Levy, O. and Goldberg, Y. (2014). Dependency-based word embeddings. In *Proceedings of ACL*, pages 302–308.

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

Louviere, J. J., Flynn, T. N., and Marley, A. A. J. (2015). *Best-Worst Scaling: Theory, Methods and Applications*. Cambridge University Press.

Lupker, S. J. (1984). Semantic priming without association: A second look. *Journal of Verbal Learning and Verbal Behavior*, 23(6):709–733.

Majewska, O., McCarthy, D., Vulić, I., and Korhonen, A. (2018). Acquiring verb classes through bottom-up semantic verb clustering. In *Proceedings of LREC*.

McRae, K., Ferretti, T. R., and Amyote, L. (1997). Thematic roles as verb-specific concepts. *Language and Cognitive Processes*, 12(2-3):137–176.

McRae, K., Khalkhali, S., and Hare, M. (2012). Semantic and associative relations in adolescents and young adults: Examining a tenuous dichotomy. In V. F. Reyna, et al., editors, *The Adolescent Brain: Learning, Reasoning, and Decision Making*, pages 39–66. American Psychological Association.

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

Milajevs, D. and Griffiths, S. (2016). A proposal for linguistic similarity datasets based on commonality lists. In *Proceedings of REPEVAL*, pages 127–133.

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41.

Mrkšić, N., Vulić, I., Séaghdha, D. Ó., Leviant, I., Reichart, R., Gašić, M., Korhonen, A., and Young, S. (2017). Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints. *Transactions of the ACL*, 5:309–324.

Mur, M., Meys, M., Bodurka, J., Goebel, R., Bandettini, P. A., and Kriegeskorte, N. (2013). Human object-similarity judgments reflect and transcend the primate-IT object representation. *Frontiers in Psychology*, 4:128.

Pennington, J., Socher, R., and Manning, C. (2014). GloVe: Global vectors for word representation. In *Proceedings of EMNLP*, pages 1532–1543.

Pilehvar, M. T., Kartsaklis, D., Prokhorov, V., and Collier, N. (2018). Card-660: Cambridge Rare Word Dataset - a reliable benchmark for infrequent word representation models. In *Proceedings of EMNLP*, pages 1391–1401.

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

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

Sauppe, S. (2016). Verbal semantics drives early anticipatory eye movements during the comprehension of verb-initial sentences. *Frontiers in Psychology*, 7:95.

Schwartz, R., Reichart, R., and Rappoport, A. (2015). Symmetric pattern based word embeddings for improved word similarity prediction. In *Proceedings of CoNLL*, pages 258–267.

Turner, J. C., Hogg, M. A., Oakes, P. J., Reicher, S. D., and

Wetherell, M. S. (1987). *Rediscovering the Social Group: A Self-Categorization Theory.* Basil Blackwell.

Turney, P. D. and Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

Turney, P. D. (2001). Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of ECML*, pages 491–502.

Turney, P. D. (2006). Similarity of semantic relations. *Computational Linguistics*, 32(3):379–416.

Tversky, A. (1977). Features of similarity. *Psychological Review*, 84(4):327.

Wieting, J., Bansal, M., Gimpel, K., and Livescu, K. (2016). CHARAGRAM: Embedding words and sentences via character n-grams. In *Proceedings of EMNLP*, pages 1504–1515.

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