# Taxonomy Beats Corpus in Similarity Identification, but Does It Matter?

**Minh Ngoc Le**
VU University Amsterdam
m.n.le@vu.nl

**Antske Fokkens**
VU University Amsterdam
antske.fokkens@vu.nl

## Abstract

We present extensive evaluations comparing the performance of taxonomy-based and corpus-based approaches on SimLex-999. The results confirm our hypothesis that taxonomy-based approaches are more suitable to identify similarity. We introduce two new measures of evaluation that show that all measures perform well on a coarse-grained evaluation and that it is not always clear which approach is most suitable when a similarity score is used as a threshold. This leads us to conclude that the inferior performance of corpus-based approaches may not (always) matter.

## 1 Introduction

Similarity measures are used in a wide variety of Natural Language Processing (NLP) tasks (see Pilehvar et al. (2013), among others for examples). They may be used, e.g. to increase coverage of an approach by using information from similar words for unseen data, or to establish average similarity between a question and a potential answer.

Due to its importance, similarity measures have received steady attention in computational linguistics. There are two widely followed, but different, schools: taxonomy-based approaches and distributional, or corpus-based, approaches. Apart from a few exceptions, these approaches have mostly been studied separately.

Our main goal is to examine how the approaches perform when identifying true similarity, in contrast to the more general relatedness, which also includes association, between word-pairs. We evaluate the approaches on the new gold-standard SimLex-999 (Hill et al., 2014b). We compare taxonomy-based approaches that use WordNet (Fellbaum, 1998) to the corpus-based approaches that performed best on SimLex-999 in Hill et al. (2014a). We hypothesize that taxonomy-based approaches outperform corpus-based approaches on a true similarity set, because corpus-based approaches tend to mix-up similarity and association.

We carry out several evaluations which investigate (i) the difference in performance on pure similarity sets and sets that combine similarity and association, (ii) the influence of associative pairs while identifying true similarity, and (iii) various evaluation metrics that compare similarity measures to the gold standard of SimLex-999.

We perform more than one evaluation metric for two reasons. First, different ranking coefficients can lead to a completely different outcome when evaluating similarity scores (Fokkens et al., 2013). Second, we want to gain more insight into the differences between individual measures. To do so, we introduced two new, more flexible, evaluation methods which reveal high results for all similarity measures. We argue that these new evaluations provide a better insight into how suitable similarity measures are to be used in NLP tasks than the commonly used Spearman's correlation (henceforth Spearman $\rho$).

Our results show that most of the evaluations confirm our hypothesis. The few cases where corpus-based methods outperformed taxonomy-based approaches reveal much smaller differences than the many cases where taxonomy-based approaches have higher results. However, all similarity measures perform very well when they are evaluated on the relative ranking of word-pairs that are further apart in the gold-standard. We therefore conclude that, even though taxonomy-based are better at identifying similarity than corpus-based approaches, this may not (always) matter.

The rest of this paper is structured as follows. In Section 2, we motivate our approach and address related work. Section 3 describes the similarity measures we investigate. In Section 4, we



*Proceedings of Recent Advances in Natural Language Processing*, pages 346–354,
Hissar, Bulgaria, Sep 7–9 2015.

outline our experimental methodology, including used datasets and evaluation methods. The results are presented in Section 5, and our conclusions and future work in Section 6.

## 2 Background and Motivation

Several gold-standards have been created that rank word-pairs based on their similarity. Agirre et al. (2009) point out that association and similarity are mixed up in these sets, where associated pairs such as *coffee* and *cup* rank higher than truly similar pairs such as *car* and *train*. The confusion directly influences the performance of corpus-based approaches, which also tend to have difficulties distinguishing association from similarity (Hill et al., 2014a).

Hill et al. (2014b) introduce a new gold standard dataset that is annotated with pure semantic similarity and larger than previously created similarity sets, such as Rubenstein and Goodenough (1965) and Agirre et al. (2009)'s sets. Hill et al. (2014a) evaluate corpus-based approaches and show that they indeed have trouble identifying similarity, performing well-below the upperbound of agreement between human annotators.

It is not surprising that corpus-based approaches confuse similarity and association: semantically related words tend to occur close to each other and hence in similar contexts. Approaches that make use of a relatively narrow context window perform slightly better, because they can capture more subtle differences in context to some extend.

Taxonomies represent word meanings in hypernym and hyponym hierarchies, directly capturing their similarity. The closer two terms are in the hierarchy, the more similar they are. Similarity measures that make use of this structure are less likely to confuse whether two terms are similar or related in some other way.

These well-known properties of corpus-based and taxonomy-based approaches led to the following hypothesis:

> Taxonomy-based approaches are better suited to identify similarity than corpus-based approaches

Agirre et al. (2009) seem to contradict this hypothesis showing that corpus-based approaches can be as good at identifying similarity (when the right model is based on enough data). However, Hill et al. (2014b) point out that Agirre et al.'s evaluation set does not form a representative set for measuring similarity, even after they made an alternative set that separates association and similarity. We therefore expected that the hypothesis would nevertheless hold on SimLex-999.

The outcome of our experiments confirmed our hypothesis, thus contradicting Agirre et al. (2009)'s results and being, to our knowledge, the first to show this on such a large and reliable benchmark. Banjade et al. (2015) also applies WordNet-based and corpus-based similarity measures to SimLex-999, but do not examine or discuss the difference between taxonomy-based approaches and corpus-based approaches in detail. Instead, they focus on the strength of combining several approaches to yield better results.[1] We investigate the difference between the approaches in various evaluations showing that taxonomy-based approaches outperform corpus-based approaches, a conclusion that cannot be drawn (clearly) from Banjade et al. (2015)'s results. It should be noted that our conclusions only apply to the task of identifying pure similarity. Markert and Nissim (2005) show, for instance, that a corpus-based approach with sufficiently large corpus works better than WordNet for anaphora resolution.

The next step in our investigation was to determine the strengths and weaknesses of each approach. The original idea was to investigate pairs that are ranked more or less correctly by one approach, but are far off in the other to identify patterns of errors in each approach. We did not find such patterns, partially because the examples that have large differences in ranking compared to the gold are relatively rare.

We therefore developed two alternative evaluation methods that are less sensitive to minor differences in ranking. The first evaluation directly tests the comparison of pairs and, more importantly, allows us to study the contribution of partitions of the dataset. The second evaluation revolves around thresholds for similarity. In this evaluation, we set thresholds to establish a binary distinction between highly similar pairs and other pairs. The pairs above the similarity threshold are compared to those falling above the threshold in the gold (see Section 4.2).

Many studies compare similarity measures (see Baroni et al. (2014) and Pedersen (2010), among

[1] We independently confirmed this result in our own experiments, but decided to leave it out of this paper because our results did not add much to Banjade et al. (2015).

others) but, to our knowledge, Agirre et al. (2009) and Banjade et al. (2015) are the only ones that look at both taxonomy-based approaches and distributional approaches. As mentioned above, they do not dive into the details of the differences between the two. Furthermore, apart from Fokkens et al. (2013), who do not propose new rankings, we are not aware of studies applying multiple evaluation metrics for similarity-based rankings.

## 3 Similarity Measures

This section describes the similarity measures compared in this paper.

### 3.1 Taxonomy-based Similarity Measures

WordNet (Fellbaum, 1998) organizes nouns and verbs in hierarchies of hypernym-hyponym relations. We selected WordNet for our taxonomy-based experiments, because it is widely used and probably the most popular taxonomy when it comes to determining word similarity. Many measures of similarity based on WordNet have been proposed over the years. Early work (Rada et al., 1989) advocates the use of *is-a* hierarchy and later approaches continue to use it heavily. In order to make a clean comparison between WordNet and distributional models, we do not include in our study measures that make use of a corpus such as Resnik (1995) and Jiang and Conrath (1997).

**Path length similarity** takes the inverse of the path length (i.e. the distance in number of nodes) from $s_1$ to $s_2$ plus one.

$$\text{PL} = \frac{1}{\text{d}(s_1, s_2) + 1}$$

**Wu and Palmer's similarity** (Wu and Palmer, 1994) takes the fact into account that senses deeper in the hierarchy tend to be more specific than those high up. It therefore incorporates the depth of the hierarchy in their similarity calculation:

$$\text{WUP} = \frac{2\text{depth}(lcs)}{\text{d}(s_1, lcs) + \text{d}(s_2, lcs) + 2\text{depth}(lcs)}$$

**Leacock and Chodorows similarity** (Leacock and Chodorow, 1998) normalizes path-based scores by the maximum depth $D$ of the hierarchy. This corrects for the difference in the depth of verb and noun hierarchy:

$$\text{LCH} = -\log \frac{\text{d}(s_1, s_2) + 1}{2D}$$

### 3.2 Distributional Semantic Models

We selected two representative models from the large and growing literature on corpus-based models of lexical semantics: Word2vec (Mikolov et al., 2013, W2V) and dependency-based word embeddings (Levy and Goldberg, 2014a, DEPS).

Word2vec is the first model to use a Skip-Gram with Negative Sampling (SGNN) algorithm for constructing semantic models and performed best on SimLex-999 in Hill et al. (2014a). Levy and Goldberg (2014b) argue that SGNN implicitly factorizes a shifted positive mutual information word-context matrix, not unlike traditional distributional semantic models. The use of a small window size and the weighting scheme that favors nearby contexts are supported by a systematic study of Kiela and Clark (2014) that shows the superiority of small windows. Moreover, Sahlgren (2006) presents empirical evidence that smaller windows lead to a cleaner distinction between syntagmatic and paradigmatic relations (which can be considered the linguistic version of similarity and association).

Levy and Goldberg (2014a) extend SGNN to work with arbitrary contexts and experiment with dependency structures. It is generally believed that dependency structures are better at capturing similarity (Padó and Lapata, 2007) although Kiela and Clark (2014) found mixed results.

The Skip-gram model captures the distribution $p(c|t)$ of a context word $c$ within a certain window around a target word $t$. For a vocabulary of millions, computing normalized probabilities (i.e. summing to one) for each example can be prohibitively expensive. Negative sampling was used to avoid the cost.

For each context-target pair $(c, t)$ taken from training data, we replace the context by random words drawn from the vocabulary to obtain new pairs $\{(c', t)\}$. We call $D \ni (c, t)$ *positive distribution* and $N \ni (c', t)$ *negative distribution*. The task of the model is to identify which pairs come from $D$ and which from $N$. Formally. that is to maximize the negative log likelihood:

$$\ell = -\left(\sum \log p(D|c, t) + \sum \log p(N|c', t)\right)$$

The probability is calculated using *target embeddings* $e_t \in \mathbb{R}^d$ and *context embeddings* $\hat{e}_c \in \mathbb{R}^d$ such that:

$$p(D|c, t) = \sigma(e_t \cdot \hat{e}_c),$$

where $\sigma(x) = 1/(1+e^{-x})$ is a monotonic function that maps any value in $(-\infty, +\infty)$ to a valid probability.

The training objective encourages to increase $p(D|c,t)$ which can be achieved by aligning $e_t$ and $\hat{e}_c$ in similar directions. On the other hand, the objective also encourages a small $p(N|c,t)$, creating an uniform "repelling force" between all pairs of words. After a lot of updating iterations, similar words come close together while dissimilar words are pulled apart.

We used the trained embeddings from Mikolov et al. (2013) and Levy and Goldberg (2014a).[2] Word2vec embeddings are 300-dimensional vectors obtained by training on 100 billion words of Google News dataset. Dependency-based embeddings were harvested from English Wikipedia automatically annotated with dependency structures. Although the dependency-based model was trained on a significantly smaller corpus, it achieves comparable results as we will show in Section 5.

## 4 Experimental Setup

In this section, we describe the experimental setup used in our evaluations. We first describe the datasets and then the evaluation metrics we use.

### 4.1 Gold-standard Datasets

We evaluate the approaches on three datasets. WordSim-353 and MEN allow us to compare performance on sets that mix association and similarity. SimLex-999's ranking is based on similarity only.

**WordSim-353** (Finkelstein et al., 2001) includes 353 word pairs scored for *relatedness* on a scale from 0 to 10 by 13 or 16 subjects. The inter-annotator agreement is 0.611 defined as the average pairwise Spearman's correlation. Researchers have reported correlation as high as 0.81 (Yih and Qazvinian, 2012). Agirre et al. (2009) later divided WordSim-353 into a "similarity" and "relatedness" set. However, Hill et al. (2014b) rightly point out that both remain relatedness datasets, because this is what the annotators rated.

**MEN** (Bruni et al., 2012) is composed of 3,000 word pairs, sampled to include a balanced range of relatedness. Annotators were asked to choose which of two pairs of words is more related, an arguably more intuitive task than assigning a score.

**SimLex-999** (Hill et al., 2014b) carefully distinguishes between similarity and association and provides a balanced range of similarity, concreteness and parts-of-speech. The authors sampled 900 associated pairs from the University of South Florida Free Association Database (Nelson et al., 2004) and randomly coupled them to create 999 unassociated pairs. Subjects were asked to judge the similarity of word pairs on a 0-6 scale. Their answers were averaged to produce the final score.

All three datasets are lemma-based. The way two words can be compared, however, is more likely via their *senses* (e.g. *queen* is not similar to *princess* when referring to a chess piece). We follow Resnik (1995) in using maximally similar senses in our taxonomy-based approaches.

### 4.2 Evaluation Metrics

The first evaluation measure we use compares between the gold ranking and a measurement's ranking using **Spearman's $\rho$** , the most widely used evaluation metric for similarity score.

Hill et al. (2014b) report performance on a subset of highly associated word pairs, but its contribution to the overall performance is unclear. We wish to gain deeper insight into how different subsets in the data contribute to the overall score. This is not possible with Spearman's $\rho$ due to its holistic nature. We overcome this by using **ordering accuracy** following Agirre et al. (2009). The scale is defined as:

$$a = a_{G,G} = \frac{1}{|G|^2} \sum_{(u,v)\in G} \sum_{(x,y)\in G} m_{s,G}(u,v,x,y)$$

where $G$ stands for the gold standard and $m_{s,G}(\cdot)$ is a matching function that returns 1 for those two word-pairs whose relative ranking is the same in the gold standard and in the ranking of the similarity measure and 0 otherwise. We also experiment with a variation of $m$ where ties get half score. As shown in Figure 1, ordering accuracy highly correlates with Spearman's $\rho$.

If $G$ can be partitioned into $n$ subsets $g_i$ (i.e. $\bigcap g_i = \emptyset$ and $\bigcup g_i = G$) then $a$ can be decomposed as the weighted sum of the accuracy on different subsets. The weights are proportional to their size:

$$a = \frac{1}{|G|^2} \sum_i \sum_j |g_i||g_j| a_{g_i,g_j}$$

[2] The models are available at: https://code.google.com/p/word2vec/ and https://levyomer.wordpress.com/2014/04/25/dependency-based-word-embeddings

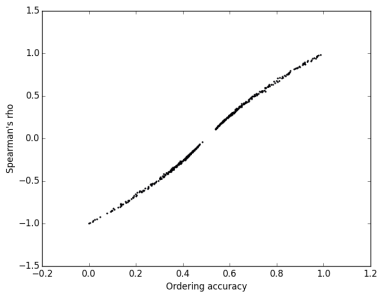


Figure 1: Ordering accuracy and Spearman's $\rho$ on a synthesized dataset of 100 word pairs.

| Model | SL-999$_{nv}$ | MEN$_{nv}$ | WS-353 |
|---|---|---|---|
| WUP | 0.47 | 0.39 | 0.35 |
| PL | 0.52 | 0.39 | 0.30 |
| LCH | 0.55 | 0.39 | 0.31 |
| W2V | 0.42 | 0.77 | 0.70 |
| DEPS | 0.45 | 0.61 | 0.63 |

Table 1: Spearman's correlation of models to similarity benchmarks.

The final evaluation measure is based on the observation that many approaches use a threshold to determine which words are similar enough to be used for contributing features or approximations, or to be candidates for lexical substitution (McCarthy and Navigli, 2009; Biran et al., 2011, e.g.). **Threshold accuracy** sets a similarity threshold and determines how many of the $n$-highest ranking word pairs in a given measurement are also in the top-$n$ pairs of the gold standard. In other words, this evaluation determines whether the right word-pairs would end up above the threshold of being similar.

## 5 Results

We calculated the similarity scores of all noun and verb pairs in SimLex-999 (a set of 888 pairs), MEN (2,034 pairs), and all pairs in WordSim-353 using the measures outlined in Section 3 and ranked the word pairs according to the outcome.

### 5.1 Spearman's Rank Correlation

Table 1 shows the performance of models on all three benchmarks. Taxonomy based approaches perform higher on SimLex-999, whereas corpus-based approaches reveal high performance on MEN and WordSim-353 and score significantly lower on SimLex-999. This result confirms that taxonomy-based approaches capture similarity rather than association, whereas corpus-based approaches do not clearly distinguish the two.

| Model | SL-999 *nv* | SL-999 *nv* | SL-999 *nv,assoc* | Diff. *assoc* |
|---|---|---|---|---|
| | | Using tie corrections | | |
| WUP | 64.9 | 66.6 | 67.3 | +0.7 |
| PL | 61.1 | 68.0 | 68.2 | +0.2 |
| LCH | 65.1 | 69.2 | 69.1 | -0.1 |
| W2V | 64.4 | 64.6 | 57.5 | -7.1 |
| DEPS | 65.5 | 65.6 | 60.9 | -4.7 |

Table 2: Ordering accuracy (percentage) of similarity measures on SimLex-999$_{nv}$.

### 5.2 Ordering Accuracy

Table 2 presents the evaluation of our metrics using ordering accuracy. The first column indicates the standard score. The scores in the second and third column are calculated while giving partial credits to ties. Note that this only affects the performance of taxonomy-based approaches, where it is common for word pairs to have identical scores.

Without correction for ties, scores for taxonomy-based and corpus-based measures are highly similar, with the corpus-based DEPS leading to the highest results. Taxonomy-based approaches uniformly beat corpus-based approaches again when we do correct for ties, confirming the outcome of our Spearman $\rho$ evaluation.

We also evaluate on a subset of highly-associated words. The results are presented in column 3 of Table 2. Sizeable decrease is observed in corpus-based measures for highly associated terms while taxonomy-based measures remain largely unaffected. This result confirms our hypothesis once more that taxonomy-based measures are more suited to capture similarity and that corpus-based methods tend to have difficulties separating similarity from association.

### 5.3 Decomposition of Ordering Accuracy

Palmer et al. (2007) showed that making subtle sense distinction is hard for human subjects leading to evaluations where both coarse-grained and fine-grained word senses are considered (Palmer et al., 2007; Navigli et al., 2007). Similarly, establishing which word-pair is more similar than another is challenging when pairs are close in sim-

| | |
|---|---|
| $\Delta = 0$ | |
| pollution-president | forget-learn |
| take-leave | succeed-try |
| army-squad | girl-child |
| emotion-passion | collect-save |
| sheep-lamb | attention-awareness |
| $\Delta = 1$ | |
| spoon-cup | argue-differ |
| remind-sell | apple-candy |
| book-topic | argument-agreement |
| corporation-business | kidney-organ |
| alcohol-wine | beach-island |

Table 3: Is the pair in the left or in the right more similar? (All pairs are extracted from SimLex-999)

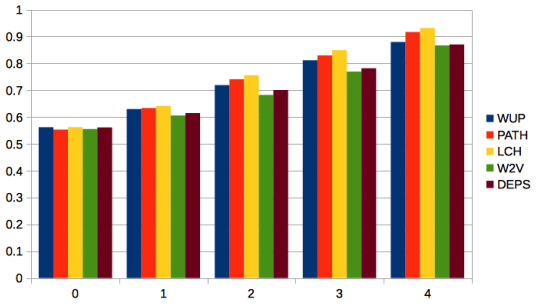


Figure 2: Ordering accuracy varies with degrees of granularity on SimLex-999$_{nv}$. $\Delta = 0$ means two pairs fall in the same range of similarity (e.g. 0-2); $\Delta = 1$ means they fall in neighboring ranges of similarity (e.g. 0-2 and 2-4), etc.

ilarity. This is illustrated by the sample pairs in Table 3. The fact that ranking such pairs is highly challenging for humans leads to the question how meaningful differences in performance of similarities measures on these pairs actually are.

To overcome this issue and gain deeper insight into how often low performance is the result of many small errors piling up and how often it is the result of a set of pairs being ranked completely wrongly, we apply our ordering accuracy to a decomposed dataset. We divide SimLex-999$_{nv}$ into five equal similarity ranges $\{g_i\}$ based on SimLex-999's original ranges. The first range $g_1$ contains highly dissimilar pairs of words with a similarity between 0 and 2. Final set $g_5$ contains very similar or synonymous pairs with a similarity from 8 to 10.

We use different granularity levels $\Delta$ ($\Delta = 0, ..., 4$). Component accuracy is calculated by comparing each pair in $g_i$ to every pair in $g_j$ such that $|i - j| = \Delta$.

The results reported in Figure 2 show that all models perform consistently well on coarse-grained similarity while only marginally beating chance-level at the most fine-grained level. Furthermore, taxonomy-based approaches only outperform corpus-based approaches when comparing pairs that are further apart in the gold ranking.

Because the two most fine-grained components ($\Delta = 0$ and $\Delta = 1$) together have a weight of 58%, the ordering accuracy as reported in Table 2 is dominated by fine-grained similarity comparison. Spearman's $\rho$ highly correlates with ordering accuracy, indicating that fine-grained differences also had a major impact on previous work. It is questionable whether it is really necessary for these measures to capture the small differences in similarity that are even difficult for humans to find. This outcome shows that similarity measures perform better than they seem to do according to recent evaluations in the literature.

### 5.4 Threshold Evaluation

The final evaluation we carry out is the so-called *threshold* evaluation. It evaluates how well a threshold performs that separates highly similar terms from less similar terms based on a specific score. We use the 10% and 20% most similar terms as a starting point. In a total set of 888 examples, this means we compare the top 89 and top 178 pairs of each measurement's output with the top pairs of the gold data. We report on the accuracy (i.e. percentage of pairs correctly classified as highly similar) of each scores. As mentioned above, taxonomy-based approaches often assign the same score to multiple pairs. If this was the case for the pairs around the threshold, we extended the range of comparison as to include all pairs with an identical score. Table 4 provides an overview of the results.

The top-$n$ sets increase significantly for taxonomy-based approaches. Because approaches tend to fare better when the size of the group changes, we calculated the scores for W2V and DEPS with the top-$n$ ranks found in the taxonomy-based scores. Table 5 shows the results of this analysis. The scores of the relevant taxonomy-based approach are repeated in the third row.

The threshold based evaluation shows more

| Model | 10%-based | | 20%-based | |
|---|---|---|---|---|
| | *n* | % | *n* | % |
| WUP | 94 | 42.6 | 191 | 50.3 |
| PATH | 172 | 43.5 | 645 | 80.8 |
| LCH | 172 | 53.5 | 305 | 61.0 |
| W2V | 89 | 32.6 | 178 | 38.2 |
| DEPS | 89 | 33.7 | 178 | 43.8 |

Table 4: Threshold based evaluation, comparing the set of top-$n$ similar pairs

| model | $n$-value | | | | |
|---|---|---|---|---|---|
| | 94 | 172 | 191 | 305 | 645 |
| W2V | 33.0 | 38.4 | 39.8 | 48.5 | **82.0** |
| DEPS | 31.9 | 43.6 | 42.9 | 52.8 | 81.4 |
| taxo. | **42.6** | 43.5/**53.5** | **50.3** | **61.0** | 80.8 |

Table 5: Scores of corpus-based methods on the $n$-values used for taxonomy-based scores.

variation than our other metric. In three out of twelve cases,[3] the corpus-based approach leads to more accurate results than the taxonomy-based score. In combination with the outcome of the accuracy ordering result, this outcome underlines the importance of using a variety of evaluation metrics.

Overall, the outcome seems to confirm that taxonomy-based approaches are better at identifying similarity. First, taxonomy-based approaches outperformed corpus-based approaches on identifying the most accurate pairs. Second, corpus-based approaches only beat taxonomy-based ones in few measures and with comparatively small margins (the largest difference being 1.2%, compared to differences up to 15.1%).

## 6 Discussion and Conclusions

This paper investigated the difference in performance of taxonomy-based approaches and corpus-based approaches on identifying similarity. The outcome of our experiments confirmed our hypothesis that taxonomy-based approaches are better at identifying similarity. This is mainly due to the fact that corpus-based approaches have difficulties distinguishing association from similarity, as also noted by Hill et al. (2014a).

We presented several results that confirm our hypothesis by (i) comparing performance of taxonomy-based and corpus-based methods on a dataset designed to capture similarity, (ii) relating this to the results of the same measures on evaluation sets that measure both association and relatedness, and (iii) looking what the influence is of testing against a set that consists of associated terms.

The results show that taxonomy-based approaches excel at identifying similarity whereas corpus-based approaches yield high results when similarity and association are not distinguished. Furthermore, taxonomy-based approaches are not influenced by association between words whereas performance of corpus-based measures drop when their task is to identify similarity.

We applied more than one evaluation to compare the models' performance on SimLex-999. This was done for two reasons. First, different evaluation measures can sometimes lead to different conclusions even if they are meant to address the same question on the same dataset. This also happened in our evaluation, where ordering accuracy without tie-correction and some thresholds led to different results. Second, the evaluation metrics revealed different aspects of the performance. Most notably, the results of our decomposed ordering accuracy showed that all similarity measures are quite good in a coarse-grained setting.

Together with the mixed outcome of the threshold-evaluation, this shows that corpus-based approaches have good potential to be used when similarity needs to be detected. In particular, when taxonomy-based approaches run into coverage issues, they may be the preferred choice. We therefore believe that it will ultimately depend on the application which approach works best. Future work will need to show whether and how these approaches differ when used in actual applications.[4]

## Acknowledgments

The research for this paper was supported by the Netherlands Organisation for Scientific Research (NWO) via the Spinoza-prize Vossen projects (SPI 30-673, 2014-2019) and the BiographyNet project (Nr. 660.011.308), funded by the Netherlands eScience Center (`http://esciencecenter.nl/`). We would like to thank the anonymous reviewers for their feedback.

[3] We compare eight corpus-based outcomes with one taxonomy score and two with two scores for $n$=172, leading to twelve comparisons in total.

[4] All our code is published on `https://bitbucket.org/ulm4/kcsim`.

## References


Eneko Agirre, 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 *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 19–27, Stroudsburg, PA, USA. Association for Computational Linguistics.

Rajendra Banjade, Nabin Maharjan, Nobal B. Niraula, Vasile Rus, and Dipesh Gautam. 2015. Lemon and tea are not similar: Measuring word-to-word similarity by combining different methods. In *Computational Linguistics and Intelligent Text Processing*, pages 335–346. Springer.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Dont count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 238–247.

Or Biran, Samuel Brody, and Noémie Elhadad. 2011. Putting it simply: a context-aware approach to lexical simplification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 496–501. Association for Computational Linguistics.

Elia Bruni, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. 2012. Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 136–145. Association for Computational Linguistics.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Lev Finkelstein, 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.

Antske Fokkens, Marieke Erp, Marten Postma, Ted Pedersen, Piek Vossen, and Nuno Freire. 2013. Offspring from reproduction problems: What replication failure teaches us. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1691–1701. Association for Computational Linguistics.

Felix Hill, KyungHyun Cho, Sébastien Jean, Coline Devin, and Yoshua Bengio. 2014a. Not all neural embeddings are born equal. *NIPS 2014 Workshop on Learning Semantics*.

Felix Hill, Roi Reichart, and Anna Korhonen. 2014b. SimLex-999: Evaluating Semantic Models with (Genuine) Similarity Estimation. *ArXiv e-prints*, August.

Jay J Jiang and David W Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the 10th Research on Computational Linguistics International Conference*, pages 19–33. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Douwe Kiela and Stephen Clark. 2014. A Systematic Study of Semantic Vector Space Model Parameters. In *Proceedings of EACL 2014, Workshop on Continuous Vector Space Models and their Compositionality (CVSC)*, pages 21–30. Association for Computational Linguistics.

Claudia Leacock and Martin Chodorow. 1998. Combining local context and WordNet similarity for word sense identification. *WordNet: An electronic lexical database*, 49(2):265–283.

Omer Levy and Yoav Goldberg. 2014a. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 302–308, Baltimore, Maryland, June. Association for Computational Linguistics.

Omer Levy and Yoav Goldberg. 2014b. Neural word embedding as implicit matrix factorization. In Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2177–2185. Curran Associates, Inc.

Katja Markert and Malvina Nissim. 2005. Comparing Knowledge Sources for Nominal Anaphora Resolution. *Computational Linguistics*, 31(3):367–402, September.

Diana McCarthy and Roberto Navigli. 2009. The english lexical substitution task. *Language Resources and Evaluation*, 43(2):139–159.

Tomas Mikolov, Greg Corrado, Kai Chen, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *ICLR Workshop*.

Roberto Navigli, Kenneth C. Litkowski, and Orin Hargraves. 2007. Semeval-2007 task 07: Coarse-grained english all-words task. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 30–35, Prague, Czech Republic, June. Association for Computational Linguistics.

Douglas L. Nelson, 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.

Sebastian Padó and Mirella Lapata. 2007. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199.

Martha Palmer, Hoa Trang Dang, and Christiane Fellbaum. 2007. Making fine-grained and coarse-grained sense distinctions, both manually and automatically. *Natural Language Engineering*, 13(02):137–163.

Ted Pedersen. 2010. Information content measures of semantic similarity perform better without sense-tagged text. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 329–332. Association for Computational Linguistics.

Taher Mohammad Pilehvar, David Jurgens, and Roberto Navigli. 2013. Align, disambiguate and walk: A unified approach for measuring semantic similarity. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1341–1351. Association for Computational Linguistics.

Roy Rada, Hafedh Mili, Ellen Bicknell, and Maria Blettner. 1989. Development and application of a metric on semantic nets. *Systems, Man and Cybernetics, IEEE Transactions on*, 19(1):17–30, Jan.

Philip Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th international joint conference on Artificial intelligence-Volume 1*, pages 448–453. Morgan Kaufmann Publishers Inc.

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

Magnus Sahlgren. 2006. The word-space model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces.

Zhibiao Wu and Martha Palmer. 1994. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133–138. Association for Computational Linguistics.

Wen-tau Yih and Vahed Qazvinian. 2012. Measuring word relatedness using heterogeneous vector space models. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 616–620. Association for Computational Linguistics.