The Construction and Evaluation of Word Space Models

Yves Peirsman[*†], **Simon De Deyne**[‡], **Kris Heylen**[*], **Dirk Geeraerts**[*]

[*]QLVL, University of Leuven, Belgium
[†]Research Foundation – Flanders
[‡]Dept. of Psychology, University of Leuven, Belgium
{yves.peirsman,kris.heylen,dirk.geeraerts}@arts.kuleuven.be
simon.dedeyne@psy.kuleuven.be

Abstract

Semantic similarity is a key issue in many computational tasks. This paper goes into the development and evaluation of two common ways of automatically calculating the semantic similarity between two words. On the one hand, such methods may depend on a manually constructed thesaurus like (Euro)WordNet. Their performance is often evaluated on the basis of a very restricted set of human similarity ratings. On the other hand, corpus-based methods rely on the distribution of two words in a corpus to determine their similarity. Their performance is generally quantified through a comparison with the judgements of the first type of approach. This paper introduces a new Gold Standard of more than 5,000 human intra-category similarity judgements. We show that corpus-based methods regularly outperform (Euro)WordNet on this data set, and that the use of the latter as a Gold Standard for the former, is thus often far from ideal.

1. Introduction

One of the great challenges in computational linguistics is the modelling of natural language semantics. For instance, practical applications like Question Answering or Query Expansion often need to know the semantic similarity between two words in order to decide if their results are indeed relevant to the question or query at hand. There are two common ways of addressing this problem. The first makes use of manually compiled resources, like machine-readable dictionaries or thesauri, in order to determine the similarity or dissimilarity between two terms (Rada and Bicknell, 1989; Wu and Palmer, 1994; Jiang and Conrath, 1997; Leacock and Chodorow, 1998). The second relies on large corpora of texts, and calculates the semantic similarity between two words on the basis of their distributional similarity in such a corpus (Landauer and Dumais, 1997; Schütze, 1998; Lin, 1998; Padó and Lapata, 2007).

In this paper, we will focus on the evaluation of these two approaches, which has so far been suffering from a number of weaknesses. We will use each of the two methods to automatically determine the semantic similarity between words from fifteen semantic categories in Dutch, and compare those scores with similarity ratings given by the participants in a psycholinguistic experiment. For the first approach, we turn to Dutch EuroWordNet (Vossen, 1998). While quite a few previous studies have evaluated (English) WordNet against human similarity scores, the data sets were very small and mostly consisted of similarity judgements across categories. We will now extend the Gold Standard from a few dozen to a few thousand similarity scores, and focus on judgements between two words that belong to the same category, like *blackbird* and *robin*.

For the second approach we will construct two so-called Word Space Models: one on the basis of a popular corpus of Dutch newspaper text, the other on the basis of a tailor-made web corpus that was specifically designed to cover the word pairs we are investigating here. Again, we will compare their similarity judgements to the psycholinguistic data. Here we are concerned with two questions. First, we want to find out whether a tailor-made web corpus indeed leads to the expected improvement in performance, as compared to a standard newspaper corpus. Second, we want to contrast the results of these Word Space Models with those of EuroWordNet. We would specifically like to determine whether it is justified to use the latter as a Gold Standard for the evaluation of the former, as is often done in the literature.

2. Dutch EuroWordNet as a source of semantic similarity

2.1. WordNet and Dutch EuroWordNet

It is hardly possible to imagine Natural Language Processing today without WordNet (Fellbaum, 1998). WordNet is a lexical database that brings together groups of synonyms (so-called synsets) in large networks, which show the semantic relationships between them. On the basis of the position of two words in such a network, it is in theory possible to determine how different or similar the concepts are that they represent. WordNet is used in NLP in many different ways: on the one hand, it is often employed as an external knowledge source for systems in Word Sense Disambiguation, Information Retrieval and many other applications; on the other it often also serves as an evaluation framework for algorithms in the field of thesaurus extraction, for example.

Thanks to this success, WordNets for many other languages have been developed, of which Dutch EuroWordNet is one example (Vossen, 1998). Although it is smaller than the English WordNet (it contains almost 34,000 nouns) and not freely available, it is also often relied on as a knowledge source or a Gold Standard for evaluation. For instance, Van de Cruys (2006) uses it to evaluate his clusters of semantically similar words, while Van der Plas and Tiedemann (2006) rely on the database to measure the success of their algorithm that automatically retrieves synonyms from parallel corpora.

While most researchers agree that (Euro)WordNet may not be an ideal Gold Standard for such applications, this fact has triggered hardly any research. Not only is there no thorough evaluation of EuroWordNet's strengths and weaknesses as a Gold Standard for semantic similarity ratings; there has not yet been any serious attempt to propose another Gold Standard that may serve the needs of the NLP community better, at least for a number of tasks. This is precisely the goal of our first set of experiments. By comparing similarity judgements based on the EuroWordNet hierarchy with those given by the participants of a psycholinguistic experiment, we will find that EuroWordNet is not able to give reliable similarity scores for quite a number of categories.

2.2. Measures of semantic similarity in (Euro)WordNet

There are many ways to measure the semantic similarity between two words on the basis of a lexical hierarchy like Dutch EuroWordNet. An excellent overview is given in Budanitsky and Hirst (2006). In particular, there is an important distinction between those measures that use only information contained in the thesaurus on the one hand, and those that combine it with statistics gathered from corpora on the other.[1] These last ones bring together information from two knowledge sources, and could thus be more robust than measures on the basis of the lexical hierarchy only.

In this paper, we will investigate four semantic similarity measures. The first three use information from EuroWordNet only, the fourth adds a web corpus as an extra knowledge source.

Path length The path length measure is probably the most intuitive way of determining the semantic distance between two words w_1 and w_2 on the basis of their position in a hierarchical tree. It simply looks for the shortest path that connects any meaning i of w_1 with any meaning j of w_2 and counts the number of steps in this path:[2]

$$d_{PL}(w_1, w_2) = min(len(w_{1i}, w_{2j})) \quad (1)$$

This metric was applied successfully to MeSH (Medical Subject Headings) by Rada and Bicknell (1989), among other studies.

Inverse path length Before we evaluate the path length measure, we turn it into a measure of semantic similarity instead of distance. We do this by taking its inverse, and refer to this similarity metric as the *inverse path length* measure:

$$s_{IPL}(w_1, w_2) = \frac{1}{d_{PL}(w_1, w_2)} \quad (2)$$

This transformation has the added advantage that it reduces the influence of individual steps in longer paths.

Leacock and Chodorow Of course, a lexical hierarchy like EuroWordNet contains more relevant information than just the number of steps between two words. Leacock and Chodorow (1998) formalize the intuition that the semantic similarity between two words does not only depend on the number of steps between them, but also on the maximum depth of the hierarchy, D. They moreover transform the normalized path lengths by taking their negative natural logarithm:

$$s_{LC}(w_1, w_2) = -log\frac{d_{PL}(w_1, w_2)}{2D} \quad (3)$$

Wu and Palmer Not only the depth of the entire hierarchy plays a role. Intuitively, two words that are two steps from each other should be rated more semantically similar as they lie deeper in the hierarchy. This idea was operationalized by Wu and Palmer (1994), who took into account the lowest hypernym shared by the two words, w_l. Their measure divides twice the depth of w_l by the sum of the path lengths from w_1 and w_2 to the top of the hierarchy:

$$s_{WP}(w_1, w_2) = \frac{2 \times depth(w_l)}{d_{PL}(w_1, w_l) + d_{PL}(w_2, w_l) + 2 \times depth(w_l)} \quad (4)$$

Jiang and Conrath Finally, Jiang and Conrath's (1997) similarity measure combines information from EuroWordNet with word frequency statistics gathered from a large corpus. It specifically takes into account the information content of a concept c, $-\log(p(c))$, where $p(c)$ is the probability of encountering an instance of concept c in a corpus. For any given word, this involves the word itself, together with all its synonyms and hyponyms. We calculated this metric on the basis of our tailor-made web corpus, which we present in the next section. The semantic distance between two words is then defined as the sum of their information contents, minus twice the information content of their lowest shared hypernym, w_l:

$$d_{JC}(w_1, w_2) = IC(w_1) + IC(w_2) - 2 \times IC(w_l) \quad (5)$$

The philosophy behind this idea is that the more information w_1 and/or w_2 add to their lowest shared hypernym, the more dissimilar they are. Jiang and Conrath is thus a measure of semantic distance. We again took its inverse to quantify semantic similarity.

2.3. An evaluation of Dutch EuroWordNet

2.3.1. Relevant work

An obvious way of evaluating the distance or similarity scores obtained from EuroWordNet is by comparing them to human judgements of semantic similarity. In research on English WordNet, this has been done a few times before. Rubenstein and Goodenough's (1965) or Miller and Charles' (1991) psychological experiments often serve as a human Gold Standard. The former study uses 65 word pairs, while the latter focuses on a subset of a mere 30 instances. All pairs were rated by people for semantic similarity on a scale ranging from 'highly synonymous' to 'semantically unrelated'. A replication by Resnik (1995) of

[1] For English, many of these measures are implemented in the package WordNet::Similarity. This is not compatible with Dutch EuroWordNet, however.

[2] As with all other measures, we restrict the paths to the hyponym–hypernym hierarchy in the thesaurus. This excludes loose semantic relationships of the type *doctor — hospital*, where the two words are semantically related, but not similar.

Miller and Charles' experiment gave a correlation of 0.8848 with the original scores. This figure indicates that the human ratings are very reliable. Moreover, it can also be used as an upper bound for the computer ratings, since we should not expect a computer to agree better with human ratings than human participants themselves do.

One investigation that compares the judgements by five WordNet measures with this psychological data is Budanitsky and Hirst (2006). Two of the three best measures in this study return in our experiments. Leacock and Chodorow's measure led to a correlation of .816 with Miller and Charles' results and of .838 with Rubenstein and Goodenough's. Jiang and Conrath's metric, combining two knowledge sources, gave a correlation of .850 and .781, respectively. Overall, these are impressive scores that do not lie far below the upper bound formulated above.

This type of evaluation, however, has its disadvantages. The first problem with the data set is its size — 65 word pairs is an absolute minimum for the evaluation of a specific similarity metric. This weakness has been noted before, for instance by Budanitsky and Hirst (2006, p.32): "While comparison with human judgments is the ideal way to evaluate a measure of similarity or semantic relatedness, in practice the tiny amount of data available [...] is quite inadequate." Moreover, Rubenstein and Goodenough's data set presents a rather easy rating task. Among its least similar words are *cord* and *smile* or *automobile* and *wizard*; among its most similar we find *gem* and *jewel* or *automobile* and *car*. With such a varied set of word pairs, it is not surprising that WordNet is able to produce similarity judgements that correlate well with the human ratings. It would therefore be interesting to see how robust (Euro)WordNet's similarity ratings are when it is asked to rate the similarity between two words in the same category. In the case of vehicles, for instance, we would like it to find that a car and a truck are more semantically similar than a car and an airplane.

As a human Gold Standard for this task, we used a set of intra-category similarity ratings for fifteen categories, which were obtained through a series of psycholinguistic experiments at the University of Leuven, Belgium. Part of this data is presented in Ruts et al. (2004).[3] Ruts and colleagues asked people to rate on a scale from 1 to 20 the semantic similarity between two exemplars from the same category, like *pigeon* and *sparrow* or *guitar* and *piano*. Each category contained between 5 and 30 exemplars, and all possible pairs were rated by at least 14 and at most 17 different participants. Split-half correlations showed the reliability of these ratings to be high or extremely high. People thus agree very well on these intra-category similarity ratings.

This data set is not only far richer than the English sets of a few dozen words; the differences and similarities between two concepts from the same category are also more fine-grained than those between concepts from different categories. We can therefore expect a thesaurus such as EuroWordNet to give results that are clearly inferior to those for inter-category judgements. Yet, the task is not unreasonable, either. Given that (Euro)WordNet is often used to evaluate algorithms that return possible synonyms of a given word, it should be able to discern relatively subtle semantic differences.

2.3.2. Results

On the basis of Dutch EuroWordNet, we thus computed the pairwise exemplar similarities for the following fifteen categories: *musical instruments*, *vehicles*, *tools*, *weapons*, *professions*, *mammals*, *birds*, *reptiles*, *fish*, *insects*, *fruit*, *vegetables*, *kitchen utensils*, *clothing* and *sports*. For each category, this gave between 78 and 465 similarity ratings, with a total of 4,263. If a word did not occur in EuroWordNet, the word pairs of which it was a part were simply ignored. The Pearson correlation of the EuroWordNet results with the human judgements can be found in table 1.[4]

As anticipated, the correlation figures in this table lie much lower than those we referred to above. One reason for this lies in the difficulty of the task, another in the fact that Dutch EuroWordNet is a less fine-grained thesaurus than English WordNet. On the basis of their average correlation, Inverse Path Length (IPL) and Jiang & Conrath (JC) emerge as the most successful metrics for the calculation of semantic similarity or distance. The correlations on the basis of Wu and Palmer's (WP) and Leacock and Chodorow's (LC) measures lie about five to ten per cent lower, on average.

The results of the three similarity measures that use only EuroWordNet as a knowledge source display the same pattern. They all give poor similarity scores for categories like fruit, insects or birds, but much better figures for classes like musical instruments, sports or tools. This clearly indicates that for a number of the investigated classes, EuroWordNet as a thesaurus is simply not able to give similarity ratings that approximate the human scores. Interestingly, the simplest measure, Inverse Path Length (IPL) clearly beats the more advanced ones. This is probably due to the fact that we are dealing with intra-category judgements. The more advanced features that Wu and Palmer's (WP) and Leacock and Chodorow's (LC) metrics add might be better geared towards inter-category judgements, by contrast, for which similarity scores are much more diverse, and on average much lower.

Why could this be the case? Remember that Leacock and Chodorow's measure is merely the negative natural logarithm of a normalized path length score. Often log-transformed values are useful, since the difference between path distances of two and three steps is much larger intuitively than that between path distances of fourteen or fifteen steps, at least in terms of semantic similarity. However, such a log-transformation might lose its usefulness when we stay within the same category, where the number of steps between two exemplars will generally be much lower. Wu and Palmer's measure, next, takes into account the depth of the lowest shared hypernym of the two words. The fact that this does not lead to an improvement for intra-category judgements suggests that the subclassification of

[3] This data is freely available from the Psychonomic Society's Norms, Stimuli, and Data archive, at `http://www.psychonomic.org/archive`.

[4] To be precise, we used the z-scores of the results for each category rather than the original similarity figures to compute the correlations.

Category	n	IPL	WP	LC	JC
Professions	377	.32	.20	.22	.41
Fruit	406	.07	.11	.005	.25
Vegetables	325	.29	.25	.28	.27
Insects	253	.08	-.06	-.02	.24
Kitchen Utensils	465	.46	.25	.36	.37
Clothing	378	.25	.05	.11	.31
Musical Instruments	276	.68	.70	.67	.51
Reptiles	78	.49	.09	.27	.44
Sports	105	.53	.45	.50	.39
Fish	120	.44	.27	.37	.37
Vehicles	351	.49	.55	.48	.44
Birds	300	-.01	-.05	-.03	.19
Weapons	153	.39	.22	.30	.38
Tools	325	.50	.49	.50	.03
Mammals	351	.11	.10	.08	.29
average	284	.34	.24	.28	.33

Table 1: Correlation between EuroWordNet measures and participants' judgements of semantic similarity for 15 categories. IPL: Inverse Path Length; WP: Wu & Palmer; LC: Leacock & Chodorow; JC: Jiang & Conrath

many categories in EuroWordNet does not correspond to the folk model that people have of these classes. This is a hypothesis that clearly deserves further investigation, since it could be of importance for the development of thesauri like EuroWordNet in the future.

As we noted, Inverse Path Length (IPL) scores on average about equally well as Jiang and Conrath (JC), the measure that calculates the information content of the relevant concepts on the basis of a corpus. This does not mean, however, that the two are interchangeable. Table 1 shows that they each have their own strengths. In particular, Inverse Path Length scores at least 5% better on the categories *kitchen utensils*, *musical instruments*, *reptiles*, *sports*, *fish*, *vehicles* and *tools*. Jiang & Conrath, by contrast, scores more than 5% better on the categories *professions*, *fruit*, *insects*, *clothing*, *birds* and *mammals*. Interestingly, *clothing* is the only artifact category for which Jiang & Conrath outperforms Inverse Path Length. Artifact categories seem to be rather well-represented in EuroWordNet on average.

Indeed, the performance of the EuroWordNet measures seems to be influenced by the type of category. The results in Table 1 can be classified in two groups. One contains the natural categories: *fruit*, *vegetables*, *insects*, *reptiles*, *fish*, *birds* and *mammals*. The other categories — *professions*, *kitchen utensils*, *clothing*, *musical instruments*, *sports*, *vehicles*, *weapons* and *tools* — instead consist of artifacts or cultural concepts like professions and sports. All three similarity measures based only on EuroWordNet display a large difference in performance between these groups. Inverse Path Length, for instance, gives an average correlation of .21 for the natural categories, but .45 for the "cultural" group. The JC measure reduces this distinction: it returns an average correlation of .29 for the natural group and of .36 for the cultural group. The cultural group seems thus best represented in Dutch EuroWordNet.

A better look at the EuroWordNet hierarchy indeed confirms that many categories are taken up in the database only in a very coarse way. Musical instruments, for instance, are neatly subcategorized into *string instruments*, *keyboard instruments*, *wind instruments*, etc., many of which still have their own subtypes. This contrasts clearly with biological categories like insects and birds, where most exemplars are listed as direct hyponyms of the category name. For instance, *fazant* (*peacock*), *ekster* (*magpie*), *papegaai* (*parrot*), *roodborstje* (*robin*) and many others all have *vogel* (*bird*) as their immediate hypernym. It is clear that this structure cannot lead to reliable similarity ratings.

2.4. Discussion

In summary, our results clearly show that often, Dutch EuroWordNet is not able to approximate human similarity ratings for two words from the same category. For a total of fifteen categories, the best-performing single measure gave an average correlation of .34 with the human similarity ratings. If we choose for each category the measure with the best result, this gives an average correlation of .40. The implications of this finding are twofold: first, as a computational model of intra-category semantic similarity, Dutch EuroWordNet only performs well for a number of categories — those where its structure is sufficiently rich. Second, computational approaches that are meant to predict or model such human similarity scores are thus better compared against these human judgements directly than against an intermediate Gold Standard like EuroWordNet.

3. Large corpora as a source of semantic similarity

Next, we turn to the second type of computational models of human similarity ratings. In contrast to thesauri like EuroWordNet, Word Space approaches are not designed manually: they are constructed on the basis of a large corpus of data in order to model the distribution of the target words.

3.1. Word Space Models

Word Space Models are inspired by the so-called *distributional hypothesis* (Harris, 1954), which states that words that occur in similar contexts will also be semantically similar. The semantic similarity between two words is thus operationalized as their distributional similarity in a corpus. In practice, the methods work as follows. For each of the target words, they build a so-called context vector, which records how often each contextual feature co-occurs with the target word. The nature of these contextual features may vary. Some models keep track of the documents a target word appears in (e.g., *Latent Semantic Analysis*, Landauer and Dumais (1997)), others look at the context words of the target within a window of a pre-defined size (Schütze, 1998; Levy and Bullinaria, 2001), still others rely on the syntactic relationships in which the target takes part (Lin, 1998; Padó and Lapata, 2007). They all, however, compute the similarity between two words by comparing their context vectors. In general, the more features the two words share, and the more similar their values for those features, the higher the estimate of semantic similarity will be. For a comparison of the results given by different context definitions, see Peirsman, Heylen and Speelman (2007; 2008).

3.2. Experimental setup

In our second series of experiments, we compare two *bag-of-word* approaches, which model a word on the basis of its context words. They are called *bag-of-word* because they treat the context of a target as an unstructured set of words. Our model used a window of two words on either side of the target word as its definition of context. With the exception of semantically empty words in our stop list, it recorded how many times each word in the corpus co-occurred with the target word within this window. On the basis of this frequency, the log-likelihood scores between a target word and all of its context words were computed. This log-likelihood score is a statistical measure that expresses whether the two words co-occur together more or less often than we expect on the basis of their individual frequencies, and is generally more informative than simple frequency counts. It is these scores that we used as the value of each context word in the target vectors. Context words that occurred less than two times together with the target were ignored. The similarity between two target words, finally, was calculated as the cosine of the angle between their two context vectors.

We investigated the performance of this Word Space Model based on two corpora. The first is the Twente Nieuws Corpus (TwNC), a Dutch corpus of about 300 million words, mainly consisting of newspaper articles from between 1999 and 2002. It was compiled at the University of Twente and later parsed by the Alpino parser at the University of Groningen (van Noord, 2006). In recent years, it has become a very popular resource in NLP research. One problem with the Twente Nieuws Corpus is that it focuses on one genre only: newspaper texts. Obviously, newspaper articles will lack frequent reference to many of the concepts we are investigating here, like reptiles and insects. For this reason, we also built a Word Space Model on the basis of a web corpus, which was specifically compiled for the categories in the psycholinguistic experiments we introduced above. The main advantage of such a tailor-made web corpus is that all words in the investigation will occur frequently enough for the model to find context vectors with a large number of non-zero entries. The downside, however, is that data on the internet may be rather noisy and thus less reliable a basis for the construction of Word Space Models. The web corpus has also not been lemmatized and tagged for part of speech.

The web corpus was constructed as follows. For each of the words in the study, 1,000 documents were retrieved from the Internet. Doubles or documents with very similar text to another URL in the set were ignored, as were instances with less than 2 or more than 1,000 sentences. Sentences longer than 3 but shorter than 100 words were tokenized and checked for language: a sentence was allowed into the corpus only if more than 60% of its word forms occurred in the Dutch CELEX word-form dictionary (Baayen et al., 1993). This process resulted in a corpus of more than 750 million words.

3.3. Results

For each of the fifteen categories in the psycholinguistic data, we computed the similarity judgements on the basis of the two Word Space Models. As above, we determined the correlation between the computers' judgements and the human similarity ratings. The results are given in Table 2. Not surprisingly, for thirteen out of fifteen categories, the judgements on the basis of the tailor-made corpus correlate better with the human similarity ratings than those on the basis of the Twente Nieuws Corpus. Sometimes the difference is only marginal, as with birds and vegetables, but usually it is far larger, with regular gains of over .20. On average, the correlation of the newspaper corpus amounts to .31, far below the .43 of the web corpus.

Note also that the Twente Nieuws Corpus does not always give a similarity score for all word pairs in the categories. This was due to one of two reasons: either the word did not occur in the newspaper corpus, or it did occur, but without context words with a frequency of two or more. The percentage of word pairs in each category for which the model returns a similarity rating are given in Table 2 as the *coverage* of the Word Space Model. Obviously, since the web corpus was specifically targeted towards the investigated words, its coverage is always 100%. In conclusion, despite the fact that newspaper corpora generally contain less noisy data than web corpora, quantity trumps quality here.

Let us see if there are interesting differences between categories. Again, there appears to be a discrepancy between the "natural" and "cultural" categories, particularly for the results of the model based on the web corpus. The natural categories give an average correlation of .36 with the human ratings, the cultural categories display an average correlation of .48. For the Twente Nieuws Corpus, the gap between the two cases has decreased: the natural and cultural categories now display mean correlations of .26 and .35 with the human ratings, respectively.

It is rather interesting that the difference between the two groups of categories we found for Dutch EuroWordNet also crops up for the Word Space Model on the basis of our

Category	n	web corpus		newspaper corpus	
		r	coverage	r	coverage
Professions	435	.42	100%	.46	93%
Fruit	435	.39	100%	.40	69%
Vegetables	378	.34	100%	.30	93%
Insects	325	.22	100%	.03	85%
Kitchen Utensils	496	.54	100%	.31	94%
Clothing	406	.39	100%	.29	93%
Musical Instruments	351	.56	100%	.32	100%
Reptiles	171	.31	100%	.08	89%
Sports	406	.47	100%	.20	80%
Fish	231	.46	100%	.33	82%
Vehicles	435	.54	100%	.52	87%
Birds	435	.44	100%	.43	100%
Weapons	171	.60	100%	.54	89%
Tools	378	.35	100%	.15	67%
Mammals	435	.37	100%	.27	100%
average	366	.43	100%	.31	88%

Table 2: Correlation between the human similarity judgements and those based on a web corpus and a newspaper corpus.

web corpus. Obviously, the reasons must be different here. Possibly, the prototypical nature of natural categories is less well reflected in the linguistic contexts of its exemplar names, but this is an issue that certainly deserves further investigation.

Finally, compare the results in Table 2 with those in Table 1. Interestingly, the performance of the two investigated corpora with respect to the human similarity judgements often surpasses that of Dutch EuroWordNet. The judgements on the basis of the Twente Nieuws Corpus show a higher correlation with the human judgements for five out of fifteen categories. For the tailor-made web corpus, this is true for nine of out fifteen classes. This substantiates our claim from the previous section: for computational models of human similarity ratings, direct comparison with such ratings should be preferred over the use of EuroWordNet as a Gold Standard.

4. Conclusions

In recent years, there has been considerable interest in computational models of the semantic similarity between two words. Both machine-readable thesauri like (Euro)WordNet and large text corpora often serve as resources for such approaches. Moreover, the former are often used as a Gold Standard for the evaluation of the latter. As this paper has shown, the quality of that type of evaluation may often be questionable. We have presented the first comparison of Dutch EuroWordNet intra-category similarity judgements with human ratings from a large psycholinguistic experiment. We showed that the performance of EuroWordNet differs considerably from category to category, depending on the amount of detail that the EuroWordNet tree contains for that category. In particular, we noted a difference in performance between "natural" categories like insects and birds, for which there was hardly any correlation, and the more "cultural" categories like musical instruments and vehicles, for which EuroWordNet judgements were much more reliable.

Next, we investigated the performance of two Word Space Models. While parameters like context size or similarity metric were kept constant, they were constructed from very different data. One was based on a standard corpus of Dutch newspaper text of about 300 million words, the other on a corpus of possibly noisy data from the web, totalling 750 million tokens. This web corpus was specifically tailored for an investigation of the categories in this paper, so that each target word occurred frequently enough. Not surprisingly, in the rule the similarity judgements of this tailor-made web corpus correlated better with the human similarity ratings than those of the newspaper corpus. Moreover, for nine out of fifteen categories they also outperformed EuroWordNet — the Gold Standard against which these Word Space Models are usually evaluated. This shows that corpus-based models are often better able to estimate semantic similarity as rated by humans than those based on a manually constructed thesaurus. For such computational models of human similarity ratings, we thus advise the use of psycholinguistic data like those in Ruts et al.'s (2004) experiments as a Gold Standard.

5. References

Harald Baayen, Richard Piepenbrock, and Hedderik van Rijn. 1993. *The CELEX Lexical Database (Release 1) [CD-ROM]*. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.

Alexander Budanitsky and Graeme Hirst. 2006. Evaluating WordNet-based measures of lexical semantic relatedness. *Computational Linguistics*, 35(1):13–47.

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

Zellig Harris. 1954. Distributional structure. *Word*, 10(23):146–162.

Jay J. Jiang and David W. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy.

In *Proceedings of the International Conference on Research in Computational Linguistics*, pages 19–33, Taiwan.

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

Claudia Leacock and Martin Chodorow. 1998. Combining local context and WordNet similarity for word sense identification. In Christiane Fellbaum, editor, *WordNet. An Electronic Lexical Database*, pages 265–283. Cambridge, MA: The MIT Press.

Joseph P. Levy and John A. Bullinaria. 2001. Learning lexical properties from word usage patterns: Which context words should be used. In R.F. French and J.P. Sougne, editors, *Connectionist Models of Learning, Development and Evolution: Proceedings of the Sixth Neural Computation and Psychology Workshop*, pages 273–282. London: Springer.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of COLING-ACL98*, pages 768–774, Montreal, Canada.

George A. Miller and Walter G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

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

Yves Peirsman, Kris Heylen, and Dirk Speelman. 2007. Finding semantically related words in Dutch. Co-occurrences versus syntactic contexts. In *Proceedings of the CoSMO workshop, held in conjunction with CONTEXT-07*, pages 9–16, Roskilde, Denmark.

Yves Peirsman, Kris Heylen, and Dirk Speelman. 2008. Putting things in order. First and second order context models for the calculation of semantic similarity. In *Proceedings of the 9th Journées internationales d'Analyse statistique des Données Textuelles (JADT 2008)*, pages 907–916, Lyon, France.

Roy Rada and Ellen Bicknell. 1989. Ranking documents with a thesaurus. *Journal of the American Society for Information Science*, 40(5):304–310.

Philip Resnik. 1995. Using information content to evaluate semantic similarity. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 448–453, Montreal, Canada.

Herbert Rubenstein and John B. Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–393.

Wim Ruts, Simon De Deyne, Eef Ameel, Wolf Vanpaemel, Timothy Verbeemen, and Gert Storms. 2004. Dutch norm data for 13 semantic categories and 338 exemplars. *Behavior Research Methods, Instruments, & Computers*, 36(3):506–515.

Hinrich Schütze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–124.

Tim Van de Cruys. 2006. The application of Singular Value Decomposition to Dutch noun-adjective matrices. In Piet Mertens, Cédrick Fairon, Anne Dister, and Patrick Watrin, editors, *Verbum Ex Machina. Actes de la 13e Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, pages 767–772, Leuven, Belgium.

Lonneke Van der Plas and Jörg Tiedemann. 2006. Finding synonyms using automatic word alignment and measures of distributional similarity. In *Proceedings of ACL/COLING-2006*, pages 866–873.

Gertjan van Noord. 2006. At last parsing is now operational. In Piet Mertens, Cédrick Fairon, Anne Dister, and Patrick Watrin, editors, *Verbum Ex Machina. Actes de la 13e Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, pages 20–42.

Piek Vossen, editor. 1998. *EuroWordNet: a multilingual database with lexical semantic networks for European Languages*. Kluwer, Dordrecht.

Zhibiao Wu and Martha Palmer. 1994. Verb semantics and lexical selection. In *Proceedings of ACL-94*, pages 133–138, Las Cruces, NM.