Natural Language Engineering **16** (4): 469–491. © Cambridge University Press 2010
doi:10.1017/S1351324910000161

The automatic identification of lexical variation between language varieties

YVES PEIRSMAN[1,2], DIRK GEERAERTS[2]
and DIRK SPEELMAN[2]

[1]*Research Foundation – Flanders (FWO), Egmontstraat 5, 1000 Brussels, Belgium*
email: yves.peirsman@arts.kuleuven.be
[2]*Quantitative Lexicology and Variational Linguistics (QLVL), University of Leuven, Blijde-Inkomststraat 21 P.O. Box 3308, 3000 Leuven, Belgium*
email: dirk.geeraerts@arts.kuleuven.be, dirk.speelman@arts.kuleuven.be

(*Received 15 July 2009; revised 25 February 2010; accepted 14 May 2010*)

Abstract

Languages are not uniform. Speakers of different language varieties use certain words differently – more or less frequently, or with different meanings. We argue that distributional semantics is the ideal framework for the investigation of such lexical variation. We address two research questions and present our analysis of the lexical variation between Belgian Dutch and Netherlandic Dutch. The first question involves a classic application of distributional models: the automatic retrieval of synonyms. We use corpora of two different language varieties to identify the Netherlandic Dutch synonyms for a set of typically Belgian words. Second, we address the problem of automatically identifying words that are typical of a given lect, either because of their high frequency or because of their divergent meaning. Overall, we show that distributional models are able to identify more *lectal markers* than traditional keyword methods. Distributional models also have a bias towards a different type of variation. In summary, our results demonstrate how distributional semantics can help research in variational linguistics, with possible future applications in lexicography or terminology extraction.

1 Introduction

One of the most significant changes in recent linguistics is the move towards corpus-based research (Tummers, Heylen and Geeraerts 2005). As a growing number of researchers become convinced that only real usage data can give a reliable impression of the complexity of language, the use of corpora is winning ground in many branches of theoretical linguistics. This trend, however, is not as strong in the field of lexical semantics. The main problem is that word meaning is a much more elusive concept than linguistic phenomena like syntax or morphology, which can readily be observed in corpora. In particular, researchers in lexical semantics are looking for automatic ways of establishing the semantic equivalence between two words (Geeraerts 2010). At the same time, in computational linguistics, distributional models of lexical semantics are one of the success stories of recent years (Turney and

https://doi.org/10.1017/S1351324910000161 Published online by Cambridge University Press

Pantel 2010). These models are able to identify semantically similar words merely on the basis of the similarity between their contexts in a large corpus. Although they seem to provide an intuitive and conceptually simple answer to the current problem of corpus-based lexical semantics, it appears that the gulf between the two disciplines thus far has been too large for a theoretical-linguistic application of distributional semantics. We want to take the first step towards closing this gap.

In this paper, we focus on variational linguistics, the research area that is concerned with the linguistic differences between varieties of the same language. Our study deals with Dutch, and its national varieties in Belgium and the Netherlands. We show how distributional models allow us to solve variational-linguistic questions that have never been addressed, like the automatic recognition of synonymy across two language varieties, and to formulate an alternative solution to well-known research questions, such as the identification of words that are typical of one language variety.

Belgium and the Netherlands provide an ideal situation for the application of this research programme. Like the United States and the United Kingdom, these two countries share a common language. Still, the two national varieties of Dutch display considerable variation. The most conspicuous difference between the two language varieties lies in their pronunciation, but there are a fair number of lexical contrasts as well. Whereas Belgians typically use the French loanword *dessert* to refer to a dessert, the Dutch tend to use *toetje*. And while Belgians generally use *nonkel* for 'uncle', the Dutch prefer *oom*. It is this specific type of language variation that we would like to model.

Before we start, we want to make a few terminological notes. First, we use the word *lect* to generalise over dialects, regiolects, registers and other language varieties that are restricted to a specific area, discourse situation or similar language-external variable. Second, we identify the two national varieties of Dutch as *Belgian Dutch* and *Netherlandic Dutch* (and not *Dutch Dutch*), in order to avoid confusion about the meaning of the word *Dutch*. Third, we use the term *marker* to refer to a word that is typical of a specific lect, either because it has a much higher frequency in that lect as compared to other ones or because it has a different meaning. To rephrase the goal of our article in these terms, we present a method for the automatic recognition of markers of a specific lect, and the automatic identification of their synonym in another lect.

The rest of our article is structured as follows. In Section 2, we introduce the framework of distributional semantics and discuss relevant literature, both from a computational-linguistic and a corpus-linguistic point of view. In Section 3, we show how distributional models can be used to identify the Netherlandic Dutch synonyms to lectal markers of Belgian Dutch. Section 4 suggests an alternative definition of the notion of *keyword* in corpus linguistics for the identification of such lectal markers. Traditionally, keywords are defined as words that occur significantly more often in a particular corpus than in a reference corpus. We demonstrate that a new distributional approach to keywords can identify lectal markers more successfully than the traditional, frequency-based approach. Finally, Section 5 wraps up with conclusions and an outlook for future research.

https://doi.org/10.1017/S1351324910000161 Published online by Cambridge University Press

2 Background

In this study, we bring together several strands of research. Our general framework is the distributional approach to lexical semantics, which has become omnipresent in computational linguistics in recent years. Section 2.1 gives a short introduction to this family of computational models. Our specific application was inspired by two different research areas, which are presented in Section 2.2. The first is the corpus-linguistic investigation of the lexicon (Section 2.2.1). The second is the development of cross-linguistic distributional models (Section 2.2.2).

2.1 Distributional models of lexical semantics

Distributional models of lexical semantics, which are also known as word space models or semantic spaces, are a popular way of automatically looking for semantically similar and related words. They were inspired by the distributional hypothesis, which claims that words that occur in similar contexts also have a similar meaning (Wittgenstein 1953; Harris 1954; Firth 1957). These models use large corpora to collect information about the contexts in which words occur. Different types of models define context differently. Word-based models count the frequency of context words in a context window of *n* words around the target word, syntax-based models rely on the frequencies of the syntactic relations in which a target word takes part, and document-based models remember in what paragraphs or documents a target word occurs. This information is stored in a so-called context vector. In order to find the words that are semantically related or similar to a target word, we merely calculate the similarity between a target word's context vector and the context vectors of all other words in the corpus, and extract the words that are most similar distributionally.

A practical example will make this procedure a bit clearer. Suppose we want to look for words that are semantically related to the English word *plane*, from a word-based perspective. This can be done by taking a corpus and counting the frequencies of all words that occur in the context of *n* words around *plane*. This will likely result in high frequency counts for context words like *fly* and *pilot*. Next, we use these frequencies to calculate the collocational strength between the target word and each of its context words, for example in the form of point-wise mutual information. The resulting values are stored in the context vector of *plane*. We also collect the context vectors for all other words in the corpus, and calculate the similarity between these context vectors and the context vector of the target word. The words with the highest similarity are likely to be semantically related to our target. We call these words the nearest neighbours to the target. The likely nearest neighbours in our example will be *airplane* and *helicopter*.

The most important parameter in the development of distributional models is the definition of context. The syntax-based models (Padó and Lapata 2007) are generally thought to be the best models of semantic similarity and will extract taxonomically similar words like *plane–airplane* or *plane–helicopter* (Peirsman, Heylen and Speelman 2007; Van de Cruys 2008; Van der Plas 2008). The document-based

https://doi.org/10.1017/S1351324910000161 Published online by Cambridge University Press

models, by contrast, tend to identify semantically *associated* word pairs like *plane–fly* or *plane–pilot* (Sahlgren 2006; Peirsman and Geeraerts 2009). The word-based models sit somewhere in between these two extremes. Their bias towards semantic similarity or association depends on the size of the context window: the word-based models with small context sizes are better models of similarity, whereas large context sizes are better models of association. Because we are concerned with the identification of semantic similarity, mainly in the form of synonymy, we will focus on the syntax-based models and the word-based models with small context sizes.

The success of distributional models of lexical semantics is apparent from the large variety of tasks to which they are applied, in both computational linguistics and cognitive science. In computational linguistics, these approaches are used for sense discrimination and disambiguation (Schütze 1998), automatic thesaurus extraction (Lin 1998; Curran 2004), automated essay grading (Kakkonen *et al.* 2005), textual entailment (Jijkoun and de Rijke 2005; Zhitomirsky-Geffet and Dagan 2009) and query expansion (Bai *et al.* 2005), to name just a few examples. In cognitive science, distributional approaches have proved successful in topics as diverse as the modeling of language acquisition (Landauer and Dumais 1997; Baroni, Lenci and Onnis 2007), priming effects (Lowe and McDonald 2000), semantic dyslexia (Buchanan, Burgess and Lund 1996), sentence and discourse comprehension (Foltz 1996; Landauer and Dumais 1997; Burgess, Livesay and Lund 1998) and metaphor processing (Kintsch 2000). Mitchell *et al.* (2008) used distributional representations of word meaning to predict the neural activation observed in the brain when a participant reads a concrete noun.

2.2 Related literature

Our study aims to fill the gap between the corpus-linguistic investigation of lexical variation and the computational-linguistic modelling of semantic similarity. Our experiments can be linked up with earlier investigations in both of these fields.

2.2.1 Corpus linguistics

Two lines of research in corpus linguistics are relevant to our research. The first is the quantitative investigation of lexical variation between two lects. The second is the application of distributional methods to theoretical-linguistic questions.

Variational linguistics. When language users want to refer to a given concept, they often have a number of words at their disposal. One factor that influences their particular choice of words is the language community to which they belong. Speakers of Netherlandic Dutch will make other choices than speakers of Belgian Dutch, just like speakers of British English will choose different words than speakers of American English. Despite the conspicuousness of this variation, Geeraerts (2010) notes that 'lexical variation has only rarely been studied as a sociolinguistic variable'. He attributes this to two factors: the difficulty of collecting enough usage data in

https://doi.org/10.1017/S1351324910000161 Published online by Cambridge University Press

certain areas, and the problem of determining whether two words are semantically equivalent and indeed express the same concept or not. Nevertheless, a number of important studies have already paved the way for a comprehensive corpus-based study of the variation between two language varieties. Geeraerts, Grondelaers and Speelman (1999) and Speelman, Grondelaers and Geeraerts (2003) investigated the variation between Belgian and Netherlandic Dutch in the lexical fields of football and clothing terms, with Soares da Silva (2010) conducting a similar research programme in Portuguese. Glynn (2007) focuses on the lectal, semantic and constructional factors that influence the choice between the English verbs *bother*, *hassle* and *annoy*, while Wulff, Stefanowitsch and Gries (2007) and Szmrecsanyi (2010) incorporate lectal variation in their study of constructional near-synonyms. Although all these previous studies deal with lexical or constructional variation between two lects, their research objectives differ quite significantly from ours. Instead of investigating the precise factors that determine lexical variation, we target the automatic identification of such differences.

Distributional methods. With the move towards usage-based linguistics, some recent studies in corpus linguistics have focused on word usage from a distributional point of view. These approaches are usually less technical than distributional models in computational linguistics. First, some studies have focused on a particular polysemous word and investigated the syntactic or lexical features that correlate with the occurrence of its several meanings. This approach is found in Gries' (2006) paper on the English verb *to run*. Second, it is also possible to study a set of words and the differences and similarities between them. This perspective is taken by Divjak and Gries (2006), who cluster Russian verbs with the meaning 'to try' according to their behaviour in a corpus. Third, there have also been some first attempts at the application of full-fledged distributional models in corpus linguistics. Sagi, Kaufmann and Clark (2009) use distributional models to investigate meaning change, and Boussidan, Sagi and Ploux (2009) study the relatedness in meaning between all English words that share the phonaestheme *gl—*. Bertels (2008) applies distributional models to the study of polysemy in French technical corpora. Despite these first efforts, distributional models are still unknown to most researchers in corpus linguistics. This gap between the theoretical and more technical disciplines is regrettable, given the many advantages of the distributional approach to the study of word meaning.

In our application of distributional semantics to lexical variation, we address the two problems identified by Geeraerts (2010). First, inspired by a large corpus of newspaper articles from the Netherlands that is freely available, we compiled a similar corpus of Belgian Dutch newspaper material. These two comparable corpora allow us to study lexical variation between the standard varieties of Dutch in the two countries. Second, we investigate whether distributional semantics can help establish the semantic equivalence between words across two lects. To this end, we apply the popular distributional approach to a situation with two corpora rather than one.

https://doi.org/10.1017/S1351324910000161 Published online by Cambridge University Press

2.2.2 Cross-linguistic distributional models

The construction of distributional models from two corpora is not completely new. In computational linguistics, this idea has been exploited in the automatic development of bilingual lexicons. On the basis of an initial set of translation pairs, researchers have been able to cross-linguistically compare the distribution of words. As very strong collocations appear to be rather constant in comparable corpora of different languages (Rapp 1995), it is possible to calculate the distributional similarity between two words in different languages. In this way, we can automatically identify likely translations of a target word. This general approach has been put into practice in a number of studies of the cross-lingual semantic similarity between concepts (Mohammad *et al.* 2007) or words (Rapp 1995; Fung and McKeown 1997; Fung and Yee 1998; Rapp 1999; Gamallo Otero 2008; Gamallo Otero and Pichel Campos 2008).

In a similar vein, we will use distributional models to compare the contextual distribution of words across corpora. The main difference with these earlier studies lies in the fact that we work with two corpora of the same language, as we are interested in what we could call intra-language (but cross-lectal) translation pairs. This actually makes our method easier than those mentioned above: as we work with two corpora of the same language, we do not need a set of initial translation pairs to compare the distributions. After all, most of the words in Dutch are shared between the two national lects and have the same meaning in Belgium and the Netherlands. This means that the cross-lectal computation of distributional similarity does not require any external knowledge source.

3 Experiment 1: Cross-lectal synonymy retrieval

One of the traditional applications of distributional semantics is synonymy identification. Generally, the distributional models are used to search for semantically similar words in the same corpus or in a comparable corpus of a different language. We will now apply the same methodology to two corpora of different language varieties to identify possible synonyms across lects.

3.1 Experimental setup

Data. We use two large corpora of Dutch. The Twente Nieuws Corpus (TwNC) represents the Netherlandic Dutch lect (Ordelman *et al.* 2007). It contains newspaper articles published in five Dutch newspapers between 1999 and 2002, totalling more than 260 million word tokens. To allow for a cross-lectal study, we compiled a comparable corpus of Belgian Dutch, with Belgian newspaper articles from the same years. It contains just below 500 million word tokens. We will refer to it as the Leuvens Nieuws Corpus (LeNC).

Gold standard. We compiled a gold standard on the basis of the *Reference List of Belgian Dutch* (*Referentiebestand Belgisch Nederlands*, RBBN) (Martin 2005). This is a list of almost 4,000 words and expressions that are typical of Belgian Dutch, together

https://doi.org/10.1017/S1351324910000161 Published online by Cambridge University Press

with their synonyms in Netherlandic Dutch, if any exist. For this first experiment, we selected all one-word adjective, noun and verb markers of Belgian Dutch with a one-word Netherlandic Dutch synonym. We required all Belgian markers to occur at least a hundred times in the Leuvens Nieuws Corpus and the Netherlandic synonyms to appear at least fifty times in the Twente Nieuws Corpus. This resulted in a test set of 939 words: 656 nouns, 121 adjectives and 162 verbs.

Models. We implement a syntax-based distributional model and word-based models with context sizes ranging from two to six words on both sides of the target word. The syntax-based model takes into account eight syntactic relations for the modelling of noun meaning: subject of verb v, direct object of verb v, head of prepositional complement of verb v with preposition p, head of adverbial PP of verb v with preposition p, modified by adjective a, post-modified by PP with preposition p and head noun n, modified by apposition with head n, coordinated with head n. For verb and adjective meaning, we take the inverse of the relevant relations. The models contain the context vectors of all words in the Belgian corpus (with a minimum frequency of 100) and all words in the Netherlandic corpus (with a minimum frequency of 50). They combine all context features – either context words or syntactic relations – from both corpora. Co-occurrence frequencies between a word and its contextual features are weighted with point-wise mutual information. We measure distributional similarity with Lin's metric (Lin 1998), weighted Jaccard (Grefenstette 1994) and the cosine of the angle between two context vectors (Salton and McGill 1983).

For each of the markers of Belgian Dutch, we identify the one hundred Netherlandic nearest neighbours with the same part of speech in the relevant model. We will focus on the results of the cosine, as this was the metric with the best results, albeit by a small margin. We evaluate each model by computing the accuracy of its nearest neighbours. This accuracy is defined as the proportion of Belgian markers that have at least one correct Netherlandic synonym among their n nearest neighbours.

3.2 Results

Figure 1 plots the observed accuracy of the n nearest neighbours for the syntax-based model, and the word-based models with context sizes two, three and five. This accuracy climbs more or less linearly with the logarithm of the number of nearest neighbours. In the syntax-based model, 234 (24.9%) of the Belgian markers have a correct Netherlandic Dutch synonym as their single nearest neighbour. Fifty percent of the markers have a correct synonym among their eight nearest neighbours.

Although the syntax-based model attains the best performance, the difference with the word-based models is not statistically significant ($\chi^2 = 0.4177$, $df = 1$, $p = .52$ for the best word-based model). The word-based model based on context windows of five words on both sides of the target is able to correctly identify 221 (23.5%) of the cross-lectal nearest neighbours. Other word-based models give a slightly lower number. Only the word-based model with context size two performs

https://doi.org/10.1017/S1351324910000161 Published online by Cambridge University Press

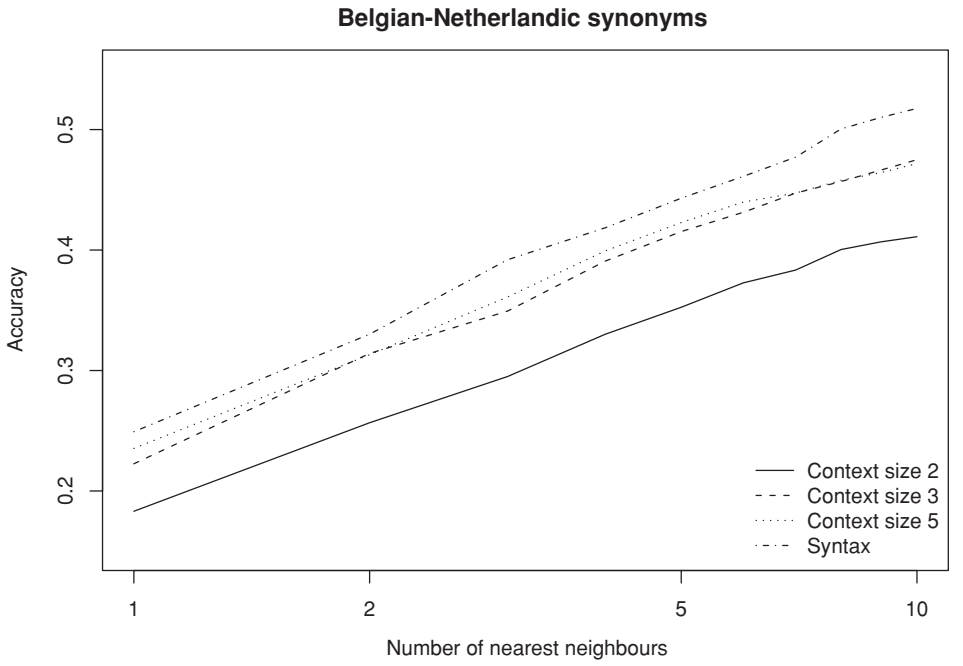


Fig. 1. Proportion of lectal markers of Belgian Dutch with at least one correct Netherlandic Dutch synonym among their *n* cross-lectal nearest neighbours.

Table 1. *Examples of Belgian markers and their Netherlandic synonym identified correctly by the word-based distributional model with context size five*

Belgian marker	Netherlandic synonym	Meaning
microgolf	magnetron	'microwave oven'
schoonbroer	zwager	'brother-in-law'
confituur	jam	'jam'
bankbriefje	bankbiljet	'bank note'
fier	trots	'proud'
gamma	assortiment	'assortment'
job	baan	'job'
living	woonkamer	'living room'
proper	schoon	'clean'
uitbaten	exploiteren	'exploit'

below par. Its accuracy is significantly lower than that of all other models, including the word-based model with context size three ($\chi^2 = 4.2673$, $df = 1$, $p = .039$).

Table 1 gives a few examples of Belgian markers whose Netherlandic synonym was correctly identified by the word-based distributional model with context size five. For instance, a microwave oven is called *microgolf* in Belgium, but *magnetron* in the Netherlands. Similarly, *schoonbroer* is a typically Belgian word for 'brother-in-law', with *zwager* as its Netherlandic Dutch synonym, and *confituur* is the Belgian

https://doi.org/10.1017/S1351324910000161 Published online by Cambridge University Press

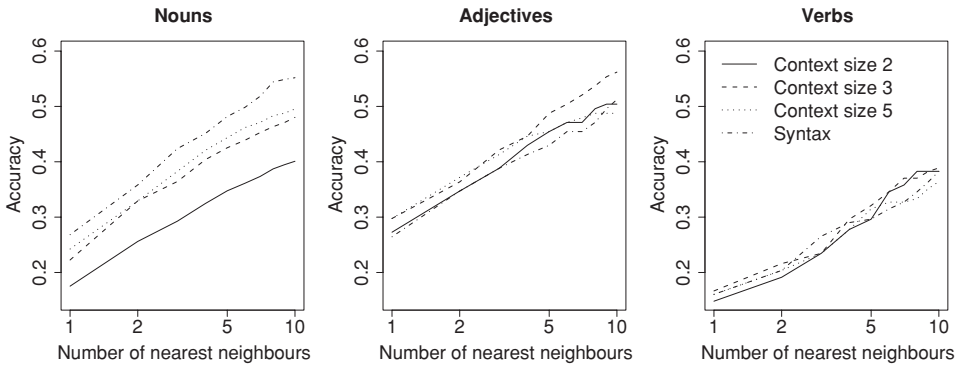


Fig. 2. Proportion of lectal markers of Belgian Dutch with at least one correct Netherlandic Dutch synonym among their *n* cross-lectal nearest neighbours for nouns, adjectives and verbs.

synonym to Netherlandic *jam*. Most of these Belgian words occur very often in the Belgian newspaper corpus, and can thus be considered 'standard Belgian Dutch'.

Because the Reference List of Belgian Dutch was targeted specifically at the identification of lectal markers of Belgian Dutch, we started with the identification of the Netherlandic Dutch synonyms to these Belgian markers. When we use the same models to try and identify the markers of Belgian Dutch that are synonyms to the Netherlandic words in the Reference List, the results are less impressive. We attribute this to the nature of the gold standard. In the word-based model with context size four, for example, the Belgian Dutch single nearest neighbours correspond to the intended Belgian markers in 15.6 percent of the cases (20.4% for adjectives, 15.9% for nouns and 10.6% for verbs). One advantage of this complementary perspective is that symmetric nearest neighbours, i.e. those cases in which the Belgian lectal marker is also the single nearest neighbour of its Netherlandic single nearest neighbour, have a much higher likelihood of being correct. Of the 255 symmetric pairs in the model with a Belgian marker, 43 percent also include a correct Netherlandic synonym of that marker.

Part of speech. The success of the distributional models depends on a number of factors. The first is the part of speech of the word involved. As shown in Figure 2, it is easiest to find the Netherlandic synonym to a Belgian adjective or noun. Verbs present the most difficult task. The accuracy of verb neighbours also climbs less steeply with the number of nearest neighbours taken into account. Also note that the difference between the different distributional models is larger for the nouns than for the verbs and the adjectives. This can be explained by the fact that the most informative clue about the meaning of an adjective is the noun that it modifies, which in text mostly follows immediately. Similarly, the meaning of Dutch verbs reveals itself mainly in arguments that tend to occur close to it. The separate analyses in Figure 2 also show that the syntax-based model mainly thanks its success to a gain in performance for the nouns. Its performance for adjectives and verbs is basically the same as that of the word-based models. The explanation may lie in our choice of syntactic relationships (see Section 3.1).

https://doi.org/10.1017/S1351324910000161 Published online by Cambridge University Press

Table 2. *The influence of frequency on the automatic identification of Netherlandic synonyms to Belgian markers in the syntax-based distributional model. The observed accuracy is shown for nine frequency groups, which distinguish between low-frequency, mid-frequency and high-frequency Belgian markers and low-frequency, mid-frequency and high-frequency Netherlandic synonyms*

	nl.low	nl.mid	nl.high
be.low	17.0%	31.8%	10.0%
be.mid	27.0%	45.8%	30.1%
be.high	2.6%	24.1%	26.3%

Frequency. The second factor that influences the success of automatic recovery is the frequency of the relevant words. In order to gauge its effect, we divided the Belgian markers with exactly one cross-lectal synonym in three bins with more or less the same number of words: one for low-frequency ($f < 400$), one for mid-frequency ($400 \leq f < 1600$) and one for high-frequency words ($f \geq 1600$). We separated their Netherlandic Dutch synonyms in the same way. This gave nine possible pairwise combinations, for each of which we calculated the accuracy of the models. Table 2 shows the effect of frequency on the accuracy of the single nearest neighbours in the syntax-based model. Two situations lead to poor performance. First, the model generally fails to capture the similarity between two words with very different frequencies. This is obvious in the low accuracies of the classes `be.low–nl.high` and `be.high–nl.low`. Second, the syntax-based model is generally not able to extract cross-lectal word pairs with two infrequent words. It shares this weakness with the word-based models with low context sizes. Word-based models with larger context sizes are able to collect more data for infrequent words and therefore perform better in this category.

3.3 Error analysis

When we manually go through the Belgian markers for which the distributional models fail to find the correct Netherlandic Dutch synonym, we find two interesting sets of problematic cases. The first is made up of polysemous words that have a typically Belgian minority sense. The second group is that of markers of colloquial Belgian Dutch.

Polysemy. A first problematic group is constituted by those words that mark Belgian Dutch in one minority sense only. Table 3 gives three such examples, and their three cross-lectal nearest neighbours as identified by the word-based distributional model with context size five. The other models, including the syntax-based model, gave similar results. Because these words have the same majority meaning in Belgium as in the Netherlands, their nearest Netherlandic neighbour is often the Belgian marker itself, but in its capacity of supralectal standard Dutch word. The Dutch word *stelling*, for instance, means ‘position’ in both of the lects

https://doi.org/10.1017/S1351324910000161 Published online by Cambridge University Press

Table 3. *Examples of words that are markers of Belgian only in a minority sense, together with their cross-lectal nearest neighbours in a word-based distributional model with context size five*

Belgian marker	Synonym	Cross-lectal nearest neighbours
stelling 'position'	steiger 'scaffold'	stelling 'position' opvatting 'opinion' standpunt 'position'
katholiek 'catholic'	goed 'good'	katholiek 'catholic' rooms-katholiek 'roman catholic' protestants 'protestant'
tenor 'tenor'	topper 'steersman'	tenor 'tenor' sopraan 'soprano' alt 'alto'

under investigation. However, in Belgium it can also refer to scaffolding. Because this meaning is far less frequent, it is not (or scarcely) picked up by the context vector that represents its distribution. As a result, the Netherlandic nearest neighbours are all related to the 'position' sense of *stelling*. Similarly, the word *katholiek*, 'catholic', can mean something like 'good' or 'kosher' in Belgium, but only in contexts like *zijn methode was niet echt katholiek*, 'his method was not exactly kosher'. This meaning is again far less frequent than the dominant one, which all Netherlandic nearest neighbours are related to. A possible solution to this problem involves two substantial changes to the current approach. First, we obviously need a corpus in which the relevant meaning is well-represented. Second, we need to make our distributional model sensitive to the several meanings of polysemous words, for instance by grouping distributional neighbours to represent senses (Pantel and Lin 2002) or by assembling the individual contexts of the target words into distributional clusters that correspond to its several meanings (Schütze 1998).

Colloquial words. A second problematic group is made up of those Belgian markers that are generally considered colloquial (see Table 4). Here the synonyms taken up in the Reference List are often the counterparts to these words in Standard Dutch, rather than in Netherlandic Dutch only. This is the case for Belgian *boerenbuiten*, for instance, a colloquial word for 'countryside'. Due to its colloquial status it occurs in the Belgian corpus only 115 times, and its Netherlandic nearest neighbours are not even vaguely related to the concept it refers to. They also vary widely between the several distributional models that we investigated. Similarly, the Belgian verb *flessen*, a homograph with the plural of *fles* 'bottle', is a colloquial Belgian synonym to Netherlandic *zakken*, 'to fail'. Here, the fact that *fles* occurs as a verb only 178 times in the corpus results from two facts. First, there is the colloquial status of the word, which makes it less appropriate for newspaper language. Second, the nearest Netherlandic neighbours show that the lemmatiser, which was developed in the Netherlands, tends to break down when *fles* is used as a verb. Hence, these neighbours are all linked to the homograph *fles*, 'bottle'. Note, finally, that this group

https://doi.org/10.1017/S1351324910000161 Published online by Cambridge University Press

Table 4. *Examples of colloquial Belgian markers, together with their cross-lectal nearest neighbours in a word-based distributional model with context size five*

Belgian marker	Synonym	Cross-lectal nearest neighbours
boerenbuiten 'countryside'	platteland	drumsolo 'drum solo' tussenspel 'intermezzo' polyfonie 'polyphony'
flessen 'fail'	zakken	uitschenken 'pour out' aanrukken 'have (another bottle)' recyclen 'recycle'
bal 'ball/franc'	frank 'franc'	bal 'ball' doelman 'goalkeeper' schot 'shot'

of problematic Belgian markers overlaps with the previous one. This is illustrated by a word like *bal*, 'ball/franc', which marks colloquial Belgian only in its minority sense of 'franc'.

3.4 Discussion

Our experiment has shown that distributional models are able to identify a correct Netherlandic synonym directly for about 25 percent of the Belgian words in the Reference List of Belgian Dutch. Another 25 percent of the Belgian markers had a correct Netherlandic synonym among their eight nearest neighbours. Noun and adjective synonyms proved easier to identify than verb synonyms, and accuracy was also higher for word pairs with similar and relatively high frequencies. A manual analysis of the results revealed two particular weaknesses of our approach. The first is related to the distributional models that we investigated. As these models construct only one context vector for each target word, they are often unable to deal with polysemy. The second problem is related to our choice of corpora. Because many of the words in the Reference List of Belgian Dutch are colloquial, they do not occur frequently in a corpus of newspaper language. In order to model the semantics of such colloquial words, we need a corpus that is representative of colloquial language.

We foresee several possible applications of this distributional method for the identification of cross-lectal synonyms. First, these bilectal distributional models can be helpful in lexicography, where they can inform lexicographers about the lectal status of certain words and suggest possible synonyms in other language varieties. Second, we aim to apply these methods to the study of language variation, where researchers are looking for methods to establish synonymy across different language varieties, in order to study differences in the use of these synonyms (Geeraerts 2010).

4 Experiment 2: The identification of lectal markers

The bilectal distributional models have more applications than the automatic identification of cross-lectal synonyms. In this section we will show how the notion

https://doi.org/10.1017/S1351324910000161 Published online by Cambridge University Press

of cross-lectal distributional similarity can help us identify the words that are typical of a given lect.

4.1 Background

The most obvious corpus-based way of identifying lectal markers consists in compiling two corpora, representative of the two relevant language varieties, and calculating the keywords of one corpus with respect to the other. Traditionally, such keywords are defined as words that occur significantly more often in one of the two corpora. These significant differences in frequency are generally identified on the basis of a chi-square or log-likelihood test (Dunning 1993; Scott 1997; Rayson, Berridge and Francis 2004). When it comes to the identification of lectal markers, however, this frequency-based keyword approach is not without its problems.

First, there are two types of lectal markers that we would like to find. The first type consists of markers that distinguish themselves on the basis of their higher frequency in the specific lect. The second type subsumes those markers with a different meaning in one lect, compared to the other. The frequency-based keyword approach is solely geared towards the identification of the first type of lexical variation, which surfaces through differences in frequency of use. Meaning differences, by contrast, do not necessarily correlate with differences in frequency. Words with a different meaning in a specific lect, but a frequency that is not much higher than that in the other lect, will never receive a high keyness value in the traditional approach.

Second, the traditional keywords do not always instantiate the linguistic phenomenon we are interested in. A first explanation for this phenomenon is of statistical nature. When we are dealing with large numbers of data, statistical procedures, such as the chi-square test, tend to return many false positives. This was already noted by Kilgarriff (2001), who claims, for instance, that '[t]he LOB-Brown differences cannot in general be interpreted as British-American differences'. He notes that 'for very common words, high χ^2 values are associated with the sheer quantity of evidence and are not necessarily associated with a pretheoretical notion of distinctiveness', and later shows that 'even where pairs of corpora are set up to be linguistically identical, the null hypothesis is resoundingly defeated' (Kilgarriff 2005). The main problem is that most statistical tests make the assumption that the data are randomly and independently drawn from the population, whereas the occurrence of words is not random at all. A second explanation is that frequency-based methods are not insensitive to differences in topic. The fact that one word occurs more often in one corpus than in another may just mean that a particular topic is treated more frequently in that first corpus. If Belgian newspapers publish more articles about cycling than Netherlandic ones, cycling-related words may receive a high 'keyness' value, even when they do not represent Belgian Dutch *per se*. This kind of differences is not the type of variation we are interested in, and should not be identified by our models.

Because of such problems, we propose an alternative keyword method, based on the distributions of the words instead of their frequencies. The idea is the following: If a word has the same meaning in the two language varieties, it will have similar

https://doi.org/10.1017/S1351324910000161 Published online by Cambridge University Press

distributions in the corpora. If, conversely, a word has two different meanings, this should be reflected in the directions of its two context vectors and in the low cosine of the angle between them.

4.2 Experimental setup

Gold standard. From the Reference List of Belgian Dutch we extracted all single-word markers of Belgian Dutch that occur at least one hundred times in the Leuvens Nieuws Corpus. As the Belgian corpus was lemmatised by a Netherlandic Dutch lemmatiser, some Belgian markers from the Reference List were found to correspond to several lemmas. This happened, for example, when the lemmatiser was not able to identify the lemma of a verb, and merely kept the original forms it met in text. Altogether we assembled 1,796 lemmas that represent lectal markers of Belgian Dutch.

Models. Our implementation of the frequency-based keyword approach is very straightforward. We compared the frequencies of all words that occurred at least one hundred times in the Belgian corpus with their frequencies in the Netherlandic corpus, both on the basis of a chi-square and a log-likelihood test. We then sorted the words according to their 'keyness' value, with the highest values first.

Like the frequency-based method, the distributional approach is developed to find statistically significant differences in behaviour between two lects. This time, however, behaviour is defined in distributional terms: We want to find words whose distributional similarity between two corpora from two different lects is significantly lower than expected. To this goal, we subdivided our Belgian and Netherlandic corpora into five subcorpora. For each word in the Belgian corpus with a minimum frequency of 100 in the entire corpus, we calculated its intra-lectal distributional similarity for each of the pairings of Belgian subcorpora. This gave ten intra-lectal similarity values. Next, we calculated its cross-lectal similarity for each pairing of a Belgian subcorpus and a Netherlandic subcorpus. This gave 25 cross-lectal similarity values. We then applied a Wilcoxon test to determine if the cross-lectal similarity values were significantly lower than the intra-lectal ones. Finally, we ordered all words with the same Wilcoxon statistic according to decreasing frequency.

When the distributional models in these experiments are constructed on the basis of all contextual features from the Belgian and the Netherlandic Dutch corpora, it is only natural for the cross-lectal distributional similarities to be smaller than the intra-lectal ones. This is because the difference in contextual features between two corpora from different lects is likely to automatically bring down distributional similarity. Therefore we constructed both the intra-lectal and the cross-lectal distributional models on the basis of dimensions that we assumed to be neutral between the two language varieties. We extracted the most frequent 5,000 words from the Belgian and the Netherlandic corpora, ignoring stop words and proper nouns, and selected as contextual features only those words that occurred in both lists. We thus assume that words that occur very frequently in both of the language varieties are unlikely to display major differences in use. Because all the different models in this experiment

https://doi.org/10.1017/S1351324910000161 Published online by Cambridge University Press

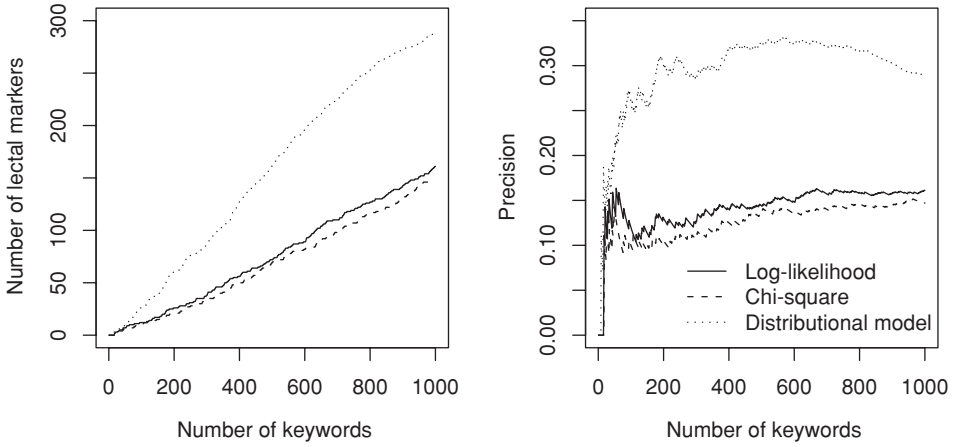


Fig. 3. Performance of the distributional keyword method, compared to the traditional log-likelihood and chi-square methods, expressed in terms of the absolute number of lectal markers identified (left) and relative precision (right).

returned very similar results, here we will focus on the word-based distributional model with a context size of three.

4.3 Results

Figure 3 shows a clear distinction between the results of the two types of keyword approaches in favour of the distributional one. Of the two frequency-based keyword methods, the log-likelihood ratio scores better than the chi-square test, although the difference is not statistically significant. We find 161 lectal markers of Belgian Dutch among the first 1,000 log-likelihood keywords, against 147 among the chi-square keywords. These low figures confirm the earlier observation that frequency-based keywords are not generally markers of the relevant language variety. The distributional method clearly beats this modest performance, with 288 lectal markers from the RBBN in its list of first 1,000 keywords. This is a statistically significant increase, both from the result of the chi-square test ($\chi^2 = 57.5814$, $df = 1$, $p < .001$) and the log-likelihood ratio ($\chi^2 = 45.5946$, $df = 1$, $p < .001$).

This first evaluation gives only an underestimate of the real figures. After all, the Reference List is not an exhaustive list of Belgian markers. To get a better insight into the success of our methods, we manually evaluated the top 500 keywords returned by the log-likelihood ratio and the distributional method. Table 5 gives the frequency of four types of markers. RBBN is the class of all Belgian markers in the Reference List. VD contains words that are either labelled as Belgian or whose meaning refers to Belgium in Van Dale, the standard dictionary of Dutch (Den Boon and Geeraerts 2005). CUL subsumes all cultural terms, like the names of Belgian cities or political parties, while REL is used for words that are morphologically related to a Belgian marker in the Reference List or Van Dale. The resulting figures are indeed considerably higher than those given above. They still clearly favour the distributional method. Around 47 percent of the first 500 keywords identified by the

https://doi.org/10.1017/S1351324910000161 Published online by Cambridge University Press

Table 5. *Results of the manual evaluation of the top 500 keywords returned by the log-likelihood ratio and the distributional method with a context size of 3. Each row gives the number of keywords that belong to one of the four groups of correct markers of Belgian Dutch*

	RBBN	VD	CUL	REL	Total
Log-likelihood	72	33	36	2	143 (28.6%)
Distributional method	162	29	44	2	237 (47.4%)

RBBN: Reference List of Belgian Dutch
VD: Van Dale dictionary
CUL: cultural terms
REL: words morphologically related to Belgian marker in RBBN or VD

distributional method are markers of Belgian Dutch, against 28.6 percent of the log-likelihood keywords. This difference is again statistically significant ($\chi^2 = 36.7105$, $df = 1$, $p < .001$).

The distributional method benefits from two differences with its frequency-based competitors. The first is its reliance on distribution rather than frequency, which allows for differences in meaning to be identified that are not accompanied by differences in frequency. The second is its use of subcorpora. This enables the method to filter out words that are keywords of one newspaper only rather than of the entire Belgian corpus. If they are very frequent, such artifacts are identified as normal keywords by the frequency-based method. A manual analysis of the results reveals the linguistic importance of the former difference in particular, as we will see immediately.

4.4 Linguistic interpretation

We have argued earlier that the two approaches to keyword identification basically target different kinds of variation. The traditional approach looks for statistically significant differences in frequency, while the second approach targets differences in distribution. This does not mean that the distributional approaches are insensitive to frequency variation: large differences in frequency between two corpora will commonly be reflected in lower distributional similarity. Yet, the low overlap between the results of the two methods indicates that they do behave very differently indeed. We calculated this overlap by taking all correct lectal markers from the RBBN in the top n keywords identified by the log-likelihood approach, and then counting how many of these keywords were also identified by the distributional approach. For the top 500 keywords, this overlap is around 60 percent. In other words, although the distributional method returns far more correct keywords than its frequency-based competitor at this stage, it still misses out on more than one-third of the lectal markers identified by the log-likelihood ratio. The overlap tends to stay between 55 percent and 65 percent between the top 200 and 2,000 keywords.

https://doi.org/10.1017/S1351324910000161 Published online by Cambridge University Press

Table 6. *Distribution of onomasiological and semasiological variation for the 200 Belgian markers with the largest difference in ranks between the two keyword methods*

Method	Onomasiological	Semasiological	Total
Frequency	83	17	100
Distribution	1	99	100

This low overlap is due to a difference in bias between the two methods. Lexical differences surface in two types of variation: onomasiological and semasiological variation. *Semasiological variation* occurs when a Dutch word means one thing in Belgium, but another in the Netherlands. This is true for a word like *patat*, which means 'chips' in the Netherlands, but 'potato' in Belgium. There are also quite a number of cases where a word has an additional meaning in Belgium that does not occur in the Netherlands. This is the case for a word like *omslag*, which has a number of shared meanings in the two varieties, along with the typically Belgian meaning 'envelope'. *Onomasiological variation* occurs when words are typical of Belgian Dutch in all of their meanings (often there is only one). This is true for a word like *containerpark*, 'civic amenity site'. Here there is no meaning difference with Netherlandic Dutch, which instead has the synonym *milieupark*. A similar type of variation appears when a Belgian marker refers to a typically Belgian entity (an institution, a political party *etc.*), like the word *schepencollege*, 'bench of (Mayor and) Aldermen'. Broadly speaking, onomasiological variation should be reflected in a difference in frequency between the two corpora, while semasiological differences should correlate with a difference in use. That is why we expect the frequency-based keyword approach to be better at the recognition of onomasiological variation, and the distributional method to be better at the identification of semasiological variation.

We tested this hypothesis by looking for the set of words that best defines the differences between the the two approaches. More particularly, we selected the words from the RBBN with a low rank in the frequency-based output but a high rank in the distributional output, or vice versa. We manually determined the type of variation represented by the hundred markers with the biggest difference in ranks, on the basis of the Van Dale dictionary. We labelled a Belgian marker as an instance of onomasiological variation when all of its meanings in the dictionary refer to Belgium or are explicitly labelled as Belgian Dutch. We treat a word as an instance of semasiological variation when it has several meanings in the dictionary, only one of which is typical of Belgian Dutch. We labelled 200 words in total: the hundred Belgian markers in the dictionary with the largest difference in ranks in favour of the log-likelihood test, and the hundred Belgian markers with the largest difference in favour of the distributional approach. This led to the figures in Table 6. Note that the number of instances of onomasiological variation is probably an underestimate of the corpus-based reality. This is because the theoretical ambiguity in a gold

https://doi.org/10.1017/S1351324910000161 Published online by Cambridge University Press

standard resource is not necessarily reflected by actual ambiguity in our corpora. A word with several possible meanings may occur in only one of its meanings in the corpus, and therefore actually represents onomasiological variation rather than semasiological variation.

The two distributions in Table 6 are significantly different at all usual p-levels, according to a chi-square test ($\chi^2 = 134.6675$, $df = 1$, $p < .0001$). They indeed follow our hypothesis: If we contrast the results of the two methods, we see a clear correlation with the two types of lexical variation. Belgian markers that score much better in the frequency-based approach more often represent onomasiological variation, while those that score much better in the distributional method generally represent semasiological variation. These results show that our distributional keyword method should not be taken as a simple replacement for the classic frequency-based approach. From a linguistic point of view, it makes more sense to compare and contrast the results of the two approaches in order to learn something about the type of lexical variation that is represented by the markers they identify.

4.5 Discussion

In this second experiment we have tested a distributional method for the automatic identification of words that are typical of a given language variety. We have shown that this distributional method retrieves more markers of Belgian Dutch than the traditional, frequency-based keyword methods like the chi-square test and the log-likelihood ratio. We argued that the new method benefits in particular from its use of subcorpora and its sensitivity to differences in meaning. This second element also surfaced in the manual evaluation of the set of words that were most typical of either of the two approaches. This evaluation showed that the frequency-based method has a relative bias towards onomasiological variation, whereas the distributional method displayed a relative bias towards semasiological variation.

Remaining problems affect both the recall and the precision of the keyword methods. As before, recall is affected in particular by the colloquial nature of many of the words in the Reference List of Belgian Dutch. These words are not represented well in our newspaper corpora and are generally missed by our current methods. Precision is influenced negatively by the fact that not all differences in distribution correspond to differences in meaning. For example, the distributional method is still susceptible to differences in topic, because these will be reflected in the distribution of the relevant words. Still, the rather conservative nature of our gold standards suggests that some of the significant differences in distribution we have identified may bring to light lexical variation that is not represented in the standard dictionaries. Apart from corpus linguistics, this variation can prove useful in fields such as terminology extraction, where distributional differences may be used to identify terms typical of a specific jargon, and stylistics, where lists of lectal markers may support the distinction of texts from different lects.

https://doi.org/10.1017/S1351324910000161 Published online by Cambridge University Press

5 Conclusions

The investigation of lexical variation between two language varieties is still in its infancy. One of the most pressing needs of the field is the development of robust models of lexical semantics that can establish the semantic equivalence between two words in different lects. In this paper, we have shown that the paradigm of distributional semantics, which has become popular in computational linguistics in the past few years, can cater to that particular need.

We have demonstrated the variational-linguistic usefulness of distributional models of lexical semantics in two related tasks. Our experiments focused on the national varieties of Dutch in Belgium and the Netherlands. The first task consisted of the identification of cross-lectal synonyms. We found that in many cases, the distributional nearest neighbour to a Belgian lectal marker indeed represented the synonym to this marker in the Netherlandic variety of Dutch. The accuracy of our method was affected in particular by the polysemy of some Belgian markers and their colloquial status. The second task addressed the automatic identification of lectal markers. We showed that an alternative, distributional approach to the traditional notion of keywords was able to identify such lectal markers more successfully than the traditional frequency-based keyword approaches. We found both methods to obtain different results: the distributional method was more sensitive to semasiological variation, whereas the traditional frequency-based methods were biased towards onomasiological variation.

We realise that this study presents only the first step towards a robust application of the distributional paradigm to variational linguistics. Some of the current weaknesses, such as the poor coverage of colloquial markers, can be solved by looking at large corpora of informal language. The ignorance of polysemy represents a deeper problem with the distributional framework that still needs to be addressed in computational linguistics as well. Nevertheless, we believe that the results of our experiments have shown that even at this stage, distributional models of lexical semantics can be a valuable tool in linguistic enquiry. Not only do they present valid alternatives to traditional models, like the frequency-based keyword methods, but they also allow researchers to address questions that they have not been able to investigate before, like the automatic identification of the synonym to a lectal marker in another language variety. This demonstrates the benefit of distributional semantics to fields like variational linguistics, lexicography or terminology extraction.

There are several directions this type of research can now take. On the more linguistic side, there is the straightforward application of distributional semantics to new corpora or new languages. British English and American English provide us with an interesting situation, and the same goes for Belgian French and French French, or German German and Swiss German. Moreover, it could be interesting to attempt a diachronic instead of a synchronic investigation, in a comparison of two corpora from different time periods. At the more computational side, we think it is important to explore some extensions of the models that we have not been able to discuss here. For example, it would be interesting to add multi-word units to the

https://doi.org/10.1017/S1351324910000161 Published online by Cambridge University Press

distributional models. The incorporation of polysemy presents another particular challenge. We hope that these and other questions find an answer in future studies of this topic.

References

Bai, J., Song, D., Bruza, P., Nie, J.-Y., and Cao, G. 2005. Query expansion using term relationships in language models for information retrieval. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management (CIKM 2005)*, pp. 688–695. New York, NY: ACM Press.

Baroni, M., Lenci, A., and Onnis, L. 2007. ISA meets Lara: an incremental word space model for cognitively plausible simulations of semantic learning. In *Proceedings of the ACL Workshop on Cognitive Aspects of Computational Language Acquisition*, pp. 49–56. Stroudsburg, PA: Association for Computational Linguistics.

Bertels, A. 2008. Sémantique quantitative et corpus technique: des analyses statistiques aux interprétations linguistiques. In *Actes des 9es Journées internationales d'Analyse statistique des Données Textuelles (JADT 2008)*, pp. 179–188. Lyon: Presses universitaires de Lyon.

Boussidan, A., Sagi, E., and Ploux, S. 2009. Phonaesthemic and etymological effects on the distribution of senses in statistical models of semantics. In *Proceedings of the CogSci Workshop on Distributional Semantics Beyond Concrete Concepts (DiSCo 2009)*, pp. 35–40. http://www.let.rug.nl/disco2009/proc/ disco2009_proceedings.pdf

Buchanan, L., Burgess, C., and Lund, K. 1996. Overcrowding in semantic neighborhoods: modeling deep dyslexia. *Brain and Cognition* **32**: 111–114.

Burgess, C., Livesay, K., and Lund, K. 1998. Explorations in context space: words, sentences, discourse. *Discourse Processes* **25**: 211–257.

Curran, J. R. 2004. *From Distributional to Semantic Similarity*. PhD thesis, University of Edinburgh, Edinburgh, UK.

Den Boon, T., and Geeraerts, D. (Eds.). 2005. *Van Dale Groot Woordenboek van de Nederlandse taal (14th ed.)*. Utrecht/Antwerp, Belgium: Van Dale Lexicografie.

Divjak, D., and Gries, S. T. 2006. Ways of trying in Russian: clustering behavioral profiles. *Corpus Linguistics and Linguistic Theory* **2**(1): 23–60.

Dunning, T. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* **19**(1): 61–74.

Firth, J. R. 1957. A synopsis of linguistic theory 1930–1955. In J. R. Firth (ed.), *Studies in Linguistic Analysis*, pp. 1–32. Oxford: Blackwell.

Foltz, P. W. 1996. Latent semantic analysis for text-based research. *Behavior Research Methods, Instruments and Computers* **28**(2): 197–202.

Fung, P., and McKeown, K. 1997. Finding terminology translations from non-parallel corpora. In *Proceedings of the 5th Workshop on Very Large Corpora*, pp. 192–202.

Fung, P., and Yee, L. Y. 1998. An IR approach for translating new words from non-parallel, comparable texts. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics (COLING-ACL 1998)*, pp. 414–420.

Gamallo Otero, P. 2008. Evaluating two different methods for the task of extracting bilingual lexicons from comparable corpora. In *Proceedings of the LREC-2008 Workshop on Comparable Corpora*, pp. 19–26. Paris, France: European Language Resources Association.

Gamallo Otero, P., and Pichel Campos, J. R. 2008. Learning Spanish-Galician translation equivalents using a comparable corpus and a bilingual dictionary. In *Computational Linguistics and Intelligent Text Processing*, pp. 423–433. Lecture Notes in Computer Science, vol. 4919. New York, NY: Springer.

https://doi.org/10.1017/S1351324910000161 Published online by Cambridge University Press

Geeraerts, D. 2010. Lexical variation in space. In P. Auer, and J. E. Schmidt (eds.), *Language and Space. An International Handbook of Linguistic Variation*, pp. 820–836. Berlin: De Gruyter Mouton.

Geeraerts, D., Grondelaers, S., and Speelman, D. 1999. *Convergentie en Divergentie in de Nederlandse Woordenschat*. Amsterdam: Meertens Instituut.

Glynn, D. 2007. *Mapping Meaning. Toward a Usage-Based Methodology in Cognitive Semantics*. PhD thesis, University of Leuven, Leuven, Belgium.

Grefenstette, G. 1994. *Explorations in Automatic Thesaurus Discovery*. Dordrecht: Kluwer.

Gries, S. T. 2006. Corpus-based methods and cognitive semantics: the many meanings of *to run*. In S. T. Gries, and A. Stefanowitsch (eds.), *Corpora in Cognitive Linguistics: Corpus-Based Approaches to Syntax and Lexis*, pp. 57–99. Berlin: Mouton de Gruyter.

Harris, Z. 1954. Distributional structure. *Word* **10**(2/3): 146–162.

Jijkoun, V., and de Rijke, M. 2005. Recognizing textual entailment: is word similarity enough? In J. Quinonero-Candela, I. Dagan, B. Magnini, and F. d'Alché-Buc (eds.), *Machine Learning Challenges, Evaluating Predictive Uncertainty, Visual Object Classification and Recognizing Textual Entailment, First PASCAL Machine Learning Challenges Workshop (MLCW 2005), Lecture Notes in Computer Science 3944*, pp. 449–460. New York, NY: Springer.

Kakkonen, T., Myller, N., Timonen, J., and Sutinen, E. 2005. Automatic essay grading with probabilistic latent semantic analysis. In *Proceedings of the 2nd Workshop on Building Educational Applications Using NLP (EdAppsNLP 05)*, pp. 29–36. Stroudsburg, PA: Association for Computational Linguistics.

Kilgarriff, A. 2001. Comparing corpora. *International Journal of Corpus Linguistics* **6**(1): 97–133.

Kilgarriff, A. 2005. Language is never ever ever random. *Corpus Linguistics and Linguistic Theory* **1**(2): 263–276.

Kintsch, W. 2000. Metaphor comprehension: a computational theory. *Psychonomic Bulletin & Review* **7**: 257–266.

Landauer, T. K., and Dumais, S. T. 1997. A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review* **104**(2): 211–240.

Lin, D. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics (COLING-ACL 1998)*, pp. 768–774. Stroudsburg, PA: Association for Computational Linguistics.

Lowe, W., and McDonald, S. 2000. The direct route: mediated priming in semantic space. In *Proceedings of the 22nd Annual Conference of the Cognitive Science Society (CogSci 2000)*, pp. 675–680. Wheat Ridge, CO: Cognitive Science Society.

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

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., and Just, M. A. 2008. Predicting human brain activity associated with the meanings of nouns. *Science* **320**(5880): 1191–1195.

Mohammad, S., Gurevych, I., Hirst, G., and Zesch, T. 2007. Cross-lingual distributional profiles of concepts for measuring semantic distance. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007)*, pp. 571–580. Stroudsburg, PA: Association for Computational Linguistics.

Ordelman, R., de Jong, F., van Hessen, A., and Hondorp, G. 2007. TwNC: a multifaceted Dutch news corpus. *ELRA Newsletter* **12**(3–4): 1–9.

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

https://doi.org/10.1017/S1351324910000161 Published online by Cambridge University Press

Pantel, P., and Lin, D. 2002. Discovering word senses from text. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2002)*, pp. 613–619. New York, NY: ACM Press.

Peirsman, Y., and Geeraerts, D. 2009. Predicting strong associations on the basis of corpus data. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009)*, pp. 648–656. Stroudsburg, PA: Association for Computational Linguistics.

Peirsman, Y., Heylen, K., and Speelman, D. 2007. Finding semantically related words in Dutch. Co-occurrences versus syntactic contexts. In *Proceedings of the Workshop on Contextual Information in Semantic Space Models (CoSMO 2007)*, pp. 34–41. http://clic.cimec.unitn.it/marco/beyond_words/proceedings/proceedingsCosmo.pdf

Rapp, R. 1995. Identifying word translations in non-parallel texts. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL 1995)*, pp. 320–322. Stroudsburg, PA: Association for Computational Linguistics.

Rapp, R. 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 1999)*, pp. 519–526. Stroudsburg, PA: Association for Computational Linguistics.

Rayson, P., Berridge, D., and Francis, B. 2004. Extending the Cochran Rule for the comparison of word frequencies between corpora. In *Le poids des mots. Actes des 7es Journées internationales d'Analyse statistique des Données Textuelles (JADT 2004)*, pp. 926–936. Louvain-la-Neuve, Belgium: Presses universitaires de Louvain.

Sagi, E., Kaufmann, S., and Clark, B. 2009. Semantic density analysis: comparing word meaning across time and phonetic space. In *Proceedings of the EACL 2009 Workshop on GEMS: Geometrical Models of Natural Language Semantics*, Stroudsburg, pp. 104–111. PA: Association for Computational Linguistics.

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

Salton, G., and McGill, M. J. 1983. *Introduction to Modern Information Retrieval*. New York, NY: McGraw-Hill.

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

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

Soares da Silva, A. 2010. Measuring and parameterizing lexical convergence and divergence between European and Brazilian Portuguese: endo/exogeneousness and foreign and normative influence. In D. Geeraerts, G. Kristiansen, and Y. Peirsman (eds.), *Advances in Cognitive Sociolinguistics*. Berlin: De Gruyter Mouton.

Speelman, D., Grondelaers, S., and Geeraerts, D. 2003. Profile-based linguistic uniformity as a generic method for comparing language varieties. *Computers and the Humanities* **37**: 317–337.

Szmrecsanyi, B. 2010. The English genitive alternation in a cognitive sociolinguistics perspective. In D. Geeraerts, G. Kristiansen, and Y. Peirsman (eds.), *Advances in Cognitive Sociolinguistics*. Berlin: De Gruyter Mouton.

Tummers, J., Heylen, K., and Geeraerts, D. 2005. Usage-based approaches in cognitive linguistics: a technical state of the art. *Corpus Linguistics and Linguistic Theory* **1**: 225–261.

Turney, P., and Pantel, P. 2010. From frequency to meaning: vector space models of semantics. *Journal of Artificial Intelligence Research* **37**: 141–188.

Van de Cruys, T. 2008. A comparison of bag of words and syntax-based approaches for word categorization. In M. Baroni, S. Evert, and A. Lenci (eds.), *Proceedings of the ESSLLI Workshop on Distributional Lexical Semantics*, pp. 47–54. http://wordspace.collocations.de/lib/exe/fetch.php/workshop:esslli:esslli_2008_lexicalsemantics.pdf

https://doi.org/10.1017/S1351324910000161 Published online by Cambridge University Press

Van der Plas, L. 2008. *Automatic Lexico-Semantic Acquisition for Question Answering*. PhD thesis, University of Groningen, Groningen, the Netherlands.

Wittgenstein, L. 1953. *Philosophical Investigations*. Oxford: Blackwell.

Wulff, S., Stefanowitsch, A., and Gries, S. T. 2007. Brutal Brits and persuasive Americans: variety-specific meaning construction in the *into*-causative. In G. Radden, K.-M. Köpcke, T. Berg, and P. Siemund (eds.), *Aspects of Meaning Construction in Lexicon and Grammar*, pp. 265–281. Amsterdam: John Benjamins.

Zhitomirsky-Geffet, M., and Dagan, I. 2009. Bootstrapping distributional feature vector quality. *Computational Linguistics* **35**(3): 435–461.

https://doi.org/10.1017/S1351324910000161 Published online by Cambridge University Press