.

SimLex-999: Evaluating Semantic Models with (Genuine) Similarity Estimation

Felix Hill
Computer Laboratory
University of Cambridge
felix.hill@cl.cam.ac.uk

Roi Reichart
Faculty of Industrial Engineering
and Management
Technion, IIT
roiri@ie.technion.ac.il

Anna Korhonen
Computer Laboratory
University of Cambridge
anna.korhonen@cl.cam.ac.uk

We present SimLex-999, a gold standard resource for evaluating distributional semantic models that improves on existing resources in several important ways. First, in contrast to gold standards such as WordSim-353 and MEN, it explicitly quantifies similarity *rather than* association *or* relatedness *so that pairs of entities that are associated but not actually similar (Freud, psychology) have a low rating. We show that, via this focus on similarity, SimLex-999 incentivises the development of models with a different, and arguably wider range of applications than those which reflect conceptual association. Second, SimLex-999 contains a range of concrete and abstract adjective, noun and verb pairs, together with an independent rating of concreteness and (free) association strength for each pair. This diversity enables fine-grained analyses of the performance of models on concepts of different types, and consequently greater insight into how architectures can be improved. Further, unlike existing gold standard evaluations, for which automatic approaches have reached or surpassed the inter-annotator agreement ceiling, state-of-the-art models perform well below this ceiling on SimLex-999. There is therefore plenty of scope for SimLex-999 to quantify future improvements to distributional semantic models, guiding the development of the next generation of representation-learning architectures.*

1. Introduction

There is very little similar about coffee and cups. *Coffee* refers to a plant, which is a living organism or a hot brown (liquid) drink. In contrast, a *cup* is a man-made solid of broadly well-defined shape and size with a specific function relating to the consumption of liquids. Perhaps the only clear trait these concepts have in common is that they are concrete entities. Nevertheless, in what is currently the most popular evaluation gold standard for semantic similarity, WordSim(WS)-353 (**?**), *coffee* and *cup* are rated as more 'similar' than pairs such as *car* and *train*, which share numerous common properties (function, material, dynamic behaviour, wheels, windows etc.). Such anomalies also exist in other gold standards such as the MEN dataset (**?**). As a consequence, these

evaluations effectively penalise models for learning the evident truth that *coffee* and *cup* are dissimilar.

Although clearly different, *coffee* and *cups* are very much related. The psychological literature refers to the conceptual relationship between these concepts as *association*, although it has been given a range of names including *relatedness* (**?**; **?**), *topical similarity* (**?**) and *domain similarity* (**?**). Association contrasts with *similarity*, the relation connecting *cup* and *mug* (**?**). At its strongest, the similarity relation is exemplified by pairs of *synonyms*; words with identical referents.

Computational models that effectively capture similarity as distinct from association have numerous applications. Such models are used for the automatic generation of dictionaries, thesauri, ontologies and language correction tools (**?**; **?**; **?**). Machine translation systems, which aim to define mappings between fragments of different languages whose meaning is similar, but not necessarily associated, are another established application (**?**; **?**). Moreover, since, as we establish, similarity is a cognitively complex operation that can require rich, structured conceptual knowledge to compute accurately, similarity estimation constitutes an effective proxy evaluation for general-purpose representation-learning models whose ultimate application is variable or unknown (**?**; **?**).

As we show in Section 2, the predominant gold standards for semantic evaluation in NLP do not measure the ability of models to reflect similarity. In particular, in both WS-353 and MEN, pairs of words with associated meaning, such as *coffee* and *cup* (rating = 6.8) *telephone* and *communication* (7.5) or *movie* and *theater* (7.7), receive a high rating regardless of whether or not their constituents are similar. Thus, the utility of such resources to the development and application of similarity models is limited, a problem exacerbated by the fact that many researchers appear unaware of what their evaluation resources actually measure.[1]

While certain smaller gold standards, those of **?** (RG) and **?** (WS-Sim), do focus clearly on similarity, these resources suffer from other important limitations. For instance, as we show, and as is also the case for WS-353 and MEN, state-of-the-art models have reached the average performance of a human annotator on these evaluations. It is common practice in NLP to define the upper limit for automated performance on an evaluation as the average human performance or inter-annotator agreement (**?**; **?**; **?**). Based on this established principle and the current evaluations, it would therefore be reasonable to conclude that the problem of representation learning, at least for similarity modelling, is approaching resolution. However, circumstantial evidence suggests that distributional models are far from perfect. For instance, we are some way from automatically-generated dictionaries, thesauri or ontologies that can be used with the same confidence as their manually-created equivalents.

Motivated by these observations, in Section 3 we present *SimLex-999*, a gold standard resource for evaluating the ability of models to reflect similarity. SimLex-999 was produced by 500 paid native English speakers, recruited via Amazon Mechanical Turk,[2] who were asked to rate the similarity, as opposed to association, of concepts via a simple visual interface. The choice of evaluation pairs in SimLex-999 was motivated by empirical evidence that humans represent concepts of distinct part-of-speech (POS) (**?**) and conceptual concreteness (**?**) differently. While existing gold standards contain

---

1 For instance, **?**, pages 1,4,10 and **?**, page 4 refer to MEN and/or WS-353 as 'similarity datasets'. Others evaluate on both these association-based and genuine similarity-based gold standards with no reference to the fact that they measure different things (**?**; **?**).

2 `www.mturk.com/`

only concrete noun concepts (MEN) or cover only some of these distinctions via a random selection of items (WS-353, RG), SimLex-999 contains a principled selection of adjective, verb and noun concept pairs covering the full concreteness spectrum. This design enables more nuanced analyses of how computational models overcome the distinct challenges of representing concepts of these types.

In Section 4 we present quantitative and qualitative analyses of the SimLex-999 ratings, which indicate that participants found it unproblematic to quantify consistently the similarity of the full range of concepts and to distinguish it from association. Unlike existing datasets, SimLex-999 therefore contains a significant number of pairs, such as [*movie*, *theater*], which are strongly associated but receive low similarity scores.

The second main contribution of this paper, presented in Section 5, is the evaluation of state-of-the-art distributional semantic models using SimLex-999. These include the well known neural language models (NLMs) of **?**, **?** and **?**, which we compare with traditional vector-space co-occurrence models (VSMs) (**?**) with and without dimensionality reduction (SVD) (**?**). Our analyses demonstrate how SimLex-999 can be applied to uncover substantial differences in the ability of models to represent concepts of different types.

Despite these differences, the models we consider each share the characteristic of being better able to capture association than similarity. We show that the difficulty of estimating similarity is driven primarily by those strongly-associated pairs with a high (association) rating in gold standards such as WS-353 and MEN, but a low similarity rating in SimLex-999. As a result of including these challenging cases, together with a wider diversity of lexical concepts in general, current models achieve notably lower scores on SimLex-999 than on existing gold standard evaluations, and well below the SimLex-999 inter-human agreement ceiling.

Finally, we explore ways in which distributional models might improve on this performance in similarity modelling. To do so, we evaluate the models on the SimLex-999 subsets of adjectives, nouns and verbs, as well as on abstract and concrete subsets and subsets of more and less strongly associated pairs (Sections 5.2.2-5.2.4). As part of these analyses, we confirm the hypothesis (**?**; **?**) that models learning from input informed by dependency parsing, rather than simple running-text input, yield improved similarity estimation and, specifically, clearer distinction between similarity and association. In contrast, we find no evidence for a related hypothesis (**?**; **?**), that smaller context windows improve the ability of models to capture similarity. We do, however, observe clear differences in model performance on the distinct concept types included in SimLex-999. Taken together, these experiments demonstrate the benefit of the diversity of concepts included in SimLex-999; it would not have been possible to derive similar insights by evaluating on existing gold standards.

We conclude by discussing how observations such as these can guide future research into distributional semantic models. By facilitating better-defined evaluations and finer-grained analyses, we hope that SimLex-999 will ultimately contribute to the development of models that accurately reflect human intuitions of similarity for the full range of concepts in language.

2. Design Motivation

In this section, we motivate the design decisions made in developing SimLex-999. We begin (2.1) by examining the distinction between similarity and association. We then show that for a meaningful treatment of similarity it is also important to take a principled approach to both part-of-speech (POS) and conceptual concreteness (2.2). We

finish by reviewing existing gold standards, and show that none enables a satisfactory evaluation of the capability of models to capture similarity (2.3).

2.1 Similarity and Association

The difference between association and similarity is exemplified by the concept pairs [*car, bike*] and [*car, petrol*]. *Car* is said to be (semantically) similar to *bike* and associated with (but not similar to) *petrol*. Intuitively, *car* and *bike* can be understood as similar because of their common physical features (e.g. wheels), their common function (transport), or because they fall within a clearly definable category (modes of transport). In contrast, *car* and *petrol* are associated because they frequently occur together in space and language, in this case as a result of a clear functional relationship (**?**; **?**).

Association and similarity are neither mutually exclusive nor independent. *Bike* and *car*, for instance, are related to some degree by both relations. Since it is common in both the physical world and in language for distinct entities to interact, it is relatively easy to conceive of concept pairs, such as *car* and *petrol*, that are strongly associated but not similar. Identifying pairs of concepts for which the converse is true is comparatively more difficult. One exception is common concepts paired with low frequency synonyms, such as *camel* and *dromedary*. Since the essence of association is co-occurrence (linguistic or otherwise (**?**)), such pairs can seem, at least intuitively, to be similar but not strongly associated.

To explore the interaction between the two cognitive phenomena quantitatively, we exploited perhaps the only two existing large-scale means of quantifying similarity and association. To estimate similarity, we considered proximity in the WordNet taxonomy (**?**). Specifically, we applied the measure of **?** (henceforth *WupSim*), which approximates similarity on a [0,1] scale reflecting the minimum distance between any two synsets of two given concepts in WordNet. WupSim has been shown to correlate well with human judgements on the similarity-focused RG dataset (**?**). To estimate association, we extracted ratings directly from the University of South Florida Free Association Database (USF) (**?**). These data were generated by presenting human subjects with one of 5000 cue concepts and asking them to write the *first word that comes into their head that is associated with or meaningfully related to that concept*. Each cue concept c was normed in this way by over 10 participants, resulting in a set of associates for each cue, and a total of over 72,000 (c, a) pairs. Moreover, for each such pair, the proportion of participants who produced associate a when presented with cue c can be used as a proxy for the strength of association between the two concepts.

By measuring WupSim between all pairs in the USF dataset, we observed, as expected, a high correlation between similarity and association strength across all USF pairs (Spearman $\rho = 0.65, p < 0.001$). However, in line with the intuitive ubiquity of pairs such as *car* and *petrol*, of the USF pairs (all of which are associated to a greater or lesser degree) over 10% had a WupSim score of less than 0.25. These include pairs of ontologically different entities with a clear functional relationship in the world [*refrigerator, food*], which may be of differing concreteness [*lung, disease*], pairs in which one concept is a small concrete part of a larger abstract category [*sheriff, police*], pairs in a relationship of modification or subcategorization [*gravy, boat*] and even those whose principal connection is phonetic [*wiggle, giggle*]. As we show in Section 2.2, these are precisely the sort of pairs that are not contained in existing evaluation gold standards. Table 1 lists the USF noun pairs with the lowest similarity scores overall, and also those with the largest additive discrepancy between association strength and similarity.

Concept 1	Concept 2	USF	WupSim
hatchet	*murder*	0.013	0.091
robbery	*jail*	0.020	0.100
lung	*disease*	0.014	0.105
burglar	*robbery*	0.020	0.105
sheriff	*police*	0.333	0.133
colonel	*army*	0.303	0.111
quart	*milk*	0.462	0.235
refrigerator	*food*	0.424	0.235

Table 1
Top: Concept pairs with the lowest WupSim scores in the USF dataset overall. Bottom: Pairs with the largest discrepancy in rank between association strength (high) and WupSim (low).

2.1.1 Association and similarity in NLP. As noted in the Introduction, the similarity/association distinction is not only of interest to researchers in psychology or linguistics. Models of similarity are particularly applicable to various NLP tasks, such as lexical resource building, semantic parsing and machine translation (**?**; **?**; **?**; **?**). Models of association, on the other hand, may be better suited to tasks such as word-sense disambiguation (**?**), and applications such as text classification (**?**) in which the target classes correspond to topical domains such as *agriculture* or *sport* (**?**).

Much recent research in *distributional semantics* does not distinguish between association and similarity in a principled way (see e.g. (**?**; **?**; **?**)).[3] One exception is **?**, who constructs two distributional models with different features and parameter settings, explicitly designed to capture either similarity or association. Using the output of these two models as input to a logistic regression classifier, Turney predicts whether two concepts are associated, similar or both, with 61% accuracy. However, in the absence of a gold standard covering the full range of similarity ratings (rather than a list of pairs identified as being similar or not) Turney cannot confirm directly that the similarity-focused model does indeed effectively quantify similarity.

? explicitly examine the distinction between association and similarity in relation to distributional semantic models. Their study is based on the partition of WS-353 into a subset focused on similarity, which we refer to as *WS-Sim*, and a subset focused on association, which we term *WS-Rel*. More precisely, WS-Sim is the union of the pairs in WS-353 judged by three annotators to be similar and the set U of entirely unrelated pairs, and WS-Rel is the union of U and pairs judged to be associated but not similar. Agirre et al. confirm the importance of the association/similarity distinction by showing that certain models perform relatively well on WS-Rel while others perform comparatively better on WS-Sim. However, as shown in the following section, a model need not be an exemplary model of similarity in order to perform well on WS-Sim since an important class of concept pair (associated but not similar entities) is not represented in this dataset. Therefore the insights that can be drawn from the results of the **?** study are limited.

---

3 Several papers that take a knowledge-based or symbolic approach to meaning do address the similarity/association issue (**?**).

Several other authors touch on the similarity/association distinction in inspecting the output of distributional models (**?**; **?**; **?**). While the strength of the conclusions that can be drawn from such qualitative analyses is clearly limited, there appear to be two broad areas of consensus concerning similarity and distributional models:

- Models that learn from input annotated for syntactic or dependency relations better reflect similarity, whereas approaches that learn from running-text or bag-of-words input better model association (**?**; **?**).

- Models with larger context windows may learn representations that better capture association, whereas models with narrower windows better reflect similarity (**?**; **?**).

2.2 Concepts, part-of-speech and concreteness

Empirical studies have shown that the performance of both humans and distributional models depends on the POS category of the concepts learned. **?** showed that children find verb concepts harder to learn than noun concepts, and **?** present evidence that different cognitive operations are employed when comparing two nouns or two verbs. **?** demonstrate differences in the ability of distributional models to acquire noun and verb semantics. Further, they show that these differences are greater for models that learn from both text and perceptual input (as with humans).

In addition to POS category, differences in human and computational concept learning and representation have been attributed to the effects of *concreteness*, the extent to which a concept has a directly perceptible physical referent. On the cognitive side, these 'concreteness effects' are well established, even if the causes are still debated (**?**; **?**). Concreteness has also been associated with differential performance in computational text-based (**?**) and multi-modal semantic models (**?**).

2.3 Existing gold standards and evaluation resources

For brevity, we do not exhaustively review all methods that have been employed to evaluate semantic models, but instead focus on the similarity or association-based gold standards that are most commonly-applied in recent work in NLP. In each case, we consider how well the dataset satisfies one of the three following criteria:

Representative. The resource should cover the full range of concepts that occur in natural language. In particular, it should include cases representing the different ways in which humans represent or process concepts, and cases that are both challenging and straightforward for computational models.

Clearly-defined. In order for a gold standard to be diagnostic of how well a model can be applied to downstream applications, a clear understanding is needed of what exactly the gold standard measures. In particular, it must clearly distinguish between dissociable semantic relations such as association and similarity.

Consistent and reliable. Untrained native speakers must be able to quantify the target property consistently, without requiring lengthy or detailed instructions. This ensures that the data reflect a meaningful cognitive or semantic phenomenon, and also enables the dataset to be scaled up or transferred to other languages at minimal cost and effort.

. We begin our review of existing evaluation with the gold standard most commonly-applied in current NLP research.

WordSim-353. WS-353 (**?**) is perhaps the most commonly-used evaluation gold standard for semantic models. Despite its name, and the fact that it is often referred to as a 'similarity gold standard',[4] in fact, the instructions given to annotators when producing WS-353 were ambiguous with respect to similarity and association. Subjects were asked to:

> *Assign a numerical similarity score between 0 and 10 (0 = words totally unrelated, 10 = words VERY closely related) ... when estimating similarity of antonyms, consider them "similar" (i.e., belonging to the same domain or representing features of the same concept), not "dissimilar".*

As we confirm analytically in Section 5.2, these instructions result in pairs being rated according to association rather than similarity.[5] WS-353 consequently suffers two important limitations as an evaluation of similarity (which also afflict other resources to a greater or lesser degree):

1.   Many dissimilar word pairs receive a high rating.

2.   No associated but dissimilar concepts receive low ratings.

As noted in the Introduction, an arguably more serious third limitation of WS-353 is low inter-annotator agreement, and the fact that state-of-the-art models such as those of **?** and **?** reach, or even surpass, the inter-annotator agreement ceiling in estimating the WS-353 scores. **?** report a Spearman correlation of $\rho = 0.713$ between their model output and WS-353. This is ten percentage points higher than inter-annotator agreement ($\rho = 0.611$) when defined as the average pairwise correlation between two annotators, as is common in NLP work (**?; ?; ?**). It could be argued that a different comparison is more appropriate: Since the model is compared to the gold-standard average across all annotators, we should compare a single annotator with the (almost) gold-standard average over all other annotators. Based on this metric the average performance of an annotator on WS-353 is $\rho = 0.756$, which is still only marginally better than the best automatic method.[6]

Thus, at least according to the established wisdom in NLP evaluation (**?; ?; ?**), the strength of the conclusions that can be inferred from improvements on WS-353 is limited. At the same time, however, state-of-the-art distributional models are clearly not perfect representation-learning or even similarity estimation engines, as evidenced by the fact they cannot yet be applied, for instance, to generate flawless lexical resources (**?**).

WS-Sim. WS-Sim is the set of pairs in WS-353 identified by **?** as either containing similar or unrelated (neither similar nor associated) concepts. The ratings in WS-Sim are mapped directly from WS-353, so that all concept pairs in WS-Sim that receive a high rating are associated and all pairs that receive a low rating are unassociated.

---

4 See e.g. (**?; ?**)
5 This fact is also noted by the dataset authors. See
   `www.cs.technion.ac.il/~gabr/resources/data/wordsim353/`.
6 Individual annotator responses for WS-353 were downloaded from
   `www.cs.technion.ac.il/~gabr/resources/data/wordsim353`.

Consequently, any model that simply reflects association would score highly on WS-Sim, irrespective of how well it captures similarity.

Such a possibility could be excluded by requiring models to perform well on WS-Sim and poorly on WS-Rel, the subset of WS-353 identified by **?** as containing no pairs of similar concepts. However, while this would exclude models of pure association, it would not test the ability of models to quantify the similarity of the pairs in WS-Sim. Put another way, the WS-Sim/WS-Rel partition could in theory resolve limitation (1) of WS-353 but it would not resolve limitation (2): models are not tested on their ability to attribute low scores to associated but dissimilar pairs.

In fact, there are more fundamental limitations of WS-Sim as a similarity-based evaluation resource. It does not, strictly-speaking, reflect similarity at all, since the ratings of its constituent pairs were assigned by the WS-353 annotators, who were asked to estimate association, not similarity. Moreover, it inherits the limitation of low inter-annotator agreement from WS-353. The average pairwise correlation between annotators on WS-Sim is $\rho = 0.667$, and the average correlation of a single annotator with the gold standard is only $\rho = 0.651$, both below the performance of automatic methods (**?**). Finally, the small size of WS-Sim renders it poorly representative of the full range of concepts that semantic models may be required to learn.

Rubenstein & Goodenough. Prior to WS-353, the smaller RG dataset, consisting of 65 pairs, was often used to evaluate semantic models. The 15 raters employed in the data collection were asked to rate the 'similarity of meaning' of each concept pair. Thus RG does appear to reflect similarity rather than association. However, while limitation (1) of WS-353 is therefore avoided, RG still suffers from limitation (2): By inspection, it is clear that the low similarity pairs in RG are not associated. A further limitation is that distributional models now achieve better performance on RG (correlations of up to Person $r = 0.86$ (**?**)) than the reported inter-annotator agreement of $r = 0.85$ (**?**). Finally, the size of RG renders it an even less comprehensive evaluation than WS-Sim.

The MEN Test Collection. A larger dataset, MEN (**?**), is used in a handful of recent studies (**?**; **?**). As with WS-353, both of the terms *similarity* and *relatedness* are used by the authors when describing MEN, although the annotators were expressly asked to rate pairs according to relatedness.[7]

The construction of MEN differed from RG and WS-353 in that each pair was only considered by one rater, who ranked it for relatedness relative to 50 other pairs in the dataset. An overall score out of 50 was then attributed to each pair corresponding to how many times it was ranked as more related than an alternative. However, because these rankings are based on relatedness, with respect to evaluating similarity MEN necessarily suffers from both of the limitations (1) and (2) that apply to WS-353. Further, there is a strong bias towards concrete concepts in MEN because the concepts were originally selected from those identified in an image-bank (**?**).

Synonym detection sets. Multiple-choice synonym detection tasks, such as the TOEFL test questions (**?**), are an alternative means of evaluating distributional models. A question in the TOEFL task consists of a cue word and four possible answer words, only one of which is a true synonym. Models are scored on the number of true synonyms identified out of 80 questions. The questions were designed by linguists to evaluate

---

7 http://clic.cimec.unitn.it/ elia.bruni/MEN.html

synonymy, so, unlike the evaluations considered thus far, TOEFL-style tests effectively discriminate between similarity and association. However, since they require a zero-one classification of pairs as synonymous or not, they do not test how well models discern pairs of medium or low similarity. More generally, in opposition to the fuzzy, statistical approaches to meaning predominant in both cognitive psychology (**?**) and NLP (**?**), they do not require similarity to be measured on a continuous scale.

3. The SimLex-999 Dataset

Having considered the limitations of existing gold standards, in this section we describe the design of SimLex-999 in detail.

3.1 Choice of Concepts

Separating similarity from association. To create a test of the ability of models to capture similarity as opposed to association, we started with the $\approx 72,000$ pairs of concepts in the USF dataset. As the output of a free-association experiment, each of these pairs is associated to a greater or lesser extent. Importantly, inspecting the pairs revealed that a good range of similarity values are represented. In particular, there were many examples of hypernym / hyponym pairs [*body, abdomen*] cohyponym pairs [*cat, dog*], synonyms or near synonyms [*deodorant, antiperspirant*] and antonym pairs [*good, evil*]. From this cohort, we excluded pairs containing a multiple-word item [*hot dog, mustard*], and pairs containing a capital letter [*Mexico, sun*]. We ultimately sampled 900 of the SimLex-999 pairs from the resulting cohort of pairs according to the stratification procedures outlined in the following sections.

To complement this cohort with entirely unassociated pairs, we paired up the concepts from the 900 associated pairs at random. From these random parings, we excluded those that coincidentally occurred elsewhere in USF (and therefore had a degree of association). From the remaining pairs, we accepted only those in which both concepts had been subject to the USF norming procedure, ensuring that these non-USF pairs were indeed unassociated rather than simply not normed. We sampled the remaining 99 SimLex-999 pairs from this resulting cohort of unassociated pairs.

POS category. In light of the conceptual differences outlined in Section 2.2, SimLex-999 includes subsets of pairs from the three principle meaning-bearing POS categories, nouns, verbs and adjectives. To classify potential pairs according to POS, we counted the frequency with which the items in each pair occurred with the three possible tags in the POS-tagged British National Corpus (**?**). To minimise POS ambiguity, which could lead to inconsistent rating, we excluded pairs containing a concept with lower than 75% tendency towards one particular POS. This yielded three sets of potential pairs : [A,A] pairs (of two concepts whose majority tag was Adjective), [N,N] pairs and [V,V] pairs.

Given the likelihood that different cognitive operations are employed in estimating the similarity between items of different POS-category (Section 2.2), concept pairs were presented to raters in batches defined according to POS. Unlike both WS-353 and MEN, pairs of concepts of mixed POS ([*white, rabbit*], [*run,marathon*]) were excluded. POS categories are generally considered to reflect very broad ontological classes (**?**). We thus felt it would be very difficult, or even counter-intuitive, for annotators to quantify the similarity of mixed POS pairs according to our instructions.

Concreteness. Although a clear majority of pairs in gold standards such as MEN and RG contain concrete items, perhaps surprisingly, the vast majority of adjective, noun and verb concepts in everyday language are in fact abstract (**?; ?**).[8] To facilitate the evaluation of models for both concrete and abstract concept meaning, and in light of the cognitive and computational modelling differences between abstract and concrete concepts noted in Section 2.2, we aimed to include both concept types in SimLex-999.
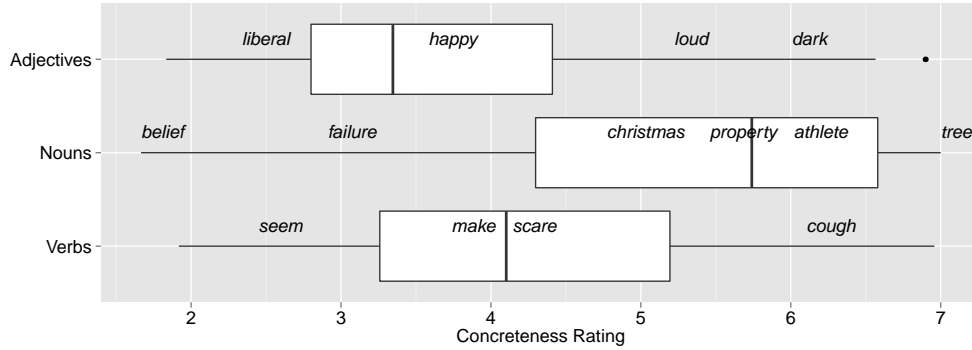
Figure 1
Boxplots showing the interaction between concreteness and POS for concepts in USF. The white boxes range from the first to third quartiles and the central vertical line indicates the median.

Unlike the POS distinction, concreteness is generally considered to be a gradual phenomenon. One benefit of sampling pairs for SimLex-999 from the USF dataset is that most items have been rated according to concreteness on a scale of 1-7 by at least 10 human subjects. As Figure 1 demonstrates, concreteness (as the average over these ratings) interacts with POS on these concepts: nouns are on average more concrete than verbs which are more concrete than adjectives. However, there is also clear variation in concreteness within each POS category. We therefore aimed to select pairs for SimLex-999 that spanned the full abstract-concrete continuum within each POS category.
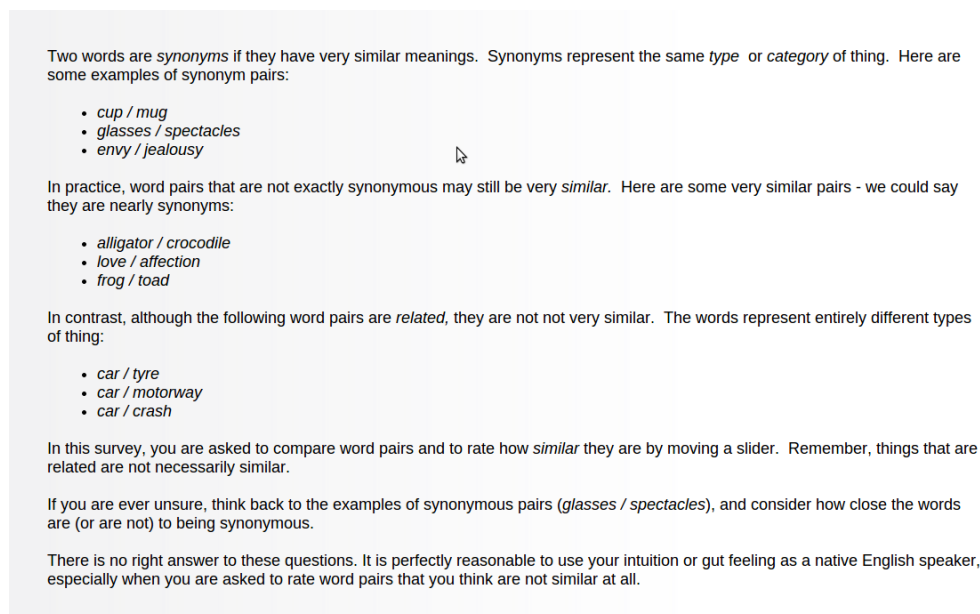
After excluding any pairs that contained an item with no concreteness rating, for each potential SimLex-999 pair we considered both the concreteness of the first item and the additive difference in concreteness between the two items. This enabled us to stratify our sampling equally across four classes: (C_1) concrete first item (rating > 4) with below-median concreteness difference, (C_2) concrete first item (rating > 4), second item of lower concreteness and the difference being greater than the median, (C_3) abstract first item (rating ≤ 4) with below-median concreteness difference, and (C_4) abstract first item (rating ≤ 4) with the second item of greater concreteness and the difference being greater than the median.

Final sampling. From the associated (USF) cohort of potential pairs we selected 600 noun pairs, 200 verb pairs and 100 adjective pairs, and from the unassociated (non-USF) cohort, we sampled 66 nouns pairs, 22 verb pairs and 11 adjective pairs. In both cases, the sampling was stratified such that, in each POS subset, each of the four concreteness classes $C_1 - C_4$ was equally represented.

---

8 According to the USF concreteness ratings, 72% of noun or verb types in the British National Corpus are more abstract than the concept *war*, a concept many would already consider quite abstract.

3.2 Question Design

The annotator instructions for SimLex-999 are shown in Figure 2. We did not attempt to formalise the notion of similarity, but rather introduce it via the well-understood idea of synonymy, and in contrast to association. Even if a formal characterisation of similarity existed, the evidence in Section 2 suggests that the instructions would need separate cases to cover different concept types, increasing the difficulty of the rating task. Therefore we preferred to appeal to intuition on similarity, and to verify post-hoc that subjects were able to interpret and apply the informal characterization consistently for each concept type.

Two words are *synonyms* if they have very similar meanings. Synonyms represent the same *type* or *category* of thing. Here are some examples of synonym pairs:

- *cup / mug*
- *glasses / spectacles*
- *envy / jealousy*

In practice, word pairs that are not exactly synonymous may still be very *similar*. Here are some very similar pairs - we could say they are nearly synonyms:

- *alligator / crocodile*
- *love / affection*
- *frog / toad*

In contrast, although the following word pairs are *related,* they are not not very similar. The words represent entirely different types of thing:

- *car / tyre*
- *car / motorway*
- *car / crash*

In this survey, you are asked to compare word pairs and to rate how *similar* they are by moving a slider. Remember, things that are related are not necessarily similar.

If you are ever unsure, think back to the examples of synonymous pairs (*glasses / spectacles*), and consider how close the words are (or are not) to being synonymous.

There is no right answer to these questions. It is perfectly reasonable to use your intuition or gut feeling as a native English speaker, especially when you are asked to rate word pairs that you think are not similar at all.

Figure 2
Instructions for SimLex-999 annotators.

Immediately following the instructions in Figure 2, participants were presented with two 'checkpoint' questions, one with abstract examples and one with concrete examples. In each case the participant was required to identify the *most similar* pair from a set of three options, all of which were associated, but only one of which was clearly similar (e.g. [*bread, butter*] [*bread, toast*] [*stale, bread*]). After this, the participants began rating pairs in groups of 6 or 7 pairs by moving a slider, as shown in Figure 3.

This group size was chosen because the (relative) rating a set of pairs implicitly requires pairwise comparisons between all pairs in that set. Therefore, larger groups would have significantly increased the cognitive load on the annotators. Another advantage of grouping was the clear break (submitting a set of ratings and moving to the next page) between the tasks of rating adjective, noun and verb pairs. For better inter-group calibration, from the second group onwards the last pair of the previous group became the first pair of the present group, and participants were asked to re-assign the rating previously attributed to the first pair before rating the remaining new items.

3.3 Context-free rating

As with MEN, WS-353 and RG, SimLex-999 consists of pairs of concept words together with a numerical rating. Thus, unlike in the small evaluation constructed by **?**, words are not rated in a phrasal or sentential context. Such meaning-in-context evaluations are motivated by a desire to disambiguate words that otherwise might be considered to have multiple senses.

We did not attempt to construct an evaluation based on meaning-in-context for several reasons. First, determining the set of senses for a given word, and then the set of contexts that represent those senses, introduces a high degree of subjectivity into the design process. Second, ensuring that a model has learned a high quality representation of a given concept would have required evaluating that concept in each of its given contexts, necessitating many more cases and a far greater annotation effort. Third, in the (infrequent) case that some concept c_1 in an evaluation pair (c_1, c_2) is genuinely (etymologically) polysemous, c_2 can provide sufficient context to disambiguate c_1.[9] Finally, the POS grouping of pairs in the survey can also serve to disambiguate in the case that the conflicting senses of the polysemous concept are of differing POS category.

Figure 3
A group of noun pairs to be rated by moving the sliders. The rating slider was initially at position 0, and it was possible to attribute a rating of 0, although it was necessary to have actively moved the slider to that position to proceed to the next page.

3.4 Questionnaire structure

Each participant was asked to rate 20 groups of pairs on a 0-6 scale of integers (non-integral ratings were not possible). Checkpoint multiple-choice questions were inserted at points between the 20 groups in order to ensure the participant had retained the correct notion of similarity. In addition to the checkpoint of three noun pairs presented before the first group (which contained noun pairs), checkpoint questions containing adjective pairs were inserted before the first adjective group and checkpoints of three verb pairs were inserted before the first verb group.

From the 999 evaluation pairs, 14 noun pairs, 4 verb pairs and 2 adjective pairs were selected as a *consistency set*. The dataset of pairs was then partitioned into 10 tranches, each consisting of 119 pairs, of which 20 were from the consistency set and the remaining 99 unique to that tranche. To reduce workload, each annotator was asked to rate the pairs in a single tranche only. The tranche itself was divided into 20 groups, with

---

9 This is supported by the fact that the WordNet-based methods that perform best at modeling human ratings model the similarity between concepts c_1 and c_2 as the minimum of all pairwise distances between the senses of c_1 and the senses of c_2 (**?**; **?**).

each group consisting of 7 pairs (with the exception of the last group of the 20, which had 6). Of these 7 pairs, the first pair was the last pair from the previous group, and the second pair was taken from the consistency set. The remaining pairs were unique to that particular group and tranche. The design enabled control for possible systematic differences between annotators and tranches, which could be detected by variation on the consistency set.

3.5 Participants

500 residents of the USA were recruited from Mechanical Turk, each with at least 95% approval rate for work on the web service. Each participant was required to check a box confirming that he or she was a native speaker of English and warned that work would be rejected if the pattern of responses indicated otherwise. The participants were distributed evenly to rate pairs in one of the ten question tranches, so that each pair was rated by approximately 50 subjects. Participants took between 8 and 21 minutes to rate the 119 pairs across the 20 groups, together with the checkpoint questions.

3.6 Post-processing

In order to correct for systematic differences in the overall calibration of the rating scale between respondents, we measured the average (mean) response of each rater on the consistency set. For 32 respondents, the absolute difference between this average and the mean of all such averages was greater than one (though never greater than two); i.e. 32 respondents demonstrated a clear tendency to rate pairs as either more or less similar than the overall rater population. To correct for this bias, we increased (or decreased) the rating of such respondents for each pair by one, except in cases where they had given the maximum rating, six (or minimum rating, zero). This adjustment, which ensured that the average response of each participant was within one of the mean of all respondents on the consistency set, resulted in a small increase to the inter-rater agreement on the dataset as a whole.

After controlling for systematic calibration differences, we imposed three conditions for the responses of a rater to be included in the final data collation. First, the average pairwise Spearman correlation of responses with all other responses for a participant could not be more than one standard deviation below the mean of all such averages. Second, the increase in inter-rater agreement when a rater was excluded from the analysis needed to be smaller than at least 50 other raters (i.e. 10% of raters were excluded on this criterion). Third, we excluded the 6 participants who got one or more of the checkpoint questions wrong. A total of 99 participants were excluded based on one or more of these conditions, but no more than 16 from any one tranche (so that each pair in the final dataset was rated by a minimum of 36 raters). Finally, we computed average (mean) scores for each pair, and transformed all scores linearly from the interval $[0, 6]$ to the interval $[0, 10]$.

4. Analysis of Dataset

In this section we analyse the responses of the SimLex-999 annotators and the resulting ratings. First, by considering inter-annotator agreement we examine the consistency with which annotators were able to apply the characterization of similarity outlined in the instructions to the range of concept types in SimLex-999. Second, we verify that

a valid notion of similarity was understood by the annotators, in that they were able to accurately separate similarity from association.

4.1 Inter-annotator agreement
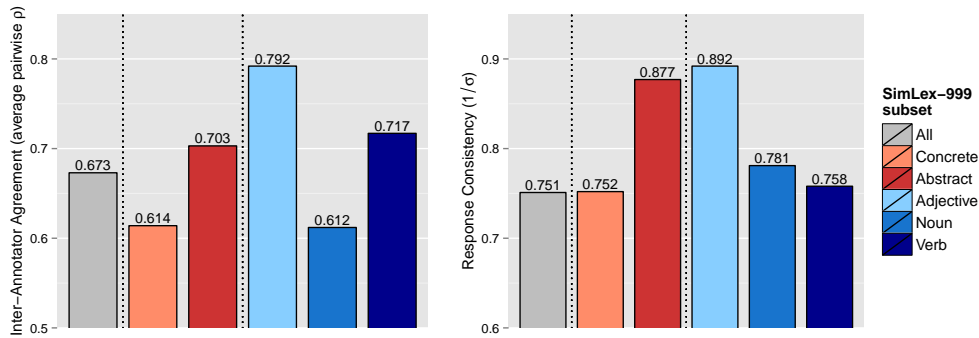
Figure 4
Left: Inter-annotator agreement, measured by average pairwise Spearman ρ correlation, for ratings of concept types in SimLex-999. **Right:** Response consistency, reflecting the standard deviation of annotator ratings for each pair, averaged over all pairs in the concept category.

As in previous annotation or data collection for computational semantics (**?**; **?**; **?**) we computed the inter-rater agreement as the average of pairwise Spearman ρ correlations between the ratings of all respondents. Overall agreement was $\rho = 0.67$. This compares favourably with the agreement on WS-353 ($\rho = 0.61$ using the same method). The design of the MEN rating system precludes a conventional calculation of inter-rater agreement (**?**). However, two of the creators of MEN who independently rated the dataset achieved an agreement of $\rho = 0.68$.[10]

The SimLex-999 inter-rater agreement suggests that participants were able to understand the (single) characterization of similarity presented in the instructions and to apply it to concepts of various types consistently. This conclusion was supported by inspection of the brief feedback offered by the majority of annotators in a final text field in the questionnaire: 78% expressed sentiment that the test was clear, easy to complete or some similar sentiment.

Interestingly, as shown in Figure 4 (left), agreement was not uniform across the concept types. Contrary to what might be expected given established concreteness effects (**?**), we observed not only higher inter-rater agreement but also less per-pair variability for abstract rather than concrete concepts[11].

Strikingly, the highest inter-rater consistency and lowest per-pair variation (defined as the inverse of the standard deviation of all ratings for that pair) was observed on adjective pairs. While we are unsure exactly what drives this effect, a possible cause is that many pairs of adjectives in SimLex-999 cohabit a single salient, one-dimensional scale (*freezing > cold > warm > hot*). This may be a consequence of the fact that many pairs

---

10 Reported at http://clic.cimec.unitn.it/ elia.bruni/MEN. It is reasonable to assume that actual agreement on MEN may be somewhat lower than 0.68 given the small sample size and the expertise of the raters.

11 Per-pair variability was measured by calculating the standard deviation of responses for each pair, and averaging these scores across the pairs of each concept type.

C1	C2	POS	USF*	USF rank (of 999)	SimLex	SimLex rank (of 999)
dirty	*narrow*	A	0.00	999	0.30	996
student	*pupil*	N	6.80	12	9.40	12
win	*dominate*	V	0.41	364	5.68	361
smart	*dumb*	A	2.10	92	0.60	947
attention	*awareness*	N	0.10	895	8.73	58
leave	*enter*	V	2.16	89	1.38	841

Table 2
Top: Similarity aligns with association Pairs with a small difference in rank between USF (association) and SimLex-999 (similarity) scores for each POS category. **Bottom: Similarity contrasts with association** Pairs with a high difference in rank for each POS category. *Note that the distribution of USF association scores on the interval [0,10] is highly skewed towards the lower bound in both SimLex-999 and the USF dataset as a whole.

in SimLex-999 were selected (from USF) to have a degree of association. On inspection, pairs of nouns and verbs in SimLex-999 do not appear to occupy scales in the same way, possibly since concepts of these POS categories come to be associated via a more diverse range of relations. It seems plausible that humans are able to estimate the similarity of scale-based concepts more consistently than pairs of concepts related in a less uni-dimensional fashion.

Regardless of cause, however, the high agreement on adjectives is a satisfactory property of SimLex-999. Adjectives exhibit various aspects of lexical semantics that have proved challenging for computational models, including antonymy, polarity (**?**) and sentiment (**?**). To approach the high level of human confidence on the adjective pairs in SimLex-999, it may be necessary to focus particularly on developing automatic ways to capture these phenomena.

4.2 Response validity: Similarity not association

Inspection of the SimLex-999 ratings indicated that pairs were indeed evaluated accord-ing to similarity rather than association. Table 2 includes examples that demonstrate a clear dissociation between the two semantic relations.

To verify this effect quantitatively, we recruited 100 additional participants to rate the WS-353 pairs, but following the SimLex-999 instructions and question format. As shown in Fig 5(a), there were clear differences between these new ratings and the original WS-353 ratings. In particular, a high proportion of pairs was given a lower rating by subjects following the SimLex-999 instructions than those following the WS-353 guidelines: The mean SimLex rating was 4.07 compared with 5.91 for WS-353.

This was consistent with our expectations that pairs of associated but dissimilar concepts would receive lower ratings based on the SimLex-999 than on the WS-353 instructions while pairs that were both associated and similar would receive similar ratings in both cases. To confirm this, we compared the WS-353 and SimLex-999-based ratings on the subsets WS-Rel and WS-Sim, which were hand-sorted by **?** to include pairs connected by association (and not similarity) and those connected by similarity (but possibly also association) respectively.

As shown in Figure 5(b-c), the correlation between the SimLex-999-based and WS-353 ratings was notably higher ($\rho = 0.73$) on the WS-Sim subset than the WS-Rel subset

Figure 5
(a) Pairs rated by WS-353 annotators (blue points, ranked by rating) and the corresponding rating of annotators following the SimLex-999 instructions (red points). **(b-c)** The same analysis, restricted to pairs in the WS-Sim or WS-Rel subsets of WS-353.

($\rho = 0.38$). Specifically, the tendency of subjects following the SimLex-999 instructions to assign lower ratings than those following the WS-353 instructions was far more pronounced for pairs in WS-Sim (Figure 5(b)) than for those in WS-Rel (5(c)). This observation suggest that the associated but dissimilar pairs in WS-353 were an important driver of the overall lower mean for SimLex-999-based ratings, and thus provide strong evidence that the SimLex-999 instructions do indeed enable subjects to distinguish similarity from association effectively.

4.3 Finer-grained Semantic Relations

We have established the validity of similarity as a notion understood by human raters and distinct from association. However, much theoretical semantics focuses on relations between words or concepts that are finer-grained than similarity and association. These include *meronymy* (a part to its whole, e.g. *blade - knife*), *hypernymy* (a category concept to a member of that category, e.g. *animal - dog*) and *cohyponymy* (two members of the same implicit category, e.g. the pair of animals *dog - cat*) (**?**). Beyond theoretical interest, these relations can have practical relevance. For instance, hypernymy can form the basis of semantic entailment and therefore textual inference: the proposition *a cat is on the table* entails that *an animal is on the table* precisely because of the hypernymy relation from animal to cat.

We chose not to make these finer-grained relations the basis of our evaluation for several reasons. At present, detecting relations such as hypernymy using distributional methods is challenging, even when supported by supervised classifiers with access to labelled pairs (**?**). Such a designation can seem to require specific world-knowledge (is a *snale* a *reptile*?), can be gradual, as evidenced by typicality effects (**?**), or simply highly subjective. Moreover, a fine-grained relation R will only be attested (to any degree) between a small subset of all possible word pairs, whereas similarity can in theory be quantified for any two words chosen at random. We thus considered a focus on fine-

grained semantic relations to be less appropriate for a general-purpose evaluation of representation quality.

Nevertheless, post-hoc analysis of the SimLex annotator responses and fine-grained relation classes, as defined by lexicographers, yields further interesting insights into the nature of both similarity and association. Of the 999 word pairs in SimLex, 382 are also connected by one of the common finer-grained semantic relations in WordNet. For each of these relations, Figure 6 shows the average similarity rating and average USF free association score for all pairs that exhibit that relation.

In cases where a relationship of hypernymy/hyponymy exists between the words in a pair (not necessarily immediate - $1\_hypernym, 2\_hypernym$ etc.) similarity and association coincide. Hyper/hyponym pairs that are separated by fewer levels in the WordNet hierarchy are both more strongly associated and rated as more similar. However, there are also interesting discrepancies between similarity and association. Unsurprisingly, pairs that are classed as synonyms in WordNet (i.e. having at least one sense in some common synset) are rated as more similar than pairs of any other relation type by SimLex annotators. In contrast, antonyms are the most strongly-associated word pairs among these finer-grained relations. Further, pairs consisting of a meronym and holonym (part and whole) are comparatively strongly associated but not judged to be similar.

The analysis also highlights a case that can be particularly problematic when rating similarity; cohyponyms, or members of the same salient category (such as *knife* and *fork*). We gave no specific guidelines for how to rate such pairs in the SimLex annotator instructions, and whether they are considered similar or not seems to be a matter of perspective. On one hand, their membership of a common category could make them appear similar, particularly if the category is relatively specific. On the other hand, in the case of *knife* and *fork*, for instance, the underlying category *cutlery* might provide a backdrop against which the differences of distinct members become particularly salient.
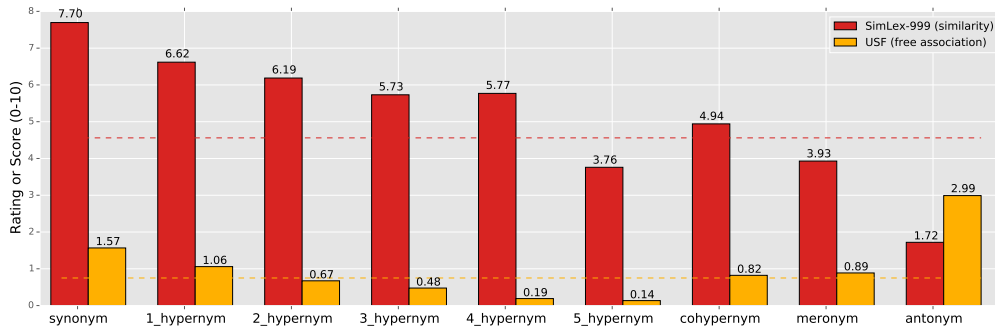
Figure 6
Average SimLex and USF free association scores across pairs representing different fine-grained semantic relations. All relations were extracted from WordNet. $n\_hypernym$ refers to a direct hypernymy path of length n. Note that the average SimLex rating across all 999 word pairs (dashed red line) is much higher than the average USF rating (dashed golden line) because of differences in the rating procedure. The more interesting differences concern the relative strength of similarity vs. association across the different relation types.

5. Evaluating Models with SimLex-999

In this section, we demonstrate the applicability of SimLex-999 by analysing the performance of various distributional semantic models in estimating the new ratings. The models were selected to cover the main classes of representation learning architectures (**?**): Vector space co-occurrence (counting) models and neural language models (NLM)s (**?**). We first show that SimLex-999 is notably more difficult for state-of-the-art models to estimate than existing gold standards. We then conduct more focused analyses on the various concept subsets defined in SimLex-999, exploring possible causes for the comparatively low performance of current models and, in turn, demonstrating how SimLex-999 can be applied to investigate such questions.

5.1 Semantic models

Collobert & Weston. ? apply the architecture of an NLM to learn a word representations v_w for each word w in some corpus vocabulary V. Each sentence s in the input text is represented by a matrix containing the vector representations of the words in s in order. The model then computes output scores $f(s)$ and $f(s^w)$, where s^w denotes an 'incorrect' sentence created from s by replacing its last word with some other word w from V. Training involves updating the parameters of the function f and the entries of the vector representations v_w such that $f(s)$ is larger than $f(s^w)$ for any w in V, other than the correct final word of s. This corresponds to minimising the sum of the following sentence objectives C_s over all sentences in the input corpus, which is achieved via (mini-batch) stochastic gradient descent:

$$C_s = \sum_{w \in V} max(0, 1 - f(s) + f(s^w)).$$

The relatively low-dimension, dense (vector) representations learned by this model and the other NLMs introduced in this section are sometimes referred to as *embeddings* (**?**). **?** train their models on 852 million words of text from a 2007 dump of Wikipedia and the RCV1 Corpus (**?**) and use their embeddings to achieve state-of-the-art results on a variety of NLP tasks. We downloaded the embeddings directly from the authors' webpage.[12]

Huang et al.. ? present a NLM that learns word embeddings to maximise the likelihood of predicting the last word in a sentence s based on (i) the previous words in that sentence (local context - as with **?**) and (ii) the document d in which that word occurs (global context). As with **?**, the model represents input sentences as a matrix of word embeddings. In addition, it represents documents in the input corpus as single-vector averages over all word embeddings in that document. It can then compute scores $g(s, d)$ and $g(s^w, d)$, where as before s^w is a sentence with an 'incorrect' randomly-selected last word. Training is again by stochastic gradient descent, and corresponds to minimising the sum of the sentence objectives $C_{s,d}$ over all of the sentences in the corpus:

$$C_{s,d} = \sum_{w \in V} max(0, 1 - g(s, d) + g(s^w, d)).$$

---

12 http://ml.nec-labs.com/senna/

The combination of local and global contexts in the objective encourages the final word embeddings to reflect aspects of both the meaning of nearby words and of the documents in which those words appear. When learning from 990m words of wikipedia text, Huang et al. report a Spearman correlation of $\rho = 71.3$ between the cosine similarity of their model embeddings and the WS-353 scores, which constitutes state-of-the-art performance for a NLM model on that dataset. We downloaded these embeddings from the authors' webpage.[13]

Mikolov et al.. ? present an architecture that learns word embeddings similar to those of standard NLMs but with no non-linear hidden layer (resulting in a simpler scoring function). This enables faster representation learning for large vocabularies. Despite this simplification, the embeddings achieve state-of-the-art performance on several semantic tasks including sentence completion and analogy modelling (**?**; **?**).

For each word type w in the vocabulary V, the model learns both a 'target-embedding' $r_w \in \mathbb{R}^d$ and a 'context-embedding' $\hat{r}_w \in \mathbb{R}^d$ such that, given a target word, its ability to predict nearby context words is maximised. The probability of seeing context word c given target w is defined as:

$$p(c|w) = \frac{e^{\hat{r}_c \cdot r_w}}{\sum_{v \in V} e^{\hat{r}_v \cdot r_w}}.$$

The model learns from a set of (target-word, context-word) pairs, extracted from a corpus of sentences as follows. In a given sentence s (of length N), for each position $n \leq N$, each word w_n is treated in turn as a target word. An integer $t(n)$ is then sampled from a uniform distribution on $\{1, \ldots k\}$, where $k > 0$ is a predefined maximum context-window parameter. The pair tokens $\{(w_n, w_{n+j}) : -t(n) \leq j \leq t(n), w_i \in s\}$ are then appended to the training data. Thus, target/context training pairs are such that (i) only words within a k-window of the target are selected as context words for that target, and (ii) words closer to the target are more likely to be selected than those further away.

The training objective is then to maximise the log probability T, defined below, across of all such examples from s, and then across all sentences in the corpus. This is achieved by stochastic gradient descent.

$$T = \frac{1}{N} \sum_{n=1}^{N} \sum_{-t(n) \leq j \leq t(n), j \neq 0} log(p(w_{n+j}|w_n))$$

As with other NLMs, Mikolov et al.'s model captures conceptual semantics by exploiting the fact that words appearing in similar linguistic contexts are likely to have similar meanings. Informally, the model adjusts its embeddings to increase the probability of observing the training corpus. Since this probability increases with $p(c|w)$, and $p(c|w)$ increases with the dot product $\hat{r}_c \cdot r_w$, the updates have the effect of moving each target-embedding incrementally 'closer' to the context-embeddings of its collocates. In the target-embedding space, this results in embeddings of concept words that regularly occur in similar contexts moving closer together.

---

13 `www.socher.org/index.php/Main/`
   `ImprovingWordRepresentationsViaGlobalContextAndMultipleWordPrototypes.`

We use the author's Word2vec software in order to train their model and use the target embeddings in our evaluations. We experimented with embeddings of dimension 100, 200, 300, 400 and 500 and found that 200 gave the best performance on both WS-353 and SimLex-999.

Vector Space Model (VSM). As an alternative to NLMs, we constructed a vector space model following the guidelines for optimal performance outlined by **?**. After extracting the 2000 most frequent word tokens in the corpus that are not in a common list of stopwords[14] as features, we populated a matrix of co-occurrence counts with a row for each of the concepts in some pair in our evaluation sets, and a column for each of the features. Co-occurrence was counted within a specified window size, although never across a sentence boundary. This resulting matrix was then weighted according to Pointwise Mutual Information (PMI) (**?**). The rows of the resulting matrix constitute the vector representations of the concepts.

SVD. As proposed initially in **?**, we also experimented with models in which Singular Value Decomposition (SVD) (**?**) is applied to the PMI-weighted VSM matrix, reducing the dimension of each concept representation to 300 (which yielded best results after experimenting, as before, with 100-500 dimension vectors).

For each model described in this section, we calculate similarity as the cosine similarity between the (vector) representations learned by that model.

5.2 Results

In experimenting with different models on SimLex-999, we aimed to answer the following questions: (i) How well do the established models perform on SimLex-999 versus on existing gold standards? (ii) Are any observed differences caused by the potential of different models to measure similarity vs. association? (iii) Are there interesting differences in ability of models to capture similarity between adjectives vs nouns vs verbs? (iv) In this case, are the observed differences driven by concreteness, and its interaction with POS, or are other factors also relevant?

Overall performance on SimLex-999. Figure 7 shows the performance of the NLMs on SimLex-999 versus on comparable datasets, measured by Spearman's ρ correlation. All models estimate the ratings of MEN and WS-353 more accurately than SimLex-999. The **?** model performs well on WS-353,[15] but is not very robust to changes in evaluation gold standard, and performs worst of all the models on SimLex-999. Given the focus of the WS-353 ratings, it is tempting to explain this by concluding that the global context objective leads the **?** model to focus on association rather than similarity. However, the true explanation may be less simple, since the **?** model performs weakly on the association-based MEN dataset. The **?** model is more robust across WS-353 and MEN, but still does not match the performance of the **?** model on SimLex-999.

---

14 Taken from the Python Natural Language Toolkit (**?**).
15 This score, based on embeddings downloaded from the authors' webpage, is notably lower than the score reported in (**?**) mentioned in Section 5.1.
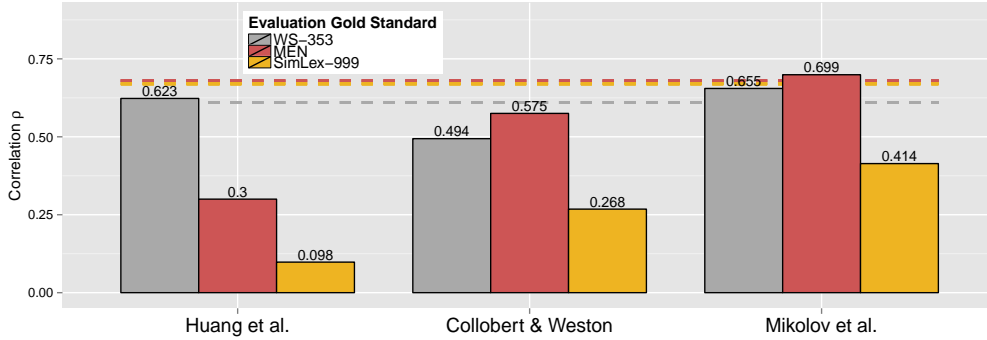
Figure 7
Performance of NLMs on WS-353, MEN and SimLex-999. All models are trained on Wikipedia; note that as Wikipedia is constantly growing, the **?** model exploited slightly more training data (≈1000m tokens) than the **?** model (≈990m), which in turn exploited more than the **?** model (≈852m). Dashed horizontal lines indicate the level of inter-annotator agreement for the three datasets.
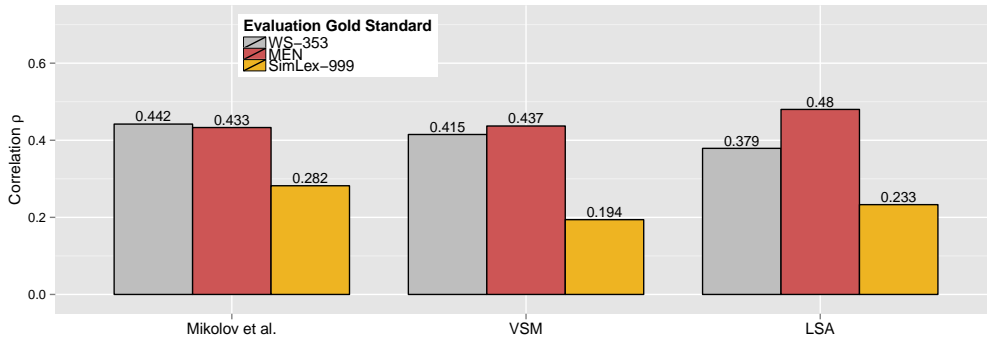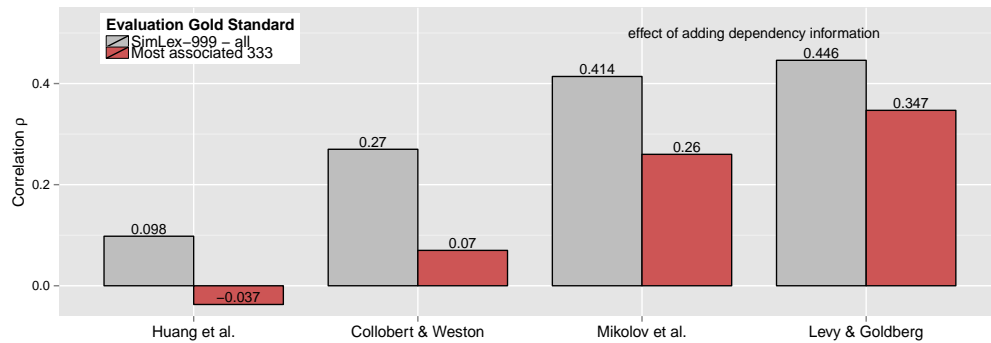

Figure 8
Comparison between the leading NLM, *Mikolov et al.*, the vector space model, *VSM*, and the *SVD* model. All models were trained on the ≈150m word RCV1 Corpus (**?**).

Figure 8 compares the best performing NLM model (**?**) with the VSM and SVD models.[16] In contrast to recent results that emphasise the superiority of NLMs over alternatives (**?**), we observed no clear advantage for the NLM over the VSM or SVD when considering the association-based gold standards WS-353 and MEN together. While the NLM is the strongest performer on WS-353, SVD is the strongest performer on MEN. However, the NLM model performs notably better than the alternatives at modelling similarity, as measured by SimLex-999.

Comparing all models in Figures 7 and 8 suggests that SimLex-999 is notably more challenging to model than the alternative datasets, with correlation scores ranging from 0.098 to 0.414. Thus, even when state-of-the-art models are trained for several days

---

16 We conduct this comparison on the smaller RCV1 Corpus (**?**) because training the VSM and SVD models is comparatively slow.

Figure 9
The ability of NLMs to model the similarity of highly-associated concepts versus concepts in general. The two models on the right hand side also demonstrate the effect of training an NLM (the **?** model) on running-text (*Mikolov et al.*) vs. on dependency-based input (*Levy & Goldberg*).

on massive text corpora, [17] their performance on SimLex-999 is well below the interannotator agreement (Figure 7). This suggests that there is ample scope for SimLex-999 to guide the development of improved models.

Modeling similarity vs. association. The comparatively low performance of NLM, VSM and SVD models on SimLex-999 compared with MEN and WS-353 is consistent with our hypothesis that modelling similarity is more difficult than modelling association. Indeed, given that many strongly-associated but dissimilar pairs, such as [*coffee, cup*], are likely to have high co-occurrence in the training data, and that all models infer connections between concepts from linguistic co-occurrence in some form or another, it seems plausible that models may overestimate the similarity of such pairs because they are 'distracted' by association.

To test this hypothesis more precisely, we compared the performance of models on the whole of SimLex-999 versus its 333 most associated pairs (according to the USF free association scores). Importantly, pairs in this strongly-associated subset still span the full range of possible similarity scores (min similarity = 0.23 [*shrink, grow*], max similarity = 9.80 [*vanish, disappear*]).

As shown in Figure 9, all models performed worse when the evaluation was restricted to pairs of strongly-associated concepts, which was consistent with our hypothesis. The **?** model was better than the **?** model at estimating similarity in the face of high association. This not entirely surprising given the global-context objective in the latter model, which may have encouraged more association-based connections between concepts. The Mikolov et al. model, however, performed notably better than both other NLMs. Moreover, this superiority is proportionally greater when evaluating on the most associated pairs only (as indicated by the difference between the red and grey bars), suggesting that the improvement is driven at least in part by an increased ability to 'distinguish' similarity from association.

To understand better how the architecture of models captures information pertinent to similarity modelling, we performed two additional experiments using SimLex-999.

---

17 Training times reported by **?** and for **?** at `http://ronan.collobert.com/senna/`.
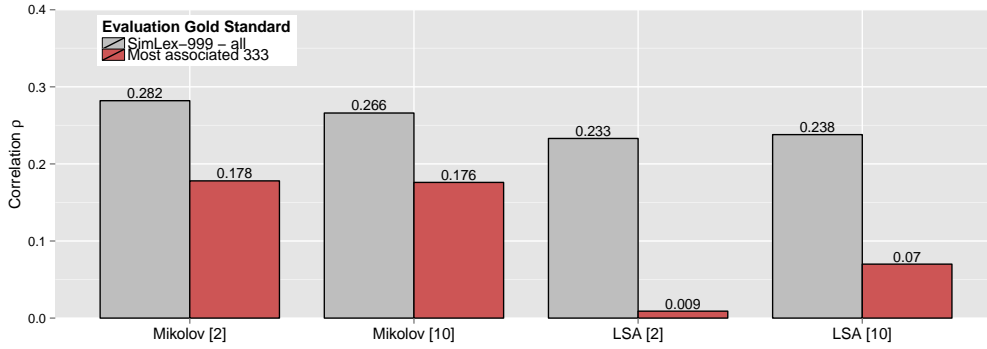
Figure 10
The effect of different window sizes (indicated in square brackets [ ]) on NLM and SVD models.
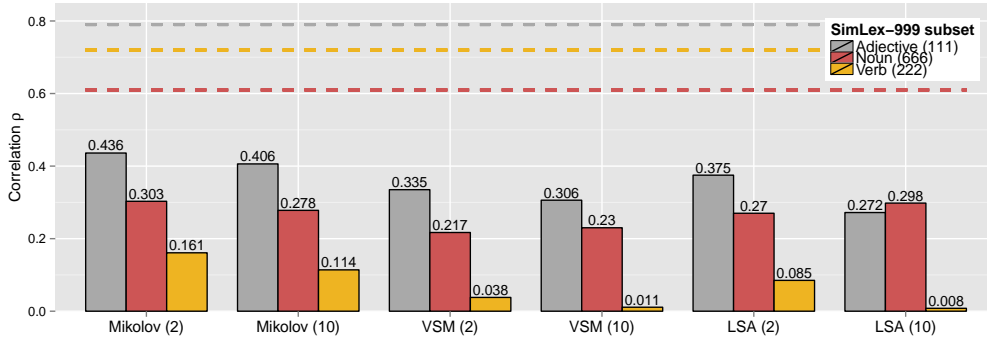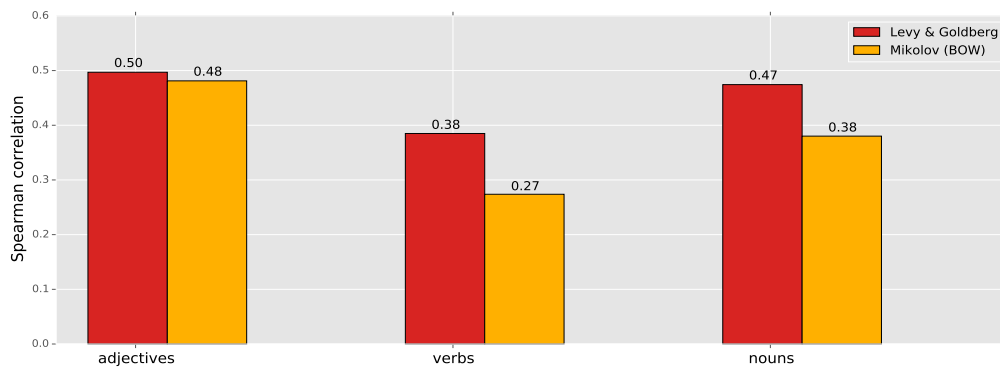

Figure 11
Performance of models on POS-based subsets of SimLex-999. The window size for each model is indicated in parentheses. Inter-annotator agreement for each POS is indicated by the dashed horizontal line.

These comparisons were also motivated by the hypotheses, made in previous studies and outlined in Section 2.1.2, that both dependency-informed input and smaller context windows encourage models to capture similarity rather than association.

We tested the first hypothesis using the embeddings of **?**, whose model extends the **?** model so that target-context training instances are extracted based on dependency-parsed rather than simple running text. As illustrated in Figure 9, the dependency-based embeddings outperform the original (running text) embeddings trained on the same corpus. Moreover, the comparatively large increase in the red bar compared to the grey bar suggests that an important part of the improvement of the dependency-based model derives from a greater ability to discern similarity from association.

Our comparisons provided less support for the second (window size) hypothesis. As shown in Figure 10, there is a negligible improvement in the performance of the model when the window size is reduced from 10 to 2. However, for the SVD model we observed the converse. The SVD model with window size 10 slightly outperforms the SVD model with window 2, and this improvement is quite pronounced on the most associated pairs in SimLex-999.

Figure 12
The importance of dependency-focussed contexts (in the Levy & Goldberg model) for capturing concepts of different POS, when compared to a standard Skipgram (BOW) model trained on the same Wikipedia corpus.

Learning concepts of different POS. Given the theoretical likelihood of variation in model performance across POS categories noted in Section 2.2, we evaluated the **?**, VSM and SVD models on the subsets of SimLex-999 containing adjective, noun and verb concept pairs.

The analyses yield two notable conclusions, as shown in Figure 11. First, perhaps contrary to intuition, all models estimate the similarity of adjectives better than other concept categories. This aligns with the (also unexpected) observation that humans rate the similarity of adjectives more consistently and with more agreement than other parts of speech (see the dashed lines). However, the parallels between human raters and the models do not extend to verbs and nouns; verb similarity is rated more consistently than noun similarity by humans, but models estimate these ratings more accurately for nouns than for verbs.

To better understand the linguistic information exploited by models when acquiring concepts of different POS, we also computed performance on the POS subsets of SimLex-999 of the dependency-based model of **?** and the standard skipgram model, in which linguistic contexts are encoded as simple bags-of-words (BOW) **?** (trained on the same Wikipedia text). As shown in Figure 12, dependency-aware contexts yield the largest improvements for capturing verb similarity. This aligns with the cognitive theory of verbs as *relational concepts* (**?**) whose meanings rely on their interaction with (or dependency on) other words or concepts. It is also consistent with research on the automatic acquisition of verb semantics, in which syntactic features have proven particularly important (**?**). While a deeper exploration of these effects is beyond the scope of this work, this preliminary analysis again highlights the how the word classes integrated into SimLex-999 are pertinent to a range of questions concerning lexical semantics.

Learning concrete and abstract concepts. Given the strong interdependence between POS and conceptual concreteness (Figure 1), we aimed to explore whether the variation in model performance on different POS categories was in fact driven by an underlying effect of concreteness. To do so, we ranked each pair in the SimLex-999 dataset according
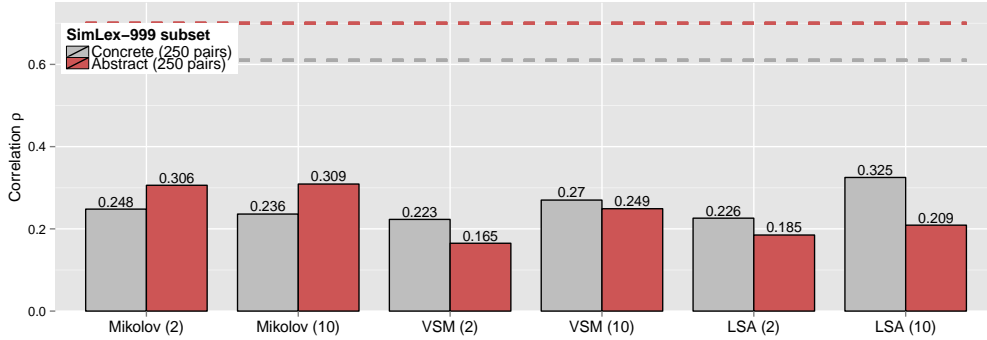
Figure 13
Performance of models on concreteness-based subsets of SimLex-999. Window size is indicated in parentheses. Horizontal dashed lines indicate inter-annotator agreement between SimLex-999 annotators on the two subsets.

to the sum of the concreteness of the two words, and compared performance of models on the most concrete and least concrete quartiles according to this ranking (Figure 13).

Interestingly, the performance of models on the most abstract and most concrete pairs suggests that the distinction characterised by concreteness is at least partially independent of POS. Specifically, while the Mikolov et al. model was the highest performer on all POS categories, its performance was worse than both the simple VSM and SVD models (of window size 10) on the most concrete concept pairs.

This finding supports the growing evidence for systematic differences in representation and/or similarity operations between abstract and concrete concepts (**?**), and suggests that at least part of these concreteness effects are independent of POS. In particular, it appears that models built from underlying vectors of co-occurrence counts, such as VSMs and SVD, are better equipped to capture the semantics of concrete entities, whereas the embeddings learned by NLMs can better capture abstract semantics.

6. Conclusion

Although the ultimate test of semantic models should be their utility in downstream applications, the research community can undoubtedly benefit from ways to evaluate the general quality of the representations learned by such models, prior to their integration in any particular system. We have presented SimLex-999, a gold standard resource for the evaluation of semantic representations containing similarity ratings of word pairs of different POS categories and concreteness levels.

The development of SimLex-999 was principally motivated by two factors. First, as we demonstrated, several existing gold standards measure the ability of models to capture association rather than similarity, and others do not adequately test their ability to discriminate similarity from association. This is despite the many potential applications for accurate similarity-focussed representation learning models. Analysis of the ratings of the 500 SimLex-999 annotators showed that subjects can consistently quantify similarity, as distinct from association, and apply it to various concept types, based on minimal intuitive instructions.

Second, as we showed, state-of-the-art models trained solely on running-text corpora have now reached or surpassed the human agreement ceiling on WordSim-353 and MEN, the most popular existing gold standards, as well as on RG and WS-Sim. These evaluations may therefore have limited use in guiding or moderating future improvements to distributional semantic models. Nevertheless, there is clearly still room for improvement in terms of the use of distributional models in functional applications. We therefore consider the comparatively low performance of state-of-the-art models on SimLex-999 to be one of its principal strengths. There is clear room under the inter-rating ceiling to guide the development of the next generation of distributional models.

We conducted a brief exploration of how models might improve on this performance, and verified the hypotheses that models trained on dependency-based input capture similarity more effectively than those trained on running-text input. The evidence that smaller context windows are also beneficial for similarity models was mixed, however. Indeed, we showed that the optimal window size depends on both the general model architecture and the part-of-speech and concreteness of the target concepts.

Our analysis of these hypotheses illustrates how the design of SimLex-999 - covering a principled set of concept categories and including meta-information on concreteness and free-association strength - enables fine-grained analyses of the performance and parameterization of semantic models. However, these experiments only scratch the surface in terms of the possible analyses. We hope that researchers will adopt the resource as a robust means of answering a diverse range of questions pertinent to similarity modelling, distributional semantics and representation learning in general.

In particular, for models to learn high-quality representations for all linguistic concepts, we believe that future work must uncover ways to explicitly or implicitly infer 'deeper', more general, conceptual properties such as intentionality, polarity, subjectivity or concreteness (**?**). However, while improving corpus-based models in this direction is certainly realistic, models that learn exclusively via the linguistic modality may never reach human-level performance on evaluations such as SimLex-999. This is because much conceptual knowledge, and particularly that which underlines similarity computations for concrete concepts, appears to be grounded in the perceptual modalities as much as in language (**?**).

Whatever the means by which the improvements are achieved, accurate concept-level representation is likely to constitute a necessary first step towards learning informative, language-neutral phrasal and sentential representations. Such representations would be hugely valuable for fundamental NLP applications such as language understanding tools and machine translation.

Distributional semantics aims to infer the meaning of words based on the *company they keep* (**?**). However, while words that occur together in text often have associated meanings, these meanings may be very similar or indeed very different. Thus, possibly excepting the population of Argentina, most people would agree that, strictly speaking, *Maradona* is not synonymous with *football* (despite their high rating of 8.62 in WordSim-353). The challenge for the next generation of distributional models may therefore be to infer what is useful from the co-occurrence signal and to overlook what is not. Perhaps only then will models capture most, or even all, of what humans know when they know how to use a language.

References

[Agirre et al.2009]Agirre, Eneko, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proccedings of NAACL*. Association for Computational Linguistics.

[Alfonseca and Manandhar2002]Alfonseca, Enrique and Suresh Manandhar. 2002. Extending a lexical ontology by a combination of distributional semantics signatures. In *Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web*. Springer, pages 1–7.

[Andrews, Vigliocco, and Vinson2009] Andrews, Mark, Gabriella Vigliocco, and David Vinson. 2009. Integrating experiential and distributional data to learn semantic representations. *Psychological review*, 116(3):463.

[Bansal, Gimpel, and Livescu2014]Bansal, Mohit, Kevin Gimpel, and Karen Livescu. 2014. Tailoring continuous word representations for dependency parsing. In *Proceedings of ACL*.

[Baroni, Dinu, and Kruszewski2014]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*, volume 1.

[Baroni and Lenci2010]Baroni, Marco and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

[Barsalou et al.2003]Barsalou, Lawrence W, W Kyle Simmons, Aron K Barbey, and Christine D Wilson. 2003. Grounding conceptual knowledge in modality-specific systems. *Trends in cognitive sciences*, 7(2):84–91.

[Beltagy, Erk, and Mooney2014]Beltagy, Islam, Katrin Erk, and Raymond Mooney. 2014. Semantic parsing using distributional semantics and probabilistic logic. In *ACL 2014 Workshop on Semantic Parsing*.

[Bengio et al.2003]Bengio, Yoshua, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *The Journal of Machine Learning Research*, 3:1137–1155.

[Bernardi et al.2013]Bernardi, Raffaella, Georgiana Dinu, Marco Marelli, and Marco Baroni. 2013. A relatedness benchmark to test the role of determiners in compositional distributional semantics. *Proceedings of ACL*.

[Biemann2005]Biemann, Chris. 2005. Ontology learning from text: A survey of methods. In *LDV forum*, volume 44, pages 75–93.

[Bird2006]Bird, Steven. 2006. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 69–72. Association for Computational Linguistics.

[Bruni et al.2012a]Bruni, Elia, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. 2012a. Distributional semantics in technicolor. In *Proceedings of ACL*. Association for Computational Linguistics.

[Bruni et al.2012b]Bruni, Elia, Jasper Uijlings, Marco Baroni, and Nicu Sebe. 2012b. Distributional semantics with eyes: Using image analysis to improve computational representations of word meaning. In *Proceedings of the 20th ACM international conference on Multimedia*. ACM.

[Budanitsky and Hirst2006]Budanitsky, Alexander and Graeme Hirst. 2006. Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47.

[Cimiano, Hotho, and Staab2005]Cimiano, Philipp, Andreas Hotho, and Steffen Staab. 2005. Learning concept hierarchies from text corpora using formal concept analysis. *J. Artif. Intell. Res.(JAIR)*, 24:305–339.

[Collobert and Weston2008]Collobert, R. and J. Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *International Conference on Machine Learning, ICML*.

[Cruse1986]Cruse, D Alan. 1986. *Lexical semantics*. Cambridge University Press.

[Cunningham2005]Cunningham, Hamish. 2005. Information extraction, automatic. *Encyclopedia of language and linguistics,*, pages 665–677.

[Fellbaum1998]Fellbaum, Christiane. 1998. *WordNet*. Wiley Online Library.

[Finkelstein et al.2001]Finkelstein, Lev, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*, pages 406–414. ACM.

[Firth1957]Firth, J.R. 1957. *Papers in Linguistics 1934âĂŞ1951*. Oxford University Press.

[Gentner1978]Gentner, Dedre. 1978. On relational meaning: The acquisition of verb meaning. *Child development*, pages 988–998.

[Gentner2006]Gentner, Dedre. 2006. Why verbs are hard to learn. *Action meets word: How children learn verbs*, pages 544–564.

[Gershman and Dyer2014]Gershman, Yulia Tsvetkov Leonid Boytsov Anatole and Eric Nyberg Chris Dyer. 2014. Metaphor detection with cross-lingual model

transfer. In *Proceedings of ACL*.

[Golub and Reinsch1970]Golub, Gene H and Christian Reinsch. 1970. Singular value decomposition and least squares solutions. *Numerische Mathematik*, 14(5):403–420.

[Griffiths, Steyvers, and Tenenbaum2007] Griffiths, Thomas L, Mark Steyvers, and Joshua B Tenenbaum. 2007. Topics in semantic representation. *Psychological review*, 114(2):211.

[Haghighi et al.2008]Haghighi, Aria, Percy Liang, Taylor Berg-Kirkpatrick, and Dan Klein. 2008. Learning bilingual lexicons from monolingual corpora. In *Proceedings of ACL 2008*.

[Hassan and Mihalcea2011]Hassan, Samer and Rada Mihalcea. 2011. Semantic relatedness using salient semantic analysis. In *AAAI*.

[Hatzivassiloglou et al.2001] Hatzivassiloglou, Vasileios, Judith L Klavans, Melissa L Holcombe, Regina Barzilay, Min-Yen Kan, and Kathleen McKeown. 2001. Simfinder: A flexible clustering tool for summarization. In *NAACL Workshop on Automatic Summarization*.

[He et al.2008]He, Xiaodong, Mei Yang, Jianfeng Gao, Patrick Nguyen, and Robert Moore. 2008. Indirect-hmm-based hypothesis alignment for combining outputs from machine translation systems. In *Proceedings of EMNLP*, pages 98–107. Association for Computational Linguistics.

[Hill, Kiela, and Korhonen2013]Hill, Felix, Douwe Kiela, and Anna Korhonen. 2013. Concreteness and corpora: A theoretical and practical analysis. *CMCL 2013*, page 75.

[Hill, Korhonen, and Bentz2013]Hill, Felix, Anna Korhonen, and Christian Bentz. 2013. A quantitative empirical analysis of the abstract/concrete distinction. *Cognitive science*.

[Hill, Reichart, and Korhonen2014]Hill, Felix, Roi Reichart, and Anna Korhonen. 2014. Multi-modal models for concrete and abstract concept meaning. *Transactions of the Association for Computational Linguistics (TACL)*.

[Huang et al.2012]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 ACL*, pages 873–882. Association for Computational Linguistics.

[Kiela and Clark2014]Kiela, Douwe and Stephen Clark. 2014. A systematic study of

semantic vector space model parameters. In *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC)@ EACL*, pages 21–30.

[Kiela et al.2014]Kiela, Douwe, Felix Hill, Anna Korhonen, and Stephen Clark. 2014. Improving multi-modal representations using image dispersion: Why less is sometimes more. In *Proceedings of ACL*. ACL.

[Landauer and Dumais1997]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.

[Leech, Garside, and Bryant1994]Leech, Geoffrey, Roger Garside, and Michael Bryant. 1994. Claws4: the tagging of the british national corpus. In *Proceedings of COLING*, pages 622–628. Association for Computational Linguistics.

[Levy and Goldberg2014]Levy, Omer and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of ACL*, volume 2.

[Levy et al.2015]Levy, Omer, Steffen Remus, Chris Biemann, and Idol Dagan. 2015. Do supervised distributional methods really learn lexical inference relations? *Proceedings of NAACL*.

[Lewis et al.2004]Lewis, David D, Yiming Yang, Tony G Rose, and Fan Li. 2004. Rcv1: A new benchmark collection for text categorization research. *The Journal of Machine Learning Research*, 5:361–397.

[Li et al.2014]Li, Changliang, Bo Xu, Gaowei Wu, Xiuying Wang, Wendong Ge, and Yan Li. 2014. Obtaining better word representations via language transfer. In *Computational Linguistics and Intelligent Text Processing*. Springer, pages 128–137.

[Li et al.2006]Li, Mu, Yang Zhang, Muhua Zhu, and Ming Zhou. 2006. Exploring distributional similarity based models for query spelling correction. In *Proceedings of ALC*, pages 1025–1032. Association for Computational Linguistics.

[Luong, Socher, and Manning2013]Luong, Minh-Thang, Richard Socher, and Christopher D Manning. 2013. Better word representations with recursive neural networks for morphology. *CoNLL-2013*, 104.

[Markman and Wisniewski1997]Markman, Arthur B and Edward J Wisniewski. 1997. Similar and different: The differentiation of basic-level categories. *Journal of*

Experimental Psychology: Learning, Memory, and Cognition, 23(1).

[Marton, Callison-Burch, and Resnik2009] Marton, Yuval, Chris Callison-Burch, and Philip Resnik. 2009. Improved statistical machine translation using monolingually-derived paraphrases. In *Proceedings of EMNLP*, pages 381–390. Association for Computational Linguistics.

[McRae, Khalkhali, and Hare2012]McRae, Ken, Saman Khalkhali, and Mary Hare. 2012. Semantic and associative relations in adolescents and young adults: Examining a tenuous dichotomy. In Valerie F Reyna, Sandra B Chapman, Michael R Dougherty, and Jere Ed Confrey, editors, *The adolescent brain: Learning, reasoning, and decision making.* American Psychological Association.

[Medelyan et al.2009]Medelyan, Olena, David Milne, Catherine Legg, and Ian H Witten. 2009. Mining meaning from wikipedia. *International Journal of Human-Computer Studies*, 67(9):716–754.

[Mikolov et al.2013a]Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *Proceedings of International Conference of Learning Representations*, Scottsdale, Arizona, USA.

[Mikolov et al.2013b]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.

[Navigli2009]Navigli, Roberto. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)*, 41(2):10.

[Nelson, McEvoy, and Schreiber2004]Nelson, Douglas L, Cathy L McEvoy, and Thomas A Schreiber. 2004. The university of south florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, 36(3):402–407.

[Padó, Padó, and Erk2007]Padó, Sebastian, Ulrike Padó, and Katrin Erk. 2007. Flexible, corpus-based modelling of human plausibility judgements. In *Proceedings of EMNLP-CoNLL*, pages 400–409, Prague, Czech Republic, June. Association for Computational Linguistics.

[Paivio1991]Paivio, Allan. 1991. Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology/Revue canadienne de psychologie*, 45(3):255.

[Pedersen, Patwardhan, and Michelizzi2004] Pedersen, Ted, Siddharth Patwardhan, and Jason Michelizzi. 2004. Wordnet:: Similarity: measuring the relatedness of concepts. In *Demonstration Papers at HLT-NAACL 2004*, pages 38–41. Association for Computational Linguistics.

[Phan, Nguyen, and Horiguchi2008]Phan, Xuan-Hieu, Le-Minh Nguyen, and Susumu Horiguchi. 2008. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *Proceedings of the 17th international conference on World Wide Web*, pages 91–100. ACM.

[Plaut1995]Plaut, David C. 1995. Semantic and associative priming in a distributed attractor network. In *Proceedings of CogSci*, volume 17, pages 37–42.

[Recchia and Jones2009]Recchia, Gabriel and Michael N Jones. 2009. More data trumps smarter algorithms: Comparing pointwise mutual information with latent semantic analysis. *Behavior research methods*, 41(3):647–656.

[Reisinger and Mooney2010a]Reisinger, Joseph and Raymond Mooney. 2010a. A mixture model with sharing for lexical semantics. In *Proceedings of EMNLP*, pages 1173–1182. Association for Computational Linguistics.

[Reisinger and Mooney2010b]Reisinger, Joseph and Raymond J Mooney. 2010b. Multi-prototype vector-space models of word meaning. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 109–117. Association for Computational Linguistics.

[Resnik1995]Resnik, Philip. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of IJCAI*.

[Resnik and Lin2010]Resnik, Philip and Jimmy Lin. 2010. 11 evaluation of nlp systems. *The handbook of computational linguistics and natural language processing*, 57:271.

[Rosch, Simpson, and Miller1976]Rosch, Eleanor, Carol Simpson, and R Scott Miller. 1976. Structural bases of typicality effects. *Journal of Experimental Psychology: Human perception and performance*, 2(4):491.

[Rose, Stevenson, and Whitehead2002]Rose, Tony, Mark Stevenson, and Miles Whitehead. 2002. The reuters corpus volume 1-from yesterday's news to tomorrow's language resources. In *LREC*,

volume 2, pages 827–832.

[Rubenstein and Goodenough1965] Rubenstein, Herbert and John B Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

[Silberer and Lapata2014]Silberer, Carina and Mirella Lapata. 2014. Learning grounded meaning representations with autoencoders. In *Proceedings of ACL*. Association for Computational Linguistics.

[Sun, Korhonen, and Krymolowski2008]Sun, Lin, Anna Korhonen, and Yuval Krymolowski. 2008. Verb class discovery from rich syntactic data. In *Computational linguistics and intelligent text processing*. Springer, pages 16–27.

[Turian, Ratinov, and Bengio2010]Turian, Joseph, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of ACL*, pages 384–394. Association for Computational Linguistics.

[Turney2012]Turney, Peter D. 2012. Domain and function: A dual-space model of semantic relations and compositions. *Journal of Artificial Intelligence Research (JAIR)*, 179(44):533–585.

[Turney and Pantel2010]Turney, Peter D and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37(1):141–188.

[Tversky1977]Tversky, Amos. 1977. Features of similarity. *Psychological review*, 84(4):327.

[Wiebe2000]Wiebe, Janyce. 2000. Learning subjective adjectives from corpora. In *AAAI/IAAI*, pages 735–740.

[Williams and Anand2009]Williams, Gbolahan K and Sarabjot Singh Anand. 2009. Predicting the polarity strength of adjectives using wordnet. In *ICWSM*.

[Wu and Palmer1994]Wu, Zhibiao and Martha Palmer. 1994. Verbs semantics and lexical selection. In *Proceedings of ACL*, pages 133–138. Association for Computational Linguistics.

[Yong and Foo1999]Yong, Chung and Shou King Foo. 1999. A case study on inter-annotator agreement for word sense disambiguation. In *Proceedings of the ACL SIGLEX Workshop on Standardizing Lexical Resources (SIGLEX99)*.