This article discusses a corpus-based method for the automatic identification of synonyms across different varieties of the same language. This method, based on the paradigm of distributional semantics, quantifies semantic similarity on the basis of contextual similarity in two comparable corpora. In two case studies for Dutch and German, we show that it automatically identifies the correct synonym for 31% and 25% of the target words, respectively. A manual error analysis moreover indicates that many additional synonyms are very close in the distributional model, while most other distributional neighbours are semantically related to the target word along other dimensions than synonymy. On the basis of these results, we argue that distributional-semantic methods can play a crucial role in the further evolution of corpus-based lexical semantics to a more quantitative discipline.

Keywords: distributional semantics, lexical variation, pluricentric languages, synonymy

1. Introduction

This article presents a distributional method for the automatic discovery of synonyms across different varieties of pluricentric languages. We will explore how the architecture of current distributional methods, which are usually applied to one corpus and one language variety, can be extended to include language variation. We will advocate the construction of a 'bilectal' semantic space on the basis of two comparable corpora and present the results of two case studies in Dutch and German.

Modelling synonymy, or semantic equivalence, is a crucial task in lexical semantics, where many studies focus on the variation in use between several words, phrases or constructions that broadly express the same concept. In the words of
Labov (1972), this field of corpus linguistics deals with “the option of saying ‘the same thing’ in several different ways: that is, the variants are identical in reference or truth value, but opposed in their social and/or stylistic significance” (Labov 1972: 271). Researchers in this field can greatly benefit from an automatic method for recognizing synonymy. It relieves them of the effort to establish identity in reference or truth value \textit{a priori}, on the basis of thesauri, synonym dictionaries, or the researcher’s own experience with the language and the relevant literature. After all, thesauri or dictionaries may lack coverage, particularly where substandard or specialized language varieties are concerned, and focusing on words whose variability is known in advance may seriously bias the results of the study (Geeraerts et al. 1999, Deygers & Van Den Heede 2000). An automatic method for recognizing synonymy can aid data selection and enable researchers to scale up their studies to a larger subset of the lexicon.

The automatic identification of synonymy is not unique to our study. In the neighbouring fields of corpus linguistics, cognitive science, computational linguistics and natural language processing, semantic similarity is often modelled as distributional similarity. This means the meaning of a word is represented as a function of the contexts in which the word occurs. At the heart of this paradigm of distributional semantics lies an idea that has become known as the distributional hypothesis (Wittgenstein 1953, Harris 1954, Firth 1957): words with a similar meaning are used in similar contexts. It has inspired a long tradition of studies and applications where two words are taken to be synonyms when they share a large number of collocations or syntactic constructions.

The application of distributional methods in lexical semantics is not without its challenges. Geeraerts (2010a) identifies two points that need particular attention. First, we need large corpora that cover the phenomena under investigation. In this study we will focus on two pluricentric languages, Dutch and German, for which comparable newspaper corpora allow us to identify cross-lectal semantic similarity. Second, we need a method that allows researchers to establish objectively that two words are semantically equivalent. We will therefore evaluate the output of a distributional-semantic model in detail, in order to determine how many of the distributionally similar words are indeed established synonyms. This will help us assess to what degree these models can complement manual efforts.

The rest of our article is structured as follows. In the next section, we will give an overview of related work in natural language processing and cognitive science, and discuss the state of the art in corpus linguistics. In Section 3, we will give a practical introduction to the distributional methods that we use for our case study. The results of this study are discussed in Section 4, where we show the success of distributional semantics in recognizing synonyms across language varieties,
and present a detailed analysis of three remaining challenges. We wrap up with conclusions and pointers for future research.

2. Related work

The success of distributional semantics is most obvious in the development of applications in natural language processing. Distributionally similar words can be used for the automatic construction of thesauri (Lin 1998, Curran 2004) or, when corpora of two different languages are involved, for the induction of bilingual lexicons (Fung & McKeown 1997, Rapp 1999, Chiao & Zweigenbaum 2002). Distributional similarity plays a crucial role in the recognition of textual entailment (Jijkoun & De Rijke 2005, Zhitomirsky-Geffet & Dagan 2009) and can therefore support query expansion (Bai et al. 2005) and automatic question answering (Van der Plas 2008). It can moreover help address the problem of data sparseness faced by many applications, because the treatment of unseen examples can be guided by distributionally similar examples in the training data. This can be helpful to tasks like parsing (Clark & Weir 2002) or language modelling (Dagan et al. 1999, Lee 1999).

However, distributional models have a wider reach of influence than natural language processing. In fact, they have been successful as models of various aspects of human cognition. These include language acquisition (Landauer & Dumais 1997, Baroni et al. 2007), semantic dyslexia (Buchanan et al. 1996) and discourse comprehension (Burgess et al. 1998, Foltz 1996, Landauer & Dumais 1997). Lowe & McDonald (2000) rely on word distributions to model mediated priming — the phenomenon whereby a word like lion primes a word like stripes, via the intermediate concept of tiger. Kintsch (2000) applies distributional semantics to metaphor comprehension, by merging the representations of the words that refer to the source and target concepts. Wettler et al. (2005), Michelbacher et al. (2007) and Peirsman & Geeraerts (2009) predict the free associations that people produce when they are presented with a given cue word, on the basis of the distributional behaviour of this cue and the possible associations. Mitchell et al. (2008) use distributional representations of word meaning to model the neural activation observed in the brain when a test person reads a concrete noun.

In linguistics, too, distributional-semantic methods are becoming more popular. For example, it is rather common for researchers to take a set of near-synonyms and study subtle distributional similarities and differences between them. Hanks (1996) uses distributional patterns to distinguish a word like urge from near-synonyms like incite. Gries (2001) quantifies the similarity between English adjectives ending in -ic or -ical (like economic and economical) on the basis of

Still, the paradigm can accommodate other research questions as well. Bertels et al. (2006) rely on distributional semantics to study the polysemy of French technical terms. Based on the variation in contextual distribution, they establish that terms in French technical corpora are often more polysemous than non-technical words. Sagi et al. (2009) apply a similar idea to investigate meaning change. They show that the mean cosine between its individual context vectors can indicate whether the meaning of a word has broadened or narrowed. In a similar vein, Boussidan et al. (2009) study the relatedness in meaning between all English words that share the phonaestheme gl-. Their results indicate that shared phonaesthemes and historic roots have a cohesive effect for the semantics of groups of words. More examples of quantitative research in semantics can be found in Glynn & Fischer (2010). For the overall position of distributional and quantitative approaches in the development of lexical semantics, see Geeraerts (2010b).

There are several reasons why corpus-based lexical semantics can benefit greatly from distributional-semantic methods for the identification of synonyms. First, distributional semantics offers a more objective way of identifying synonyms than researchers’ personal knowledge of a language or familiarity with the literature. The fact that the synonyms are extracted from the corpus under study means only relevant words will be analyzed, and words can be identified irrespective of their presence in synonym dictionaries or thesauri. This may even complement large-scale efforts like FrameNet (Baker et al. 1998), which typically target standard language only. Second, distributional semantics can help scale current studies to a larger subset of the lexicon. It can support studies like Geeraerts et al. (1999) and Soares da Silva (2010), which investigate the lexical variation in pluricentric languages for entire lexical fields, by measuring to what degree their speakers use the same words to refer to the same concept. Geeraerts et al.’s (1999) historical data, from the 1950s, 1970s and 1990s, shows a converging trend between word
use in Belgium and the Netherlands, as speakers of Belgian Dutch appeared to follow the Netherlandic Dutch standard. Distributional semantics would allow similar studies to scale up to a larger number of lexical fields.

3. Background: Distributional semantics

Distributional semantics was inspired by the so-called distributional hypothesis. This hypothesis is reflected, among other places, in Firth’s (1957: 11) maxim that “you shall know a word by the company it keeps”, and is present in varying forms in the works of Wittgenstein (1953) and Harris (1954), too. Distributional methods — semantic spaces, word space models or vector space models, as they are also often called — therefore assume that words that occur in similar contexts will be semantically similar. This allows them to quantify the semantic similarity between two given words in terms of the similarity between their contexts, and to automatically identify a large number of synonyms in a corpus.

We will start our discussion with an overview of the steps that are involved in the extraction of synonyms from a corpus. We first need to model the meaning of each word in the corpus in terms of its contexts. This means we need to settle on a precise working definition of context. This choice will greatly affect the success of our method, and whether it is targeted towards recognizing synonymy or other semantic relations. Then we start the construction of the semantic space. This involves (i) counting co-occurrence frequencies, (ii) weighting these frequencies and (iii) calculating distributional similarity.

3.1 Preparation: Definition of context

The distributional hypothesis posits that semantically similar words occur in similar contexts. However, it does not elaborate on the precise definition of context that is most well suited for the modelling of semantic similarity. Through the years, several definitions have been tried and tested. First, there are document-based models, like Latent Semantic Analysis (Landauer & Dumais 1997). These models take paragraphs or entire documents as contextual units. They assume that words will be semantically similar when they often appear in similar documents. Second, word-based models focus on smaller units of context, in the form of a context window around a word. Here two words are assumed to be semantically similar when they often have the same context words in that window. For example, two synonyms like bike and bicycle will both often co-occur with words like ride, wheel, and car. Words with a different meaning will typically have different context words. Finally, syntax-based models look at the syntactic behaviour of words. Here
two words are taken to be semantically similar when they often appear in the same syntactic relations. To return to the same example, both bike and bicycle will often be the direct object of ride, buy, and sell.

This choice of context definition is extremely important. This is because the contextual or distributional similarity between two words can correspond to many types of semantic similarity, depending on the precise definition of context (Padó & Lapata 2007, Peirsman 2008). When the goal is to recognize synonyms, syntax-based models are the best choice. Syntax-based distributional similarity tends to go hand in hand with strict, taxonomical similarity, as it is found in hyponyms, hypernyms, co-hyponyms and most importantly, synonyms. This is in line with the points made by Janda & Solovyev (2009) and Gries & Stefanowitsch (2004), who argue that syntactic structure is needed for fine-grained models of semantic similarity. Document-based models are at the opposite end of the spectrum. Document-based distributional similarity correlates with a looser type of semantic relatedness (Kilgarriff & Yallop 2000, Sahlgren 2006). This relatedness can take the form of taxonomical similarity, but also of more general association (like the relation between wave and sea or doctor and hospital). Word-based models lie in between these two extremes. Their precise position depends on the size of the context window. When the window contains only a small number of words around the target word, they will behave more like syntax-based models, with a preference for strict similarity. When the window is stretched and contains larger portions of text around the target, they will behave more like document-based models (Sahlgren 2006, Peirsman 2008).

3.2 Frequencies

When a definition of context has been chosen, we need to construct a ‘context vector’ for all words in the corpus. This context vector initially just contains information about the number of times each word occurs in the documents that make up the corpus (document-based models), about what context words it co-occurs with, and how often (word-based models), or what syntactic relations it appears in, and how often (syntax-based models). For example, in Figure 1 we have collected the syntax-based contexts of three target words: the synonyms bike and bicycle and the unrelated word coffee. In our fictitious corpus, we have found that bike occurs as the object of ride 165 times, bicycle 61 times, and coffee 2 times, possibly as the result of incorrect parses. Conversely, coffee is very often post-modified by the prepositional phrase with milk (96 times), but bike and bicycle are not. Finally, all three words can be modified by the adjective black, although this is more frequently the case with coffee than with bike or bicycle.
3.3 Strength of co-occurrence

On the basis of these frequencies, it is possible to see that bike and bicycle have a very similar contextual distribution, and both behave very differently from coffee. In absolute terms, however, the frequencies of their contextual features are still very different. This is because these frequencies are also a result of the target words themselves: bicycle appears less frequently in our corpus than bike does, which accounts for the lower frequencies of its context features. In order to control for this factor, we need to weight the frequencies of the context features.

One initially intuitive solution would be to replace the absolute frequencies by relative ones. In this case we would divide for each target word the frequencies of all its context features by the total number of context features observed together with that target word. However, this gives an unfair disadvantage to those context features that appear most frequently in the corpus. All target words would have very high values for frequent, but uninformative context features like subject of be or object of have. As a result, the vectors of most words will look alike (Lowe 2001). Instead, we want to identify those context features that are typical of a target word, irrespective of their frequency in the corpus.

To do this, we need a weighting scheme that measures the statistical strength of each co-occurrence. One such scheme is point-wise mutual information (PMI, Church & Hanks 1989), which has proved very effective in the past (Peirsman et al. 2010, Turney & Pantel 2010). PMI expresses whether a context feature c co-occurs with target word w more or less often than expected by chance. If the target word and a contextual feature occur independently from one another, their probability of co-occurrence \( p(w,c) \) equals \( p(w) \times p(c) \). Their dependency on each other can thus be measured by taking their actual probability of co-occurrence \( p(w,c) \) and dividing it by their probability of occurrence assuming independence:

\[
PMI(w, c) = -\log \frac{p(w, c)}{p(w) \times p(c)}
\]
The logarithm is added to reduce the influence of extremely high ratios, while the negative sign ensures that high attraction corresponds to a positive number. The probabilities are generally estimated by relative frequencies in the corpus (the so-called maximum likelihood estimation):

$$PMI(w, c) = -\log \frac{relfreq(w, c)}{relfreq(w) \ast relfreq(c)}$$

Figure 1 shows what happens with the context vectors of bike, bicycle and coffee after they have been weighted. As expected, bike and bicycle now have very similar context vectors, despite their different frequencies.

3.4 Similarity

Each target word corresponds to a weighted vector in semantic space. For a given target word, we finally want to identify the most similar words on the basis of these vectors. We will call these most similar words the ‘nearest neighbours’. The similarity function that is most often used for this purpose is the cosine of the angle between two vectors:

$$\cos(v_1, v_2) = \frac{\sum_i (v_{1i} \ast v_{2i})}{\sqrt{\sum_i v_{1i}^2} \ast \sqrt{\sum_i v_{2i}^2}}$$

Thus, for each target word we calculate the cosine between its context vector and those of all other words in the corpus. We then order these words by decreasing cosine, with an ordered list of neighbours as a result.

These steps form the backbone of every distributional-semantic analysis. There is considerable variation possible in the definition of the weighting and the distance functions, and there are other, optional procedures, like the reduction of the original high-dimensional spaces to models of a lower dimensionality. Interested readers can find a far more exhaustive overview of distributional methods in Turney & Pantel (2010). However, in this article we will restrict ourselves to the most popular weighting and distance functions and work in the original, unreduced spaces. In fact, we believe that the extraction of synonyms benefits from a simple, unreduced space that keeps as much information as possible about the original contexts.

4. Lexical variation

Now we will apply semantic spaces to identify synonymy across different varieties of the same language. At first sight, this is an application that current distributional
methods are not very appropriate for, as they are usually applied to one corpus, and one language variety. We will show how their architecture can be extended to include language variation, present our results and discuss in detail some of the challenges to current methods.

4.1 Background

We will focus on lexical variation in two examples of pluricentric languages. Languages are called ‘pluricentric’ when they have developed several standard varieties, for example in different countries. Two classic examples are Dutch, with different standard varieties in Belgium and the Netherlands, and German, with standard varieties in Germany and Austria, among other places. In both cases, there is a fair number of concepts that are expressed differently in the two standard languages. For example, an uncle is called oom in the Netherlands, but nonkel in Belgium. Similarly, the first month of the year is called Januar in Germany, but Jänner in Austria. It is this type of synonyms between two language varieties or ‘lects’ that we would like to recognize automatically.

Distributional methods in their most popular form are of limited use to this type of variational-linguistic study. This is because they are typically constructed on the basis of one corpus only. As most corpora are restricted to one language variety (a single regional variety like British English, or a specific genre like newspaper language), the resulting spaces will only contain frequent words in that variety. This is of course problematic for variational linguistics, where we are interested in the variation between different language varieties.

However, there is no particular reason why semantic spaces should include one corpus only. When we know the contextual distribution of a word in one corpus, we can look for words with a similar distribution in any other corpus. Of course, if this distributional similarity is to mean something in semantic terms, we need to make sure that there is a large overlap between the context features in the two corpora, and that the meaning of these context features is comparable. When we work with comparable corpora of two language varieties that are not too different, like two national varieties of the same language, this is mostly guaranteed. Despite the many lexical differences between Belgian Dutch and Netherlandic Dutch, or Austrian German and German German, most words are shared, with little variation in their use. For example, if a word in Austrian German is often the object of a verb like fahren “drive”, we can expect its German German synonym to have an equally high value for that contextual feature. Similarly, if a Belgian Dutch word often co-occurs with the word huis “house”, so should its Netherlandic Dutch synonym. While there will doubtlessly be context features that contradict this rule, like the words typical of one of the relevant lects, it should hold in the majority of
cases. Therefore the notion of distributional similarity can easily be generalized from one corpus to two corpora of the same language. We will call the resulting distributional model a 'bilectal' semantic space, and the synonyms that we are targeting 'cross-lectal synonyms'.

Figure 2 gives an overview of how a bilectal semantic space is constructed. It shows the possible distribution of the lectally neutral words vader “father” and moeder “mother” in two corpora of Belgian Dutch and Netherlandic Dutch, and of the Belgian word nonkel “uncle” and its Netherlandic Dutch synonym oom “uncle”. The two example context features that serve as dimensions are vrouw “woman” and ouders “parents”. We expect moeder “mother” to co-occur more frequently with vrouw “woman” than vader “father” or nonkel/oom “uncle” do. Similarly, vader “father” and moeder “mother” will normally co-occur more often with ouders “parents” than nonkel/oom “uncle” does.

In the traditional setting, like the one we described above, we work with monolectal spaces that are constructed on the basis of one corpus, representative of one particular lect. This situation is represented at the top of Figure 2. The

![Diagram of bilectal semantic space](image-url)
Belgian monolectal space allows us to calculate the distributional similarity between two words in the Belgian corpus; the Netherlandic monolectal space allows us to compute the similarity between two words in the Netherlandic corpus. In order to determine the similarity between nonkel, a typically Belgian word, and oom, a typically Netherlandic word, we need to merge the two monolectal spaces. The context features of the resulting space can be defined as either the intersection or the union of the context features of the original spaces — we will opt for the latter. We then populate the new bilectal space with the vectors from both original spaces.

The bilectal space at the bottom of Figure 2 now allows us to identify the Netherlandic Dutch synonyms to Belgian Dutch words. When there is little or no variation in the use of a word between the two lects, its two context vectors will generally be symmetric nearest neighbours in the bilectal space. This is the case for vader and moeder. By contrast, if we look for the cross-lectal nearest neighbour to a word typical of a specific language variety, we are likely to find the synonym to that word in the other variety. Belgian Dutch nonkel in Figure 2 has its Netherlandic Dutch synonym oom as its nearest neighbour, and vice versa. In this way, bilectal spaces can thus assist in modelling semantic identity across language varieties.

This application of distributional semantics to two corpora is not completely new. However, so far this type of spaces has only been applied to corpora from two languages rather than corpora from different language varieties. In such a bilingual setting, it is possible to use distributional similarity for the automatic identification of translations, and the construction of bilingual lexicons (Rapp 1999). Intra-lingual translation has been attempted (albeit with sub-corpora from the same language varieties), but only as a test bed for cross-lingual approaches (Diab & Finch 2000, Fung & McKeown 1997).

4.2 Setup of the case studies

We collect our distributional statistics from large newspaper corpora. For Dutch, we use the Twente Nieuws Corpus (250 million words) as our Netherlandic corpus, and the Leuven Nieuws Corpus (500 million words) as our Belgian corpus. The two corpora are very similar, as they contain newspaper articles from national newspapers from the same years (1999–2002). They have moreover been parsed automatically, so that we can model the syntactic behaviour of words. For German, we use parts of the Huge German Corpus (HGC) for German German and the Deutsches Referenzkorpus (DEREKO) for Austrian German, for a total of 200 million words of German newspaper articles and 200 million words of Austrian newspaper articles. The German corpora have not been parsed automatically, so
we restrict this part of our study to word-based models only. The full data flow is shown in Figure 3.

We construct several bilectal semantic spaces for the Dutch and German corpora, according to our description above. For the word-based spaces, we will report our results with a context window of five words to either side of the target word. We experimented with other context sizes as well, but found context size five to give the best results overall. For Dutch, we also present the results of a syntax-based semantic space based on eight syntactic relations: (i) subject of verb \( v \), (ii) direct object of verb \( v \), (iii) prepositional complement of verb \( v \) introduced by preposition \( p \), (iv) head of an adverbial prepositional phrase (PP) to verb \( v \) introduced by preposition \( p \), (v) modified by adjective \( a \), (vi) postmodified by a

Figure 3. A flowchart of the data in our case studies
PP with head $n$, introduced by preposition $p$, (vii) modified by an apposition with head $n$, and (viii) coordinated with head $n$. Each instantiation of $v$, $a$, $p$ and $n$ was responsible for a new context feature. Feature frequencies were weighted by PMI. Distributional similarity was calculated by the cosine.

For the evaluation of our method, we use established resources of words that are typical of Belgian Dutch and Austrian German. We call these words ‘markers’ of the given language variety. For Dutch, we rely on the *Referentiebestand Belgisch Nederlands* “Reference List of Belgian Dutch” (Martin 2005), a list of 4,000 words and expressions typical of Belgian Dutch, along with their Netherlandic Dutch alternative, if one exists. From this list, we extract all Belgian words that occur at least 100 times in the Belgian corpus that have a Netherlandic synonym that occurs at least 50 times in the Netherlandic corpus. This gives us a test set of 927 words typical of Belgian Dutch, together with a list of their possible cross-lectal synonyms. We moreover improve the coverage of these lists of synonyms by adding synonyms from the standard Van Dale dictionary (Den Boon & Geeraerts 2005) that are not found in the Reference List. For German, we rely on the *Variantenwörterbuch des Deutschen* “Variant Dictionary of German” (Ammon et al. 2004). From this dictionary we extract all Austrian words with one meaning in the dictionary that occur at least 50 times in the Austrian corpus and that have a German German synonym that occurs at least 50 times in the German corpus. This gives us a test set of 364 typically Austrian words. We evaluate our methods by identifying for each of the Belgian and Austrian words their nearest cross-lectal neighbour in the semantic space, i.e. the Netherlandic-Dutch or German-German word with the highest distributional similarity, and determining if this nearest neighbour is indeed a correct synonym.

4.3 Results

Figures 4 and 5 show the results for the Belgian-Dutch and Austrian-German data, respectively. For Dutch we display the performance of the syntax-based model and the word-based model with a context of five words. For German we focus on the word-based model with context five.

Each line in Figures 4 and 5 shows the evolution of precision (P) against recall (R). Precision is the number of single nearest neighbours that are correct synonyms, divided by the number of words taken into account. Recall is the number of correct synonyms divided by the size of the total test set, irrespective of how many words we take into account. If we take all words in the test set into account, 31.6% of their single nearest neighbours in the Dutch syntax-based space, and 25% of their single nearest neighbours in the German word-based space are indeed correct cross-lectal synonyms. In line with previous findings, the results of
The corpus-based identification of cross-lectal synonyms in pluricentric languages

Figure 4. Precision and recall of the cross-lectal nearest neighbours in the Dutch syntax-based and word-based spaces, with the pairs ordered by decreasing cosine, from left to right.

Figure 5. Precision and recall of the cross-lectal nearest neighbours in the German word-based space, with the pairs ordered by decreasing cosine, from left to right.

The syntax-based space for Dutch (31.6%) are slightly better than those for the word-based space (29.3%). The difference between the two languages likely results
from a variety of factors, like the size of the corpora (bigger is better) and the frequency of the words in the gold standard.

When the full gold standard is looked at, precision equals recall. Figures 4 and 5 show that this is not necessarily the case. As long as we take only one nearest neighbour into account, we cannot achieve a higher recall score. However, we can reach a higher precision score by selecting only those nearest neighbours that have a relatively high cosine value. For example, if we discard all nearest neighbours with a cosine similarity of less than .06 to their target word in the Dutch syntax-based space, the precision of the remaining nearest neighbours increases to 40%. A cosine threshold of .10 brings precision to 45%. Of course, this has an adverse effect on recall. It will generally depend on the nature of the precise study whether a higher recall or a higher precision is desirable, but even with a low recall distributional methods may suggest synonyms that are overlooked by other sources. The figure for Dutch also shows that the cosine might be a better measure of confidence in a word-based space than in a syntax-based space. However, this observation is based on a relatively small number of words, and need not apply in general. It is also important to note that cosine values are not absolute. They will depend on the size of the corpus, the frequencies of the target word and its neighbour, and the weighting function, among other factors.

Some applications may benefit from high recall rather than high precision. In this case, it is beneficial to look at a higher number of nearest neighbours for each word. In the Dutch syntax-based semantic space, for example, 38% of the Belgian markers have at least one correct Netherlandic synonym among their two nearest neighbours. 49% have at least one correct Netherlandic synonym among their five nearest neighbours. While this approach may require more manual correction than the previous one, its ability to find synonyms for a large number of words may make it attractive for researchers that target high-coverage studies.

Let us look at a few example cross-lectal synonym pairs that the models identify correctly. Tables 1 and 2 show ten random examples for Dutch and Austrian, respectively. For example, *jam* is generally called *confituur* in Belgium, but *jam* in the Netherlands. The Belgian political function of *schepen* corresponds to the Netherlandic function of *wethouder*. We also find our earlier example pair *nokkel-oom*. Similarly, as we mentioned earlier, the first month of the year is called *Januar* in Germany, but *Jänner* in Austria. A Christmas tree is a *Weihnachtsbaum* in Germany, but a *Christbaum* in Austria. We find 292 such correct pairs in our Dutch syntax-based semantic space, and 91 in our German word-based space.
Table 1. Belgian Dutch target words and their Netherlandic Dutch synonyms that were identified correctly by the Dutch syntax-based semantic space

<table>
<thead>
<tr>
<th>Target word</th>
<th>Synonym</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Confituur</td>
<td>Jam</td>
<td>“jam”</td>
</tr>
<tr>
<td>Schepen</td>
<td>Wethouder</td>
<td>“alderman”</td>
</tr>
<tr>
<td>Nonkel</td>
<td>Oom</td>
<td>“uncle”</td>
</tr>
<tr>
<td>Bankbriefje</td>
<td>Bankbiljet</td>
<td>“bank note”</td>
</tr>
<tr>
<td>Fier</td>
<td>Trots</td>
<td>“proud”</td>
</tr>
<tr>
<td>Job</td>
<td>Baan</td>
<td>“job”</td>
</tr>
<tr>
<td>Living</td>
<td>Woonkamer</td>
<td>“living room”</td>
</tr>
<tr>
<td>Proper</td>
<td>Schoon</td>
<td>“clean”</td>
</tr>
<tr>
<td>Uitbaten</td>
<td>Exploiteren</td>
<td>“exploit”</td>
</tr>
<tr>
<td>Sanctioneren</td>
<td>Bestraffen</td>
<td>“sanction”</td>
</tr>
</tbody>
</table>

Table 2. Austrian German target words and their German German synonyms that were identified correctly by the German word-based semantic space

<table>
<thead>
<tr>
<th>Target word</th>
<th>Synonym</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jänner</td>
<td>Januar</td>
<td>“january”</td>
</tr>
<tr>
<td>Christbaum</td>
<td>Weihnachtsbaum</td>
<td>“christmas tree”</td>
</tr>
<tr>
<td>Fleischhauer</td>
<td>Fleischer</td>
<td>“butcher”</td>
</tr>
<tr>
<td>Spital</td>
<td>Krankenhaus</td>
<td>“hospital”</td>
</tr>
<tr>
<td>Melanzani</td>
<td>Aubergine</td>
<td>“aubergine”</td>
</tr>
<tr>
<td>Prozентuell</td>
<td>Prozental</td>
<td>“percentage-wise”</td>
</tr>
<tr>
<td>Neuerlich</td>
<td>Erneut</td>
<td>“repeated”</td>
</tr>
<tr>
<td>Entlehenen</td>
<td>Entleihen</td>
<td>“borrow”</td>
</tr>
<tr>
<td>Erzeugen</td>
<td>Produzieren</td>
<td>“produce”</td>
</tr>
<tr>
<td>Ansuchen</td>
<td>Beantragen</td>
<td>“apply”</td>
</tr>
</tbody>
</table>

Our suggestion is that sets of synonyms that have been generated in this way can serve as the input to studies of lexical variation. For example, these cross-lectal synonyms can be the foundation of a diachronic study along the lines of Geeraerts et al. (1999), where corpora of different time periods give an insight in the changing distributions of these synonyms across the relevant language varieties. The automatically identified synonyms can serve as the input to the calculation of convergence, while the accompanying cosines can serve as a measure of their influence (Ruette et al. 2014). We will discuss further possible applications in the conclusions to this article.
4.4 Error analysis

The results above also indicate that in many cases, the single nearest neighbour to a target word is not its intended cross-lectal synonym. A manual analysis of these errors sheds light on the precise behaviour of semantic spaces, and is indispensable for a deeper understanding of how these models work. In fact, most errors can be explained by a small number of factors: polysemy, the presence of different semantic relations, and a low frequency of the target word or its synonym.

4.4.1 Polysemy

One of the most pervasive problems in distributional semantics is the modelling of polysemous words. If a word has several meanings, these all influence its contextual distribution. As a result, the position of its context vector in the semantic space will correspond to some sort of average of these meanings, weighted by their frequency. If one meaning is far more frequent than the other ones, the contextual distribution and nearest neighbours will mainly reflect this frequent meaning. In our case studies, polysemy is therefore only a problem when we deal with minority senses, either of the lectal marker or its cross-lectal synonym.

Let us take a look at a number of examples. First, there are polysemous words that have only one meaning typical of Belgian Dutch or Austrian German. In that case, our gold standards contain the cross-lectal synonym to this meaning only. If this meaning is infrequent, compared to the other meanings of the word, distributional models will generally not be able to identify that cross-lectal synonym. Tables 3 and 4 give a few examples for Dutch and German.

For instance, stelling is used both in Belgium and the Netherlands with the meaning “position”, but only in Belgium with the meaning “scaffold”. Because the meaning “position” is far more frequent in newspaper articles, however, all Netherlandic nearest neighbours are related to it. Similarly, the Dutch word katholiek “catholic” can also mean something like “good” or “kosher” in Belgium, in contexts like zijn methode was niet echt katholiek “his method was not exactly kosher”. This meaning, however, is far less frequent than the dominant one, which all Netherlandic nearest neighbours are related to. The German word Trainer, finally, means “trainer” in both Austrian and German German, but in Austrian German it can also refer to a tracksuit. Still, as this second sense is far less frequent in a newspaper corpus, the position of Trainer in the semantic space only reflects its majority meaning. This is again apparent from the nearest German German neighbours Trainer “trainer”, Coach “coach” and Mannschaft “team”. These examples illustrate how the distributional representation of a word in a semantic space indeed mainly depends on the majority meaning of that word. This majority meaning is reflected in its nearest neighbours.
Second, the problem may also lie in the polysemy of the cross-lectal synonym. This is the case, for example, with the Belgian word leiband “leash” and its Netherlandic synonym lijn “leash/line”. Leiband has only one meaning, but lijn is not in the list of nearest neighbours because it is an extremely polysemous word with a majority meaning fully unrelated to dogs and their collars. As a result, its context vector is hardly influenced by this meaning, and it ends up somewhere else in the space. Polysemy thus affects our approach in two ways, either in the representation of the lectal marker or in that of its cross-lectal synonym. Sometimes these two problems occur at the same time. This is true for the Dutch pair darm/slang, where darm “intestine” marks Belgian Dutch only in its minority meaning of “hose”, which is also only a minority meaning of Netherlandic Dutch slang “snake”.

Table 3. Polysemous Belgian Dutch markers, their correct Netherlandic Dutch synonym and Netherlandic Dutch nearest neighbours in the syntax-based semantic space

<table>
<thead>
<tr>
<th>Belgian marker</th>
<th>Dutch synonym</th>
<th>Dutch nearest neighbours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Katholiek</td>
<td>Goed</td>
<td>Katholiek “catholic”</td>
</tr>
<tr>
<td>“catholic/good”</td>
<td>“good”</td>
<td>Rooms-katholiek “roman-catholic”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Protestants “protestant”</td>
</tr>
<tr>
<td>Stelling</td>
<td>Steiger</td>
<td>Stelling “position”</td>
</tr>
<tr>
<td>“position/scaffold”</td>
<td>“scaffold”</td>
<td>Standpunt “position”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Opvatting “opinion”</td>
</tr>
<tr>
<td>Tenor</td>
<td>Topper</td>
<td>Bariton “baritone”</td>
</tr>
<tr>
<td>“tenor/steeersman”</td>
<td>“steersman”</td>
<td>Operazanger “opera singer”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sopraan “soprano”</td>
</tr>
<tr>
<td>Darm</td>
<td>Slang</td>
<td>Milt “spleen”</td>
</tr>
<tr>
<td>“intestine/hose”</td>
<td>“snake/hose”</td>
<td>Long “lung”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Lever “liver”</td>
</tr>
<tr>
<td>Leiband</td>
<td>Lijn</td>
<td>Riem “leash”</td>
</tr>
<tr>
<td>“leash”</td>
<td>“line/leash”</td>
<td>Schoenveter “shoelace”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Touwtje “piece of cord”</td>
</tr>
</tbody>
</table>

Table 4. Polysemous Austrian German markers, their correct German German synonym and German German nearest neighbour in the word-based semantic space

<table>
<thead>
<tr>
<th>Austrian marker</th>
<th>German synonym</th>
<th>German nearest neighbours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beilage</td>
<td>Anlage</td>
<td>Salat “salad”</td>
</tr>
<tr>
<td>“side dish/enclosure”</td>
<td>“enclosure”</td>
<td>Vorspeise “starter”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Beilage “side dish”</td>
</tr>
<tr>
<td>Trainer</td>
<td>Trainingsanzug</td>
<td>Trainer “trainer”</td>
</tr>
<tr>
<td>“trainer/tracksuit”</td>
<td>“tracksuit”</td>
<td>Coach “coach”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mannschaft “team”</td>
</tr>
</tbody>
</table>
The inability of many distributional models to accommodate all but one frequent meaning reliably is an active topic of research in natural language processing (Reisinger & Mooney 2010). We can thus expect it to improve in the future. Still, this typical characteristic can also be put to use. McCarthy et al. (2004) and Padó & Lapata (2007), for example, have shown that the distributional nearest neighbours to a polysemous target can be used to determine its dominant sense. In the same vein, we can rely on the cross-lectal nearest neighbours of a word to determine its most frequent sense in a specific language variety.

4.4.2 Other semantic relations

A second problem with distributional semantics is that there is no simple way of telling apart different types of semantic relations. As we discussed above, distributional similarity can correspond to a wide spectrum of semantic similarity and relatedness. While our choice of context features influences the distribution of these relations among nearest neighbours, there is no simple trick that will identify one particular type of relation only.

To gain a better insight in this problem, we performed a manual analysis of the nearest neighbours in the Dutch syntax-based semantic space. We restrict this analysis to the nouns in our dataset, as these are most easily grouped into semantic classes. The classes we looked at are synonym, hyponym, hypernym, co-hyponym, related and unrelated. ‘Synonyms’ are correct translations identified by our gold standard, and single neighbours that we deemed correct but that were not included in the gold standard. ‘Hyponym’, ‘hypernym’ and ‘co-hyponym’ are taxonomical relations which indicate that the nearest neighbour is more general than the target (e.g. rendez-vous “rendezvous”; ontmoeting “meeting”), more specific (verdiep “floor”; benedenverdieping “ground floor”), or a member of the same category as the target (pompelmoes “grapefruit”; abrikoos “apricot”). The class “related” brings together all semantically related words that do not belong to any of the previous categories (soeplepel “soup spoon”; oregano “oregano”). Finally, we classify as “unrelated” all nearest neighbours that are not straightforwardly related to their target (zichtkaart “postcard”; speldje “pin”). Orthogonal to those semantic categories, we make a distinction between nearest neighbours that are related to the typically Belgian sense of the target word (“correct sense” in Figure 6), and those that are related to a different sense (“different sense” in Figure 6).

Figure 6 shows the result of this manual analysis. 32% of the single nearest neighbours are synonymous to the intended sense of the target word, thereby constituting the largest class. Only 19% of the nearest neighbours are semantically unrelated to their target word. As we will discuss below, this lack of semantic relatedness is generally due to the low frequency of the target. A majority of all nearest neighbours (64%) is related to the intended Belgian sense of the target
The corpus-based identification of cross-lectal synonyms in pluricentric languages

73

word. Even when they are not synonymous to the target, most nearest neighbours are still taxonomically similar. The relative frequencies of hyponymy, hypernymy and co-hyponymy are greatly influenced by their distribution in the taxonomy: words tend to have most co-hyponyms and fewest hypernyms. A smaller number of nearest neighbours is not taxonomically similar to the target word, but still semantically related. Finally, 17% of the single nearest neighbours are related to a different sense of the target word, with a similar distribution over the six classes. In other words, many nearest neighbours that are normally considered incorrect still make sense semantically. One of the remaining challenges for distributional semantics is therefore classifying these words along their taxonomic or associative relations automatically.

4.4.3 Frequency
A final problem of semantic spaces is that they need a large number of observations to work well. Unfortunately, many typically Belgian and Austrian words are relatively informal, and therefore rarely show up in newspaper corpora. This may result in a low frequency of the word itself, or a low frequency of the typically Belgian or Austrian sense. Infrequent senses will cause nearest neighbours to be related to the more frequent sense (see above). If the word itself is infrequent, the neighbours tend to be entirely unrelated, because the model does not have reliable distributional statistics for these words.

Table 5 gives a few examples of such informal words and meanings in Dutch. Ambetant is a typically Belgian Dutch word for “unpleasant”. Because of its low frequency in the newspaper corpus, the model fails to identify vervelend as a synonym. Plaat and aardig both have an informal meaning in addition to their
standard one. Because this informal meaning is underrepresented in our corpus, it is the standard meaning that dominates the nearest neighbours in the semantic space. Table 6 contains three similar examples for German. Aufreger, Nachzipf and Wunderwuzzi are all labelled as informal words in the Variantenwörterbuch. As a result of their low frequency in the newspaper corpus, their cross-lectal nearest neighbours are only vaguely semantically related, at best.

<table>
<thead>
<tr>
<th>Belgian word</th>
<th>Dutch synonym</th>
<th>Dutch nearest neighbours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plaat</td>
<td>Röntgenfoto</td>
<td>Album “album”</td>
</tr>
<tr>
<td>“record/X-ray”</td>
<td>“X-ray”</td>
<td>Cd “cd”</td>
</tr>
<tr>
<td>“record/X-ray”</td>
<td>“X-ray”</td>
<td>Debutalbum “debut album”</td>
</tr>
<tr>
<td>Ambetant</td>
<td>Vervelend</td>
<td>Soepeltes “smoothly”</td>
</tr>
<tr>
<td>“unpleasant”</td>
<td>“unpleasant”</td>
<td>Onverstandig “foolish”</td>
</tr>
<tr>
<td>“unpleasant”</td>
<td>“unpleasant”</td>
<td>Beresterk “very strong”</td>
</tr>
<tr>
<td>Aardig</td>
<td>Raar</td>
<td>Leuk “nice”</td>
</tr>
<tr>
<td>“nice/odd”</td>
<td>“odd”</td>
<td>Boeiend “interesting”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Geweldig “great”</td>
</tr>
</tbody>
</table>

Table 6. Informal Austrian German markers, their correct German German synonym and nearest neighbours in the word-based semantic space

<table>
<thead>
<tr>
<th>Austrian word</th>
<th>German synonym</th>
<th>German nearest neighbours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aufreger</td>
<td>Skandal</td>
<td>Gestalter “creator”</td>
</tr>
<tr>
<td>“scandal”</td>
<td>“scandal”</td>
<td>Gründervater “creator”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Freigeist “freethinker”</td>
</tr>
<tr>
<td>Nachzipf</td>
<td>Nachprüfung</td>
<td>Lernort “place for study”</td>
</tr>
<tr>
<td>“resit exam”</td>
<td>“resit exam”</td>
<td>Schlafengehen “bedtime”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Gefängnistor “prison gate”</td>
</tr>
<tr>
<td>Wunderwuzzi</td>
<td>Alleskönner</td>
<td>Innovationsschub “innovative drive”</td>
</tr>
<tr>
<td>“jack-of-all-trades”</td>
<td>“jack-of-all-trades”</td>
<td>Mächtmensch “power seeker”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Schallwelle “sound wave”</td>
</tr>
</tbody>
</table>

5. Conclusions

In this article, we have presented a distributional-semantic method for the automatic identification of synonyms across different language varieties. We showed how the meaning of a word can be modelled in terms of its contexts in a large corpus, and how comparable corpora from two language varieties can be used
to build bilectal semantic spaces. We presented the results of two case studies on Dutch and German, arguing that the automatically identified word pairs can be used as input for studies of lexical variation. Finally, our detailed error analysis uncovered three problems that current distributional semantics still struggles with. First, the various senses of polysemous words give rise to a distributional representation that either only reflects the most frequent sense or does not represent any of the senses accurately. Second, distributional semantics does not yet offer a straightforward way of distinguishing between several types of semantic similarity, with synonymy being just one possible relation between nearest neighbours. Finally, the success of the method depends crucially on the choice of corpus, and the frequency of the relevant words in that corpus. The first two of these problems are currently being tackled in computational-linguistic research.

Despite these challenges, it is our belief that the framework of distributional semantics can play a crucial role in the further evolution of corpus-based lexical semantics to a more quantitative discipline. The first reason is that it gives researchers a more empirical way of establishing semantic equivalence that can moreover be based on the corpora they are using for their studies. This can complement the inherently limited capacity of dictionaries, taxonomies and researchers’ own intuition. Because of the imperfect nature of the results, a manual selection of the distributional neighbours may still be advisable, depending on the study at hand. The second reason is its ability to scale up corpus-based lexical-semantic studies, which so far have tended to focus on restricted sets of target words. A broader view of the lexicon will help us broaden our view from single synonym sets to full lexical fields. Ultimately it is even possible to set aside lists of target words and to compute nearest neighbours for all words with a reasonable frequency in a corpus. Even a single computer with several gigabytes of RAM is sufficient to study large portions of the lexicon.

Finally, distributional semantics allows us to address new research questions. With its graded model of semantic relatedness, synonymy is no binary concept. This means degrees of synonymy can be computed and be used, for example, to weight synonymy-based measures of lexical convergence or divergence (Ruette et al. 2014). In this way, the lectometrical approach introduced in Geeraerts et al. (1999) and Speelman et al. (2003) can be scaled up beyond manually determined synonym sets. In addition, it becomes possible to quantify the semantic distance between several language varieties and determine, for example, whether Belgian Dutch is semantically closer to Netherlandic Dutch than Austrian German is to German German. We can moreover study the precise types of contextual features that influence this difference, like specific syntactic constructions, or more semantic and pragmatic factors like connotation and formality. In this way, distributional semantics opens up entirely new areas for study in quantitative lexical semantics.
References


Sagi, E., Kaufmann, S., & Clark, B. (2009). Semantic density analysis: Comparing word meaning across time and phonetic space. In R. Basili & M. Pennacchiotti (Eds.), *Proceedings of


Authors’ addresses

Yves Peirsman
Research group Quantitative Lexicology and Variational Linguistics
University of Leuven
Blijde-Inkomststraat 21 POBox 03308
3000 Leuven
Belgium
yves.peirsman@gmail.com

Dirk Geeraerts
Research group Quantitative Lexicology and Variational Linguistics
University of Leuven
Blijde-Inkomststraat 21 POBox 03308
3000 Leuven
Belgium
dirk.geeraerts@arts.kuleuven.be

Dirk Speelman
Research group Quantitative Lexicology and Variational Linguistics
University of Leuven
Blijde-Inkomststraat 21 POBox 03308
3000 Leuven
Belgium
dirk.speelman@arts.kuleuven.be