Putting things in order. First and second order context models for the calculation of semantic similarity.

Yves Peirsman[1,2], Kris Heylen[1] and Dirk Speelman[1]

[1]Quantitative Lexicology and Variational Linguistics (QLVL)

University of Leuven – Belgium

[2]Research Foundation – Flanders

Abstract

Recent years have seen an increase of interest in vector-based approaches to lexical semantics. These are inspired by the distributional hypothesis, which states that semantic similarity can be modelled by distributional similarity in a corpus. However, while there are a variety of ways to construct the vector models, it is an open question as to what type of semantic information these contain. In this paper, we investigate three popular ways of building such models. The first two are bag-of-word methods that rely on the (weighted) frequencies of the words that co-occur with the target word. One looks at the context words directly, the other uses second-order contexts. The third technique takes syntactic relations into account. Despite the popularity of these models, very little is known about the semantic information that they capture. For a start, our experiments show a low overlap between the results of the second-order model and the other two. Second-order information thus leads to a vector space clearly different from spaces that rely on first-order information. Secondly, the overall performance of the second-order model is clearly inferior to that of its first-order competitors, compared to a gold standard like EuroWordNet. Thirdly, an investigation of the types of semantic relations that the models find shows that syntactic information seems to favour strict semantic relations, while first-order bag-of-word information results in the discovery of more loose relations.

Keywords: distributional hypothesis, word space models, semantic similarity, thesaurus extraction.

1. Introduction

One of the great challenges in computational linguistics is the modelling of natural language semantics. Since the advent of statistical methods in NLP, this problem has often been approached with word space models, which treat the meaning of a word as a function of the contexts it occurs in. They are motivated by the assumption that words with a similar meaning appear in similar contexts – a claim often referred to as the *distributional hypothesis* (Harris, 1954). In this framework, the semantic similarity between two words corresponds to the similarity between their respective context vectors.

Word space models thus fully depend on the notion of *context*. Yet, this concept can be – and has been – defined in many ways. One possibility is to just look at the context words that a certain word co-occurs with. Another is to take syntactic relations into account. In the future, it is not inconceivable that extralinguistic context will play a role as well. Naturally, we can expect these distinct types of context to capture different kinds of semantic information.

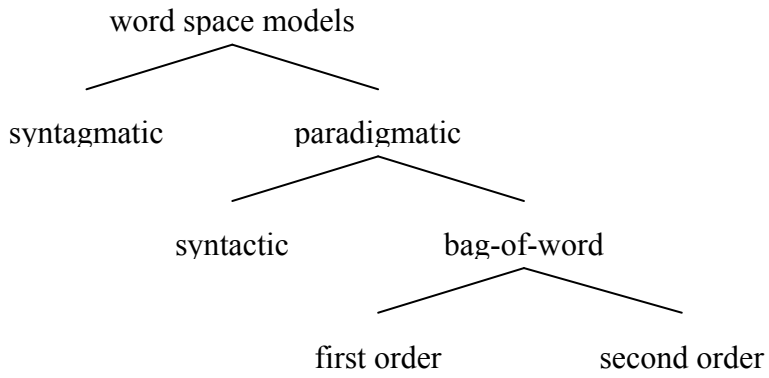

Figure 1. Overview of different types of word space models.

However, so far little is known about the influence of the context definition on the semantic information present in the vector spaces. While most researchers choose one specific word space model and apply it to their task, "comparisons between the (…) models have been few and far between in the literature" (Padó and Lapata 2007: 166). We consider this a major issue: without any knowledge of the linguistic characteristics of the models, it is impossible to know which approach is best suited for a particular task, and why. In this paper, we will take a first step towards such a comparison.

In the next section, we will introduce three of the most popular word space models and discuss some comparative literature on the topic. Section three will then describe our research questions and experimental setup. In section four, we present and discuss the experimental results. Section five wraps up with a conclusion and an outlook for future research.

2. Types of word space models

The definitions of linguistic context that word space models often use, display a hierarchical structure (see Figure 1)[1]. At the top, we have the main distinction between syntagmatic and paradigmatic contexts – two terms we borrowed from Sahlgren (2005). A syntagmatic context model records what paragraphs, articles, documents, or other context region a word appears in. A paradigmatic model, in contrast, defines a word in terms of features that occur in its context (words, syntactic relations, etc.) Thus, in a syntagmatic context model, two words will receive similar vectors when they co-occur in the same contexts. In a paradigmatic model, two words will be closely related when they share many context features.

The paradigmatic class of word space models can be further subdivided into syntactic and bag-of-word models. In a syntactic model, the only possible context features are the words that are syntactically related to the target word, possibly together with their relation. In a bag-of-word model, all words within a predefined context window are taken into account.

Finally, we can distinguish between first-order and second-order bag-of-word models. While the first looks at the context words directly, the second records the context words of these

[1] We should add that the hierarchy in Figure 1 is by no means complete. It is possible, for instance, to use a second-order syntactic model that looks at the context words of the syntactically related words. Our classification merely contains the most popular types of vector models.

direct context words – so-called *second-order context words*. Such second-order context models are claimed to deal better with data sparseness and with first-order context words that have similar meanings.

All these models appear frequently in the literature. Latent Semantic Analysis, for instance, probably the most popular word space model, relies on syntagmatic contexts for the calculation of semantic similarity (Landauer and Dumais, 1997). Schütze (1998) uses second-order bag-of-word models for unsupervised Word Sense Disambiguation, while Lin (1998) looks at syntactic context for the automatic discovery of semantically similar words. Clearly, there is no agreement on the merits of one context model compared to the others.

2.1. Comparative studies of word space models

As we said above, there is hardly any research that compares the different word space models. One notable exception is Sahlgren (2005). In his PhD dissertation, Sahlgren studied the first two types of word space models we introduced: the syntagmatic model and the first-order bag-of-word model (which he calls paradigmatic). His hypothesis was that syntagmatic word spaces contain more syntagmatic information, while first-order bag-of-word spaces contain more paradigmatic information.

These notions of syntagmatic and paradigmatic information are derived from Structuralism. In De Saussure's theory, two words are syntagmatically related when they can be combined in context, and paradigmatically related when the one can be substituted by the other. Of course, there is no sharp distinction between the two. While a relationship like synonymy, for instance, is typically paradigmatic, two synonyms can also co-occur in the same context.

Through a series of tasks, Sahlgren was able to confirm his hypothesis. First-order bag-of-word contexts outperformed syntagmatic ones for the synonym and part-of-speech tests, mainly paradigmatic in nature. Syntagmatic contexts, in contrast, gave better results for the more syntagmatic association test. With the thesaurus and antonym tests, the results varied with the evaluation procedure. Indeed, these tasks have to be situated somewhere in the middle of the continuum between the two relations.

In a previous article (Peirsman et al., 2007), we took this type of investigation one step further. We compared six word space models – syntactic and first-order bag-of-word spaces, with different context sizes and dimensionality reduction methods – both with respect to their overall performance and the semantic relations that they favoured (synonymy, hyponymy, hypernymy and co-hyponymy). We found that the syntactic model clearly outperformed the other ones, with an outspoken preference for synonymy relations. It is this line of research that we will continue in this paper.

3. Experimental setup

Our experiments compare three types of word spaces: the first-order and second-order bag-of-word models and the syntactic model. They are meant to answer three questions. First, to what degree do the results of the three models differ? Second, which of the three models is most successful at finding words that are semantically similar to the target word? Third, do the different models have a bias for different types of semantic relations?

Our data is the 300 million word Twente Nieuws Corpus of Dutch newspaper text, which was parsed with the Alpino parser at the University of Groningen (Van Noord, 2006). We

extracted from the lemmatized corpus the 10,000 most frequent nouns and retrieved for each one of them the 100 most similar nouns among the remaining 9,999 possibilities.

The parameters of the models were set as follows:

(1) *Dimensionality:* For all three approaches, we used only the 4,000 most frequent features as dimensions. With the second-order bag-of-word model, this means a second-order context word was only taken into account if the word itself and the first-order context word to which it belonged were among the 4,000 most frequent context words in the corpus.

(2) *Context window*: For the bag-of-word approaches, we used a context window of three words on either side of the target word. Preliminary experiments had shown this window size to give better overall results than wider context windows of ten or twenty words.

(3) *Weighting scheme:* Our context vectors do not contain the frequencies of the features, but rather the point-wise mutual information (PMI) scores between each feature and the target. This measure quantifies if the target word and the feature occur together more or less often than expected on the basis of their individual frequencies.

(4) *Frequency cut-off:* With the bag-of-word models (both second-order and first-order) we counted only those context words that occurred at least five times together with the target word, in order to remove noise. For the syntactic model such a cut-off was not used, as it resulted in data sparseness.

(5) *Similarity metric:* The cosine was used to measure the similarity between two vectors. This measure is more or less standard in the literature (see for instance Schütze, 1998 and Padó and Lapata, 2007).

(6) *Stop list:* For the bag-of-word models, semantically empty words were not considered candidates for the 4,000 most frequent dimensions. For the syntactic models, a stop list was not used, since it was found to hurt performance.

The syntactic model, finally, took into account eight syntactic relations, responsible for eight different types of features:

- subject of verb *v,*
- direct object of verb *v,*
- prepositional complement of verb *v* introduced by preposition *p*,
- head of an adverbial PP of verb *v* introduced by preposition *p*,
- modified by adjective *a*,
- postmodified by a PP with head *n*, introduced by preposition *p*,
- modified by an apposition with head *n*, or
- coordinated with head *n*.

Each specific instantiation of the variables *v*, *p*, *a*, or *n* led to a new context feature.

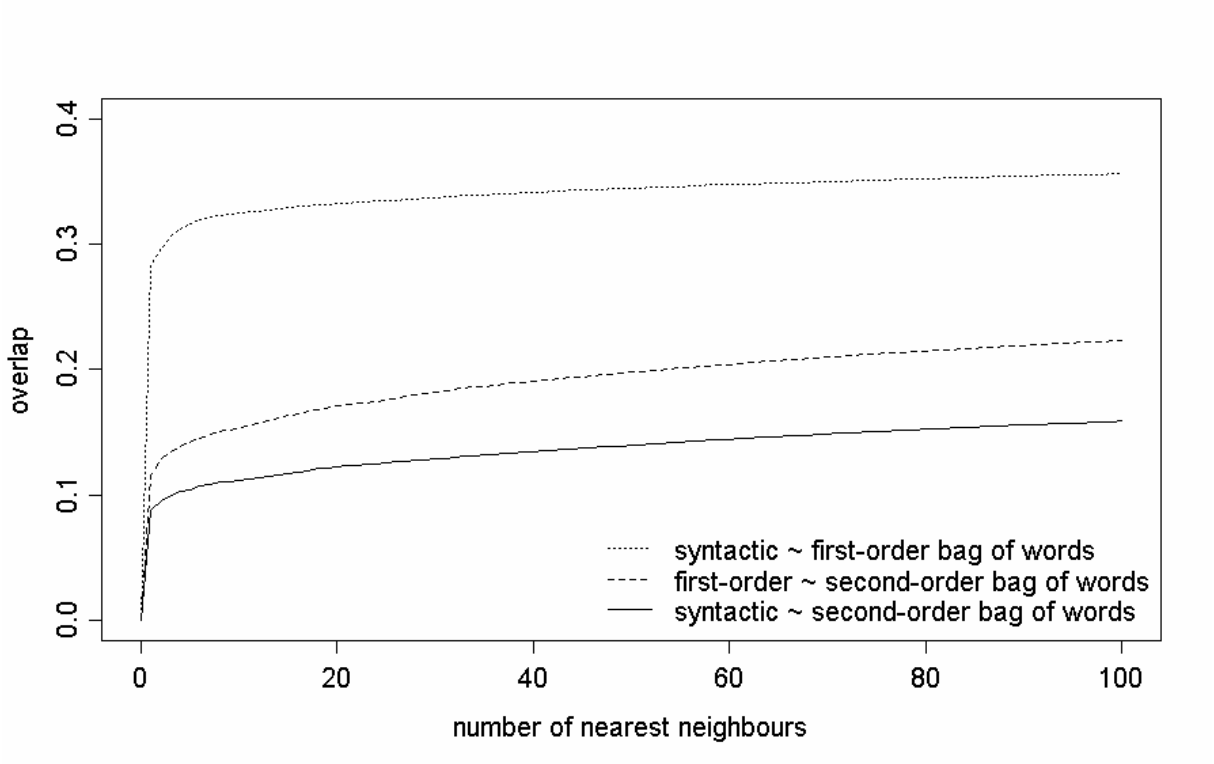

Figure 2. Overlap between the results of the different models.

4. Results and discussion

4.1. Overlap between the results

First we wanted to measure the difference between the word spaces in terms of the words they retrieve. This can be done by the overlap metric, taken from Sahlgren (2005). This metric reflects how similar two word spaces are, simply by calculating the overlap between their results: for each of our 10,000 words, we took the *n* nearest neighbours and then calculated how many neighbours the word spaces shared. This overlap is generally very low. Sahlgren, for instance, found a maximum of around 10% between the syntagmatic and first-order paradigmatic models, when looking at the ten most similar nouns.

Figure 2 maps the overlap between each pair of models, with the number of nearest neighbours ranging from 0 to 100. Clearly, the first-order bag-of-word model and the syntactic model are most similar. Their overlap lies at 28% with just one neighbour, and increases to 32.5% and 35.5% with ten and 100 neighbours respectively. The second-order bag-of-word model differs quite considerably from either of these two: at ten neighbours, the overlap with the first-order model is only 15%, and with the syntactic model a mere 11%.

A likely explanation for this contrast lies in the fact that both the first-order bag-of-word model and the syntactic model rely on first-order information. Admittedly, the syntactic model does not look at exactly the same neighbours as the first-order bag-of-word model, and contains extra information in the form of syntactic relations, but both models rely on information that is directly present in the context of the target word. This sets them apart from second-order bag-of-word models, which treat context words in terms of *their* context words. Since this is a different type of information, the lower overlap should come as no surprise.

In addition, the size of the context window is likely to be another important factor. As our first-order bag-of-word model uses a small context window of only three words on either side of the target, many of the context words it takes into account will also be syntactically related to the target. We therefore expect that the results of first-order bag-of-word models with wider context windows will overlap less with the results of the syntactic model. As long as we work with small contexts, however, the main difference does not lie between the syntactic and bag-of-word models, but between first-order and second-order approaches.

4.2. Overall performance

Of course, the overlap between the models does not tell us anything about the quality of their results. For each of the target words in our test set, we therefore looked at its semantic similarity with its nearest neighbour. This semantic similarity can be measured on the basis of an electronic thesaurus like Dutch EuroWordNet. Much like its English sibling, Dutch EuroWordNet is an electronic resource that contains around 34,000 sets of synonyms in a conceptual hierarchy. On the basis of the position of two words, it is possible to quantify their semantic similarity. One of the popular measures for this task is the Wu & Palmer similarity score (Wu and Palmer, 1994).

For all three models, we calculated the average Wu & Palmer similarity score between each target and its nearest neighbour, if present in EuroWordNet. If a word occurred at different places in the hierarchy, only the highest Wu & Palmer figure was taken into account[2]. This similarity score was then averaged over all targets. Table 1 gives the results. In line with our previous research (Peirsman et al., 2007), the syntactic model reached the highest performance, with an average Wu & Palmer score of 0.62. Both bag-of-word models score considerably lower, with the first-order approach at an average score of 0.52 and the second-order method at 0.31. Again, the first-order models give the most similar results.

	average Wu & Palmer score
syntactic	0.62
first-order bag of words	0.52
second-order bag of words	0.31

Table 1. Average Wu and Palmer score for the single nearest neighbour.

Whereas the strong result of the syntactic model is no surprise, the low figure of the second-order model is rather striking. After all, second-order bag-of-word models are fairly popular in the literature, not only for reasons of data sparseness, but also because they are said to deal better with (nearly) synonymous context words (which receive similar word vectors). Our results now suggest that when the corpus provides enough data, first-order models are the better alternative by far.

[2] In the case of a polysemous word, we are only interested in the meaning most closely related to the target word.

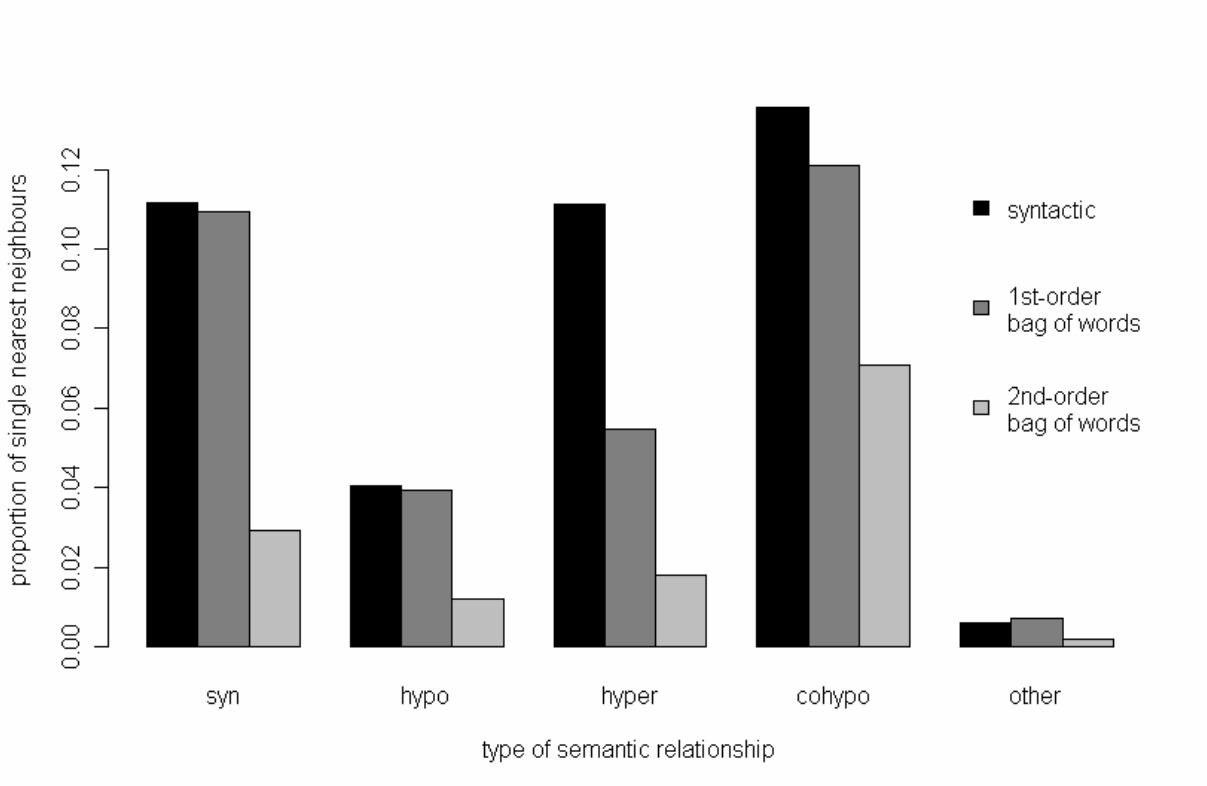

Figure 3. The distribution of semantic relationships for the three word spaces.

We should note, however, that this first evaluation metric has two weaknesses. First of all, the Wu & Palmer similarity score does not tell us what type of semantic similarity the models capture. Do they mostly find synonyms or hyponyms, or other types of semantic relations? Second, it only quantifies semantic *similarity*, or closeness in the EuroWordNet hierarchy. This relationship holds between pairs of words like *car* and *vehicle*, or *car* and *truck*. Yet, two words need not be hierarchically close to each other in order to be semantically related. Pairs like *car* and *wheel* or *doctor* and *hospital* display a clear semantic relation, even though their members are not semantically similar. We need another evaluation method to capture this type of *looser* semantic relatedness.

4.3. Semantic relations

In order to address the problem mentioned above, we investigated the distribution of semantic relations for the three models by comparing the single nearest neighbour for each target with that target's immediate vicinity in EuroWordNet[3]. This vicinity contains five types of relations: *synonyms*, *hyponyms*, *hypernyms*, *co-hyponyms*, and *other*, for all other relations. We defined hyponyms as the immediate daughters of the target, hypernyms as immediate mothers, and co-hyponyms as the sisters in the hierarchy. *Other* relations are loosely related words that are just one step away from the target, but in a relationship different from those above (*meronymy*, *antonymy, etc.)*. Although such links are very informative from a linguistic point of view, they are very poorly represented in Dutch EuroWordNet.

[3] Nearest neighbours not present in EuroWordNet were simply ignored.

4.3.1. Distribution of semantic relations

The results of these experiments are given in Figure 3.[4] The bars indicate what percentage of the single nearest neighbours present in EuroWordNet belonged to one of the predefined semantic relationships, for each of the three models we investigated. They confirm the results in the previous section, which singled out the syntactic model as the best indicator of semantic similarity, followed by the first-order bag-of-word model and then the second-order bag-of-word model. This is true for the four major relationships, although the difference between the first two models for synonyms and hyponyms is extremely small and, according to a chi-square test, also insignificant.

Obviously, this distribution of semantic relations is influenced by the frequency of these relations in EuroWordNet. The main reason the models find more co-hyponyms than any other relation is that a word generally has more co-hyponyms than synonyms, hyponyms or hypernyms. If we bear this in mind, the distributions in Figure 3 have three interesting characteristics. First of all, the syntactic model finds an unexpectedly high number of hypernyms, compared to the other two models. The general bias of the syntactic model for the strictest semantic relationships (synonymy, hyponymy and hypernymy) can thus fully be attributed to this single relationship. Next, the second-order bag-of-word model retrieves a relatively high number of co-hyponyms. Co-hyponyms are already one step further away in the hierarchy, which may suggest that relatively speaking, the second-order model is biased towards such slightly looser relations. Finally, the syntactic and first-order bag-of-word models return an almost equal proportion of *other* semantic relationships, with the second-order model clearly behind. The frequency of these nearest neighbours, however, is too low to allow for any far-reaching conclusions. We therefore applied another evaluation metric to quantify a possible preference this type of relatedness.

4.3.2. Syntagmatic relatedness

Loosely or topically related words often occur together in context. In a text where the word *hospital* occurs, we are more likely to also find *doctor* than, say, *forest*. Such a syntagmatic relationship between two words can be discovered by their point-wise mutual information score (PMI). PMI is defined as the logarithm of the joint probability of two words divided by the product of their respective probabilities. Since the chance of seeing two topically related words together will normally be higher than we expect on the basis of their frequencies, their PMI score will also be higher than that between unrelated words.

Thus, for the single nearest neighbour of each target, we calculated its point-wise mutual information with the target. We used a context window of ten words on either side of the target word. These results can be seen in the first column of Table 2. If a word never occurs in the context of the target, its PMI score is not defined. Yet, this is probably also an indication that the two words are not, or very vaguely, related. While this problem can be tackled with smoothing techniques, for now we chose to simply give these results a PMI of zero. The resulting figures are given in the second column of Table 2. All differences are statistically significant, on the basis of a paired Wilcoxon test.

[4] If there was more than one relationship between two words, only the strongest was counted, with synonymy > hyponymy > hypernymy > cohyponymy > other relations.

	average PMI value	average PMI value with penalty
syntactic	3.06	2.78
first-order bag of words	3.86	2.99
second-order bag of words	2.84	1.77

Table 2. Average PMI value for the single nearest neighbour.

Interestingly, it is the first-order bag-of-word model that scores best, with an average PMI value of 3.86 without penalty. Thus, while the syntactic model finds more results that are semantically similar to the target word, the first-order bag-of-word model retrieves words that are more syntagmatically related. This observation also holds when we penalize the model for nearest neighbours that never co-occur with the target word. The first-order bag-of-word model still scores best, but the syntactic model now follows closely behind. The second-order model in particular suffers from this penalty. While it may still be biased towards these syntagmatic relationships, the quality of its nearest neighbours appears far from stable.

5. Conclusions and future work

Our detailed investigation of three popular semantic word space models has brought to light some important differences. In all our experiments, the syntactic model and the first-order bag-of-word model were clearly set apart from the second-order bag-of-word model. This appeared mainly due to the quality of the results, with the second-order model finding fewer words that are semantically similar or related to the targets. Yet, there were also major differences between the first-order models. The syntactic model showed a particular preference for hypernyms as nearest neighbours, while the first-order bag-of-word model scored better when it comes to syntagmatic relationships.

Obviously, there are more ways of building word spaces than the ones we have explored so far. In the near future, we would like to expand our research to different context sizes and other types of syntactic information. In this way we would like to obtain a comprehensive overview of the semantic characteristics of distributional models.

References

Harris Z. (1954). Distributional structure. *Word*, 10 (23): 146-162.

Landauer T. K. and Dumais S. T. (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.

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

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

Peirsman Y., Heylen K. and Speelman D. (2007). Finding semantically related words in Dutch. Co-occurrences versus syntactic contexts. In *Proceedings of the CoSMO workshop*, Roskilde, Denmark, pages 9-16.

Sahlgren M. (2005). *The Word-Space Model. Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces.* Unpublished PhD dissertation, University of Stockholm.

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

Van Noord G. (2006). At last parsing is operational now. In Mertens P., Fairon C., Dister A., Watrin P. (eds)*, Verbum Ex Machina. Actes de la 13e conference sur le traitement automatique des langues naturelles*, Leuven, Belgium, pages 20-42.

Wu Z. and Palmer M. (1994). Verb semantics and lexical selection. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, Las Cruces, NM, pages 133-138