Word Space Models of Lexical Variation

Yves Peirsman
Research Foundation – Flanders &
QLVL, University of Leuven
Leuven, Belgium
yves.peirsman@arts.kuleuven.be

Dirk Speelman
QLVL, University of Leuven
Leuven, Belgium
dirk.speelman@arts.kuleuven.be

Abstract

In the recognition of words that are typical of a specific language variety, the classic keyword approach performs rather poorly. We show how this keyword analysis can be complemented with a word space model constructed on the basis of two corpora: one representative of the language variety under investigation, and a reference corpus. This combined approach is able to recognize the markers of a language variety as words that not only have a significantly higher frequency as compared to the reference corpus, but also a different distribution. The application of word space models moreover makes it possible to automatically discover the lexical alternative to a specific marker in the reference corpus.

1 Introduction

Different varieties of the same language often come with their lexical peculiarities. Some words may be restricted to a specific register, while other ones may have different meanings in different regions. In corpus linguistics, the most straightforward way of finding such words that are typical of one language variety is to compile a corpus of that variety and compare it to a reference corpus of another variety. The most obvious comparison takes on the form of a keyword analysis, which looks for the words that are significantly more frequent in the one corpus as compared to the other (Dunning, 1993; Scott, 1997; Rayson et al., 2004). For the purposes of a language-variational study, this classic keyword approach often does not suffice, however. As Kilgarriff has argued, keyword statistics are far too sensitive to high frequencies or topical differences to be used in the study of vocabulary differences (Kilgarriff, 2001). We therefore put forward an approach that combines keyword statistics with distributional models of lexical semantics, or word space models (Sahlgren, 2006; Bullinaria and Levy, 2007; Padó and Lapata, 2007; Peirsman, 2008). In this way, we not only check whether two words have significantly different frequencies in the two relevant language varieties, but also to what degree their distribution varies between the corpora.

In this paper, we will focus on the lexical differences between two regional varieties of Dutch. Dutch is interesting because it is the official language of two neighbouring countries, Belgium and the Netherlands. Between these two countries, there exists a considerable amount of lexical variation (Speelman et al., 2006). There are words much more frequently used in one of the two varieties as well as terms that have a different meaning in the two regions. We will call such words *markers* of a specific *lect* — a general term for regiolects, dialects, or other language varieties that are specific to a certain region, genre, etc. By constructing a word space model on the basis of two corpora instead of one, we will show how the distributional approach to lexical semantics can aid the recognition of such lectal variation.

In the next section, we will point out the weaknesses of the classic keyword approach, and show how word space models can provide a solution. In section 3, we will discuss how our approach recognizes markers of a given lect. In section 4, we will demonstrate how it can automatically find the alternatives in the other language variety. Section 5 wraps up with conclusions and an outlook for future research.

2 Bilectal Word Spaces

Intuitively, the most obvious way of looking for words that mark a particular language variety, is to take a corpus that represents this variety, and calculate its keywords with respect to a reference

Proceedings of the EACL 2009 Workshop on GEMS: GEometical Models of Natural Language Semantics, pages 9–16,
Athens, Greece, 31 March 2009.

χ^2		log-likelihood	
keyword	χ^2	keyword	log-likelihood
frank/noun (‘franc’)	262492.0	frank/noun (‘franc’)	335587.3
meer/adj (‘more’)	149505.0	meer/adj (‘more’)	153811.6
foto/noun (‘photograph’)	84286.7	Vlaams/adj (‘Flemish’)	93723.2
Vlaams/adj (‘Flemish’)	83663.0	foto/noun (‘photograph’)	87235.1
veel/adj (‘much’/’many’)	73655.5	vrijdag/noun (‘Friday’)	77865.5
Belgisch/adj (‘Belgian’)	62280.2	veel/adj (‘much’/‘many’)	74167.1
vrijdag/noun (‘Friday’)	59135.9	Belgisch/adj (‘Belgian’)	64786.0
toekomst/noun (‘future’)	42440.5	toekomst/noun (‘future’)	55879.1
dossier/noun (‘file’)	34623.3	dossier/noun (‘file’)	45570.0
Antwerps/adj (‘Antwerp’)	33659.1	ziekenhuis/noun (’hospital’)	44093.3

Table 1: Top 10 keywords for the Belgian newspaper corpus, as compared to the Twente Nieuws Corpus.

corpus (Dunning, 1993; Scott, 1997; Rayson et al., 2004). This keyword approach has two important weaknesses, however. First, it has been shown that statistically significant differences in the relative frequencies of a word may arise from high absolute frequencies rather than real lexical variation (Kilgarriff, 2001). Second, in the explicit comparison of two language varieties, the keyword approach offers no way of telling what word in the reference corpus, if any, serves as the alternative to an identified marker. Word space models offer a solution to both of these problems.

We will present this solution on the basis of two corpora of Dutch. The first is the Twente Nieuws Corpus (TwNC), a 300 million word corpus of Netherlandic Dutch newspaper articles from between 1999 and 2002. The second is a corpus of Belgian Dutch we compiled ourselves, with the goal of making it as comparable to the Twente Nieuws Corpus as possible. With newspaper articles from six major Belgian newspapers from the years 1999 to 2005, it totals over 1 billion word tokens. Here we will work with a subset of this corpus of around 200 million word tokens.

2.1 Keywords

As our starting point, we calculated the keywords of the Belgian corpus with respect to the Netherlandic corpus, both on the basis of a chi-square test (with Yates’ continuity correction) (Scott, 1997) and the log-likelihood ratio (Dunning, 1993). We considered only words with a total frequency of at least 200 that moreover occurred at least five times in each of the five newspapers that make up the Belgian corpus. This last restriction was imposed in order to exclude idiosyncratic language use in any of those newspapers. The top ten resulting keywords, listed in Table 1, show an overlap of 90% between the tests. The words fall into a number of distinct groups. *Frank*, *Vlaams*, *Belgisch* and *Antwerps* (this last word appears only in the χ^2 top ten) indeed typically occur in Belgian Dutch, simply because they are so tightly connected with Belgian culture. *Dossier* may reflect a Belgian preference for this French loanword. Why the words *meer*, *veel*, *vrijdag*, *toekomst* and *ziekenhuis* (only in the log-likelihood top ten) are in the lists, however, is harder to explain. There does not appear to be a linguistically significant difference in use between the two language varieties, neither in frequency nor in sense. The presence of *foto*, finally, may reflect certain publishing habits of Belgian newspapers, but again, there is no obvious difference in use between Belgium and the Netherlands. In sum, these Belgian keywords illustrate the weakness of this approach in the modelling of lexical differences between two language varieties. This problem was already noted by Kilgarriff (2001), who argues that “[t]he LOB-Brown differences cannot in general be interpreted as British-American differences”. One of the reasons is that “for very common words, high χ^2 values are associated with the sheer quantity of evidence and are not necessarily associated with a pre-theoretical notion of distinctiveness”. One way to solve this issue is presented by Speelman et al. (2008). In their so-called *stable lexical markers* analysis, the word frequencies in one corpus are compared to those in several reference corpora. The keyness of a word then corresponds to the number of times it appears in the resulting keyword lists of the first corpus. This repetitive test

helps filter out spurious keywords whose statistical significance does not reflect a linguistically significant difference in frequency. Here we explore an alternative solution, which scores candidate markers on the basis of their contextual distribution in the two corpora, in a so-called bilectal word space.

2.2 Bilectal Word Spaces

Word space models (Sahlgren, 2006; Bullinaria and Levy, 2007; Padó and Lapata, 2007; Peirsman, 2008) capture the semantic similarity between two words on the basis of their distribution in a corpus. In these models, two words are similar when they often occur with the same context words, or when they tend to appear in the same syntactic relationships. For our purposes, we need to build a word space on the basis of two corpora, more or less in the vein of Rapp's (1999) method for the identification of translation equivalents. The main difference is that we use two corpora of the same language, each of which should be representative of one of the language varieties under investigation. All other variables should be kept as constant as possible, so that we can attribute differences in word use between the two corpora to lexical differences between the two lects. Next, we select the words that occur in both corpora (or a subset of the n most frequent words to reduce dimensionality) as the dimensions of the word space model. For each target word, we then build two context vectors, one for each corpus. These context vectors contain information about the distribution of the target word. We finally calculate the similarity between two context vectors as the cosine of the angle between them.

One crucial parameter in the construction of word space models is their definition of *distribution*. Some models consider the syntactic relationships in which a target word takes part (Padó and Lapata, 2007), while other approaches look at the collocation strength between a target and all of the words that occur within n words to its left and right (Bullinaria and Levy, 2007). With these last *word-based* approaches, it has been shown that small context sizes in particular lead to good models of the semantic similarity between two words (Bullinaria and Levy, 2007; Peirsman, 2008). So far, we have therefore performed experiments with context sizes of one, two and three words to the left and right of the target. These all gave very similar results. Experiments with other context sizes and with syntactic features will be carried out in the near future. In this paper, we report on the results of a word-based model with context size three.

In order to identify the markers of Belgian Dutch, we start from the keyword lists above. For each of the keywords, we get their context vector from the Belgian corpus, and find the 100 most similar context vectors from the Netherlandic corpus. The words that correspond to these context vectors are called the 'nearest neighbours' to the keyword. In the construction of our word space model, we selected from both corpora the 4,000 most frequent words, and used the cross-section of 2,538 words as our set of dimensions or context features. The model then calculated the point-wise mutual information between the target and each of the 2,538 context words that occurred at least twice in its context. All words in the Netherlandic Dutch corpus with a frequency of at least 200, plus the target itself, were considered possible nearest neighbours to the target.

Generally, where there are no major differences in the use of a keyword between the two lects, it will have itself as its nearest neighbour. If this is not the case, this may identify the keyword as a marker of Belgian Dutch. For example, six words from the lists above have themselves as their nearest neighbour: *meer*, *foto*, *veel*, *vrijdag*, *toekomst* and *ziekenhuis*. These are indeed the keywords that made little sense from a language-variational perspective. *Dossier* is its own second nearest neighbour, which indicates that there is slightly less of a match between its Belgian and Netherlandic use. Finally, the words linked to Belgian culture — *frank*, *Vlaams*, *Belgisch* and *Antwerps* — are much lower in their own lists of nearest neighbours, or totally absent, which correctly identifies them as markers of Belgian Dutch. In short, the keyword analysis ensures that the word occurs much more frequently in Belgian Dutch than in Netherlandic Dutch; the word space approach checks if it also has a different distribution in the two corpora.

For markers of Belgian Dutch, we can interpret the nearest neighbour suggested by the system as the other variety's alternative to that marker. For instance, *dossier* has *rapport* as its nearest neighbour, a synonym which indeed has a high keyword value for our Netherlandic Dutch corpus. Similarly, the culture-related words have their Dutch

equivalents as their distributionally most similar words: *frank* has *gulden* ('guilder'), *Vlaams* and *Belgisch* both have *Nederlands* ('Dutch'), and *Antwerps* has *Amsterdams* ('Amsterdam (adj.)'). This makes intuitive sense if we take meaning to be a relative concept, where for instance a concept like 'currency of this country' is instantiated by the franc in Belgium and the guilder in Holland — at least in the pre-Euro period. These findings suggest that our combined method can be applied more generally in order to automatically discover lexical differences between the two language varieties.

3 Recognizing lectal differences

First we want to investigate whether a bilectal word space model can indeed contribute to the correct identification of markers of Belgian Dutch on a larger scale. We therefore had both types of approaches — the simple keyword approach and the combined method — suggest a top 2,000 of possible markers on the basis of our two corpora. The combined approach uses the same word space method we described above, with 2,538 dimensions and a context size of three. Basing itself on the lists of nearest neighbours, it then reorders the list of keywords, so as to arrive at a ranking that reflects lectal variation better than the original one. To this goal, each keyword receives a new score, which is the multiplication of two individual numbers. The first number is its rank in the original keyword list. At this point we considered only the 5,000 highest scoring keywords. The second is based on a list that ranks the words according to their difference in distribution between the two corpora. Words that do not occur in their own list of 100 nearest neighbours appear at the top of the list (rank 1), followed by words that are their own 100th nearest neighbour (rank 2), and so on to the words that have themselves as nearest neighbour (rank 101). In the future we plan to consider different numbers of neighbours in order to punish words with very different distributions more or less heavily. At this stage, however, restricting the method to 100 nearest neighbours gives fine results. These two ranks are then multiplied to give a combined score, on the basis of which a final list of candidates for lectal variation is computed. The lower this combined score (reflecting either high keyword values, very different distributions in the two corpora, or both), the higher the likelihood that the word is a marker of Belgian Dutch. This approach thus ensures that words that have very different distributions in the two corpora are promoted with respect to the original keyword list, while words with very similar distributions are downgraded.

candidate marker	evaluation
frank/noun ('franc')	culture
Vlaams/adj ('Flemish')	culture
match/noun ('match')	literature
info/noun ('info')	
rijkswacht/noun ('state police')	RBBN
weekend/noun ('weekend')	
schepen/noun ('alderman')	RBBN
fr./noun ('franc')	culture
provinciaal/adj ('provincial')	RBBN
job/noun ('job')	RBBN

Table 2: Top ten candidate markers suggested by the combined method on the basis of the log-likelihood ratio.

As our Gold Standard we used the *Reference List of Belgian Dutch* (*Referentiebestand Belgisch Nederlands*, RBBN), a list of almost 4,000 words and expressions that are typical of Belgian Dutch (Martin, 2005). These are classified into a number of groups — culturally-related terms (e.g., names of political parties), Belgian markers that are not lexicalized in Netherlandic Dutch, markers that are lexicalized in Netherlandic Dutch, etc. We used a subset of 717 one-word nouns, verbs and adjectives that appear at least 200 times in our Belgian corpus to evaluate our approach. Even if we informally explore the first ten candidate markers, the advantages of combining the log-likelihood ratio with the word space model already become clear (see table 2). Four of these candidates are in the RBBN gold standard. Similarly, *frank*, *Vlaams* and *fr.* are culturally related to Belgium, while *match* has been identified as a typically Belgian word in previous corpus-linguistic research (Geeraerts et al., 1999). *Info* and *weekend* are not present in the external sources we consulted, but nevertheless show an interesting distribution with respect to their respective synonyms. In the Belgian corpus, *info* occurs more often than the longer and more formal *information* (32,009 vs 30,171), whereas in the Dutch corpus the latter is used about 25 times as frequently as the former (1,681 vs 41,429). Similarly, the Belgian corpus

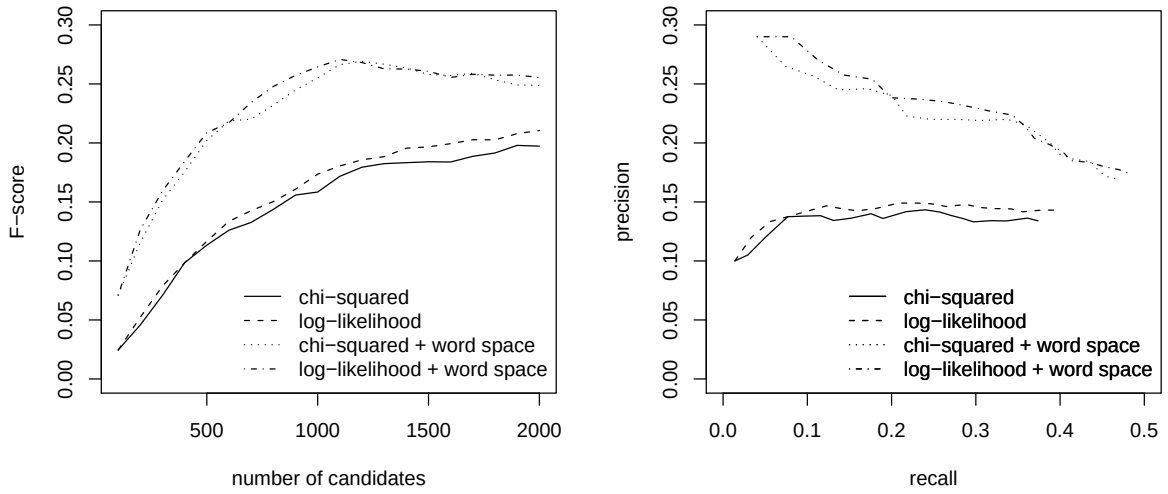

Figure 1: Precision and recall figures of the keyword methods and the combined approaches.

contains far more instances of *weekend* than of its synonym *weekeinde* (35,406 vs 6,390), while the Dutch corpus shows the reverse pattern (6,974 vs 28,234). These words are thus far better candidate markers than the original keywords *meer*, *foto*, *veel*, *vrijdag*, *toekomst* or *ziekenhuis*, which have disappeared from the top ten.

Let us now evaluate the methods more broadly, on the basis of the top 2,000 keywords they suggest. The left plot in Figure 1 shows their F-scores in function of the number of suggested markers; the right graph plots precision in function of recall. The two keyword approaches score rather similarly, with the log-likelihood ratio achieving slightly better results than the chi-square test. This superiority of the log-likelihood approach was already noted by Rayson et al. (2004). Both combined methods give a very clear advantage over the simple keyword statistics, again with the best results for the log-likelihood ratio. For example, ten of the first 100 candidates suggested by both keyword approaches are present in our Gold Standard, giving a precision of 10% and a recall of 1.4% (F-score 2.4%). Adding our word space model makes this figure rise to 29 correct markers, resulting in a precision of 29% and a recall of 4% (F-score 7.1%). This large advantage in performance is maintained further down the list. At 1000 candidates, the keyword approaches have a recall of around 20% (chi-square 19%, log-likelihood 21%) and a precision of around 14% (chi-square 14%, log-likelihood 15%). At the same point, the combined approaches have reached a recall of over 30% (chi-square 31%, log-likelihood 32%) with a precision of around 22% (chi-square 22%, log-likelihood 23%). Expressed differently, the best keyword approach needs around 500 candidates to recover 10% of the gold standard, 1000 to recover 20% and 2000 to recover 40%. This linear increase is outperformed by the best combined approach, which needs only 300, 600 and 1500 candidate markers to reach the same recall figures. This corresponds to relative gains of 40%, 40% and 25%. As these results indicate, the performance gain starts to diminish after 1000 candidates. Future experiments will help determine if this issue can be resolved with different parameter settings.

Despite these large gains in performance, the combined method still has problems with a number of Belgian markers. A manual analysis of these cases shows that they often have several senses, only one of which is typical of Belgian Dutch. The Reference List for instance contains *fout* ('mistake') and *mossel* ('mussel') as Belgian markers, with their specialized meanings 'foul (in sports)' and 'weakling'. Not only do these words have very low keyword values for the Belgian corpus; they also have very similar distributions in the two corpora, and are their own first and second neighbour, respectively. Sometimes a failure to recognize a particular marker is more due

	top 100		top 500	
class	n	%	n	%
in Gold Standard	29	29%	127	25.4%
in Van Dale	11	22%	47	9.4%
related	2	2%	23	4.6%
cultural terms	25	25%	60	12%
total	67	67%	257	51.4%

Table 3: Manual analysis of the top 500 words suggested by the combined approach.

to the results of one individual method. This is for instance the case with the correct Belgian marker *home* ('(old people's) home'). Although the word space model does not find this word in its own list of nearest Netherlandic neighbours, it remains low on the marker list due to its fairly small log-likelihood ratio. Conversely, *punt*, *graad* and *klaar* are rather high on the keyword list of the Belgian corpus, but are downgraded, as they have themselves as their nearest neighbour. This is again because their status as a marker only applies to one infrequent meaning ('school mark', 'two-year cycle of primary education' and 'clear') instead of the dominant meanings ('final stop, point (e.g., in sports)', 'degree' and 'ready'), which are shared between the two regional varieties. However, this last disadvantage applies to all markers that are much more frequently used in Belgium but still sometimes occur in the Netherlandic corpus with a similar distribution.

Finally, because our Gold Standard is not an exhaustive list of Belgian Dutch markers, the results in Figure 1 are an underestimate of real performance. We therefore manually went through the top 500 markers suggested by the best combined approach and classified them into three new groups. The results of this analysis are presented in Table 3. First, we consulted the *Van Dale Groot Woordenboek der Nederlandse taal* (Den Boon and Geeraerts, 2005), the major dictionary of Dutch, which contains about 3,000 words marked with the label "Belgian Dutch". 11% of the first 100 and 9.4% of the first 500 candidates that were initially judged incorrect carry this label or have a definition that explicitly refers to Belgium. Second, we counted the words that are morphologically related to words in the Gold Standard or to Belgian words found in Van Dale. These are for instance compound nouns one of whose parts is present in the Gold Standard, which means that

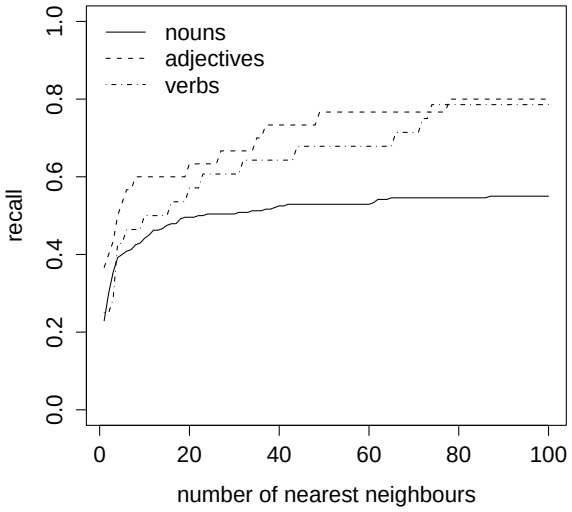

Figure 2: Percentage of markers of Belgian Dutch whose Netherlandic alternative is present among their n nearest neighbours.

they are correct markers of Belgian Dutch as well. They represent 2% of the top 100 and 4.6% of the top 500. Third, we counted the words that are inherently linked to Belgian culture, mostly in the form of place names. This group corresponds to 25% of the first 100 and 12% of the first 500 candidate markers. This suggests that the true precision of our method at 100 and 500 candidates is thus at least 67% and 51.4%, respectively.

4 Finding alternatives

The *Reference List of Belgian Dutch* not only lists Belgian Dutch words and expressions, but also gives their Netherlandic Dutch alternative, if one exists. Our word space model offers us a promising way of determining this alternative automatically, by looking at the nearest Netherlandic neighbours to a Belgian marker. As our Gold Standard, we selected from the Reference List those words with a frequency of at least 200 in the Belgian corpus whose Dutch alternative also had a frequency of at least 200 in the Dutch corpus. This resulted in a test set of 315 words: 240 nouns, 45 verbs and 30 adjectives. For each of these words, we used our word space model to find the 100 nearest Netherlandic neighbours, again with context size three but now with as dimensions all words shared between the two corpora, in order to improve performance. We then determined if the

Dutch alternative was indeed in the list of nearest neighbours to the target. We started by looking at the single nearest neighbour only, and then step by step extended the list to include the 100 nearest neighbours. If a word had itself among its nearest neighbours, this neighbour was discarded and replaced by the next one down the list. The results are shown in Figure 2. 11 out of 30 adjectives (36.7%), 10 out of 45 verbs (22.2%) and 56 out of 240 nouns (23.3%) had their Dutch alternative as their nearest neighbour. At ten nearest neighbours, these figures had risen to 60.0%, 48.9% and 44.6%. These encouraging results underpin the usefulness of word space models in language-variational research.

A manual analysis of Belgian markers for which the approach does not find the Netherlandic alternative again reveals that a large majority of these errors occur when polysemous words have only one, infrequent meaning that is typical of Belgian Dutch. For example, the dominant sense of the word *tenor* is obviously the 'male singer' meaning. In Belgium, however, this term can also refer to a leading figure, for instance in a political party or a sports discipline. Since this metaphorical sense is far less frequent than the literal one, the context vector fails to pick it up, and almost all nearest Netherlandic neighbours are related to opera or music. One way to solve this problem would be to abandon word space models that build only one context vector per word. Instead, we could cluster all individual contexts of a word, with the aim of identifying context clusters that correspond to the several senses of that word (Schütze, 1998). This is outside the scope of the current paper, however.

5 Conclusions and future research

We have presented an application of word space models to language-variational research. To our knowledge, the construction of word space models on the basis of two corpora of the same language instead of one is new to both variational linguistics and Natural Language Processing. It complements the classic keyword approach in that it helps recognize those keywords that, in addition to their different relative frequencies in two language varieties, also have a substantially different distribution. An application of this method to Belgian Dutch showed that the keywords that pass this test indeed much more often represent markers of the language variety under investigation. Moreover, often the word space model also succeeded in identifying the Netherlandic Dutch alternative to the Belgian marker.

As the development of this approach is still in its early stages, we have committed ourselves more to its general presentation than to the precise parameter settings. In the near future, we therefore aim to investigate more fully the possible variation that the method allows. First, we will focus on the implementation of the word space model, by studying word-based models with other context sizes as well as syntax-based approaches. Second, we want to examine other ways in which the word-based model and the classic keyword approach can be combined, apart from the multiplication of ranks that we have proposed here. While this large freedom in parameter settings could be seen as a weakness of the proposed method, the fact that we obtained similar results for all settings we have tried out so far, adds to our confidence that word space models present a sensible complementation of the classic keyword approaches, irrespective of the precise parameter settings.

In addition to those modelling issues, there are a number of other extensions we would like to explore. First, the Gold Standard we have used so far is rather limited in scope. We therefore plan to incorporate more sources on language variation to test the robustness of our approach. Finally, as we have observed a number of times, the method in its present form is not sensitive to possibly infrequent meanings of a polysemous word. This may be solved by the application of a clustering approach that is able to cluster a word's contexts into several sense clusters (Schütze, 1998). Still, the promising results in this paper encourage us to believe that the current approach has a future as a new method in language-variational research and as a tool for lexicography.

References

John A. Bullinaria and Joseph P. Levy. 2007. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behaviour Research Methods*, 39:510–526.

Ton Den Boon and Dirk Geeraerts. 2005. *Van Dale Groot Woordenboek van de Nederlandse taal (14e ed.)*. Van Dale Lexicografie, Utrecht/Antwerp.

Ted Dunning. 1993. Accurate methods for the statis-

tics of surprise and coincidence. *Computational Linguistics*, 19:61–74.

Dirk Geeraerts, Stefan Grondelaers, and Dirk Speelman. 1999. *Convergentie en Divergentie in de Nederlandse Woordenschat*. Meertens Instituut, Amsterdam.

Adam Kilgarriff. 2001. Comparing corpora. *International Journal of Corpus Linguistics*, 6(1):1–37.

Willy Martin. 2005. Het Belgisch-Nederlands anders bekeken: het Referentiebestand Belgisch-Nederlands (RBBN). Technical report, Vrije Universiteit Amsterdam, Amsterdam, Holland.

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

Yves Peirsman. 2008. Word space models of semantic similarity and relatedness. In *Proceedings of the 13th ESSLLI Student Session*, pages 143–152.

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 519–526, College Park, Maryland.

Paul Rayson, Damon Berridge, and Brian Francis. 2004. Extending the cochran rule for the comparison of word frequencies between corpora. In *Proceedings of the 7ièmes Journées Internationales d'Analyse Statistique des Données Textuelles (JADT 2004)*, pages 926–936, Louvain-la-Neuve, Belgium.

Magnus Sahlgren. 2006. *The Word-Space Model. Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations Between Words in High-dimensional Vector Spaces*. Ph.D. thesis, Stockholm University, Stockholm, Sweden.

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

Mike Scott. 1997. PC analysis of key words – and key key words. *System*, 25(2):233–245.

Dirk Speelman, Stefan Grondelaers, and Dirk Geeraerts. 2006. A profile-based calculation of region and register variation: The synchronic and diachronic status of the two main national varieties of Dutch. In Andrew Wilson, Dawn Archer, and Paul Rayson, editors, *Corpus Linguistics around the World*, pages 195–202. Rodopi, Amsterdam.

Dirk Speelman, Stefan Grondelaers, and Dirk Geeraerts. 2008. Variation in the choice of adjectives in the two main national varieties of Dutch. In Gitte Kristiansen and René Dirven, editors, *Cognitive Sociolinguistics. Language Variation, Cultural Models, Social Systems*, pages 205–233. Mouton de Gruyter, Berlin.