Semantic weighting mechanisms in scalable lexical sociolectometry

Tom Ruette, Dirk Geeraerts, Yves Peirsman, Dirk Speelman

1. Introduction

In sociolectometry, the goal is to analyze language varieties or *lects* that represent several sources of linguistic variation. As such, sociolectometry can be contrasted with dialectometry (Goebl 2006, Nerbonne & Kretzschmar 2003) and stylometry (Biber 1995), which only focus on a limited set of lects that represents a single source of linguistic variation, respectively a geographical source and a register source. Dialectometry and stylometry do share with sociolectometry an interest in the aggregate level calculation of linguistic distance between lects. Therefore, these three domains can be captured under the cover term *lectometry*, which has as its typical outcome a map or visualization that reflects linguistic dissimilarities between the lects under analysis. A visual outcome gives us in the case of sociolectometry an insight in the multifactorial nature of linguistic variation.

Although the main focus of a lectometric analysis lies on the production of a visualization of the linguistic distances between lects, which represent different sources of linguistic variation, it is very important to have a thorough understanding of the linguistic variables that are used to measure these distances. In the end, after all, the output is a reflection of the input. In our paper, we will use lexical variables for the distance measurements. This is a deviation from most of the previous lectometric work, where lexical variables were only marginally studied, and where the focus was primarily on phonological and phonetic variables. This is mainly due to the objections of Lavandera (1978) against the study of lexical variation. Her main argument was that it is impossible to set up a *linguistic* variable as soon as the semantics of the variable start to play a role. The first goal of our paper is therefore to acknowledge these objections, and to show how the objections can be addressed.

Next to the semantic problems that we inevitably encounter because of the use of lexical variables, we also run into the problem of *representativity* due to our ambition to analyze the multifactorial nature of linguistic variation. This problem of representativity is less important for dialectometry and stylometry, because their goal is to show *that* there is a regional or register pattern, whereas sociolectometry, in contrast, looks at multiple sources of variation and wants to show *what* the actual patterns are. This is important, because it implies that we can not start with a variable set that is unnaturally biased towards one source of variation. Instead, we need a variable set that is representative of the multifactorial linguistic variation. Admittedly, our focus on lexical variables introduces a second type of bias, because sociolectal or functional variation is already proven to occur with other types of variables, as well. Interestingly enough, to overcome the first type of bias towards a specific dimension of variation, previous sociolectometric work (Geeraerts et al. 1999, Soares de Silva 2010) increased the second type of bias. This was achieved by zooming in on a limited amount of lexical fields, so that a set of lexical variables that covers the entirety of the lexical fields could be generated manually. As the set is exhaustive for the lexical field under scrutiny, the first type of bias is undone and we

can say that the variable set is representative of the multifactorial variation in that lexical field. This approach, however, required an immense amount of manual labor, which necessarily restricted these first sociolectometric studies to a small scale. Moreover, this almost extreme bias of the second type limits the generalizability of these studies massively. Therefore, the second goal of our paper is to introduce an automatic methodology that can generate many lexical variables that cover a representative part of the vocabulary.

The paper is structured as follows: in Section 2, we discuss the theoretical and methodological side of aggregating lexical variables. Section 3 zooms in on the automatic methodology for generating lexical variables and our attempts to answer to the objections of Lavandera (1978). Our methodological and theoretical proposals are applied to a case study in Section 4. The case study looks at the multidimensional structure of the Dutch vocabulary, by measuring the distances between two national varieties of Dutch and several registers. In Section 5, we conclude the paper by situating the field of lexical sociolectometry in the area of this volume and by summing up the most important findings.

2. Aggregating lexical variables

Since Labov (1972: 271), a sociolinguistic variable has been defined so that the variants of the variable say or do the same thing (cf. Chambers and Trudgill 1980: 50), and only differ in their socio-stylistic distribution. Although the sociolinguistic variable was predominantly used for phonological and phonetic variation, its range was quickly extended to non-phonological variation, as well. For the lexicon, this implies that a sociolinguistic lexical variable consists of words that mean the same thing, but are used by different people in different circumstances. However, this practice attracted criticism:

"it is inadequate at the current state of sociolinguistic research to extend to other levels of analysis of variation the notion of sociolinguistic variable originally developed on the basis of phonological data. The quantitative studies of variation which deal with morphological, syntactic, and lexical alternation suffer from the lack of an articulated theory of meanings." (Lavandera 1978: 171)

Lavandera's critique boils down to the fundamental question of "what is meaning?". The preference for phonological and phonetic variation in sociolinguistics, even today, is obviously caused by the apparent lack of meaning of the variable, and thus effectively immune to the Lavandera criticism. However, a recent article of Campbell-Kibler (2011: 435) even questions the lack of meaning of morphophonological variables, by pointing out that the individual variants of an assumed variable carry distinct social meanings. If we zoom in on the problem of sociolinguistic lexical variables, we bump into issues with word-meaning. Word-meaning in the late 70s was mainly considered to be restricted to referential meaning, but soon, the unwieldy problem of (lexical) semantics would be under full attention (e.g. Cruse 1986, Taylor 1989), leading to the linguistic war between Generative and Interpretative Semantics. In the 80s, under the influence of the upcoming Cognitive Linguistic paradigm (for an overview, see Geeraerts and Cuyckens 2007), word-meaning was seen as a multidimensional object, without a distinction

between linguistic knowledge and encyclopedic knowledge (Taylor, 1989: Chapter 5).

The construction of sociolinguistic lexical variables or synonyms now becomes even more problematic: are two words only synonymous if they are identical for all the dimensions of meaning? The existence of a pair of words that complies with that requirement is highly unlikely. However, this does not rule out the study of lexical variation in the form of near-synonyms, which turns the concept of synonymy into something gradual. Near-synonyms are prototypically structured conceptual categories , with gradable membership along many dimensions, e.g. a referential, expressive, social or stylistic dimension. To study near-synonyms, we have to extend the sociolinguistic perspective, and take the possible variation on the other meaning dimensions of the lexical variable into account, as well. We are therefore not studying just sociolinguistic lexical variables, but lexical variables in general, with all their variational – including semantic – dimensions. It seems that a marriage of lexical semantics and variational linguistics needs to be announced. This point-of-view is an extension of the stance of Edmonds & Hirst (2002: Section 2.5), and is articulated in the field of Cognitive Sociolinguistics (Kristiansen & Dirven 2008, Geeraerts et al. 2010)

Next to these issues at the level of the individual variable, the current study will also run into issues because we *aggregate* lexical variables. This means that the variational patterns of the individual lexical variables are averaged out in order to get a grip on the more general variational patterns that play in the vocabulary. Such an aggregated perspective has as its main disadvantage the loss of detail and the destruction of subtle semantic differences. However, this disadvantage is the price that needs to be payed for getting an overarching, bird-eye's view on variation in the vocabulary. Another problem is more pressing: as every individual lexical variable carries a certain concept (the shared meaning of all the variants in the variable), the aggregation of lexical variables introduces the dimension of *conceptual variation*. Conceptual variation is the fact that some concepts are more frequent than others, that more frequent concepts might be more salient, and that these conceptual frequencies may differ per lect. A sociolectometric study has the ambition to take all these aspects and issues into account.

The first sociolectometric study of lexical variation was Geeraerts et al. (1999). In that monograph, a similarity metric *U'* was introduced which aggregated lexical variables (so-called *profiles*) to measure lexical similarity of varieties of Dutch. The similarity metric *U'* was transformed into a distance metric by Speelman et al. (2003: Section 2.2 and 2.3). As we will use this metric for the current paper, as well, we will introduce the metric here.

Given two subcorpora V_i and V_j that represent the varieties under scrutiny, a profile L and x_1 to x_n (the exhaustive list of variants in the profile L), then we refer to the absolute frequency F of the usage of x_i for L in V_j with:

$$F_{V_j,L}(x_i) \tag{1}$$

Subsequently, we introduce the relative frequency R of x_i, part of profile L, in V_j:

$$R_{V_j,L}(x_i) = \frac{F_{V_j,L}(x_i)}{\sum_{k=1}^{n} F_{V_j,L}} \tag{2}$$

Now we can define the City-Block distance V_1 between and V_2 on the basis of L as follows (the division by two is for normalization, mapping the results to the interval [0,1]):

$$D_{CB,L}(V_1, V_2) = \frac{1}{2}\sum_{i=1}^{n} \left| R_{V_1,L}(x_i) - R_{V_2,L}(x_i) \right| \quad (3)$$

The City-Block distance is a straightforward descriptive dissimilarity measure that assumes the absolute frequencies in the sample-based profile to be large enough for the relative frequencies to be good estimates for the relative frequencies in the underlying population-based profiles. As such, the City-Block distance accounts for lexical variation, by taking the frequencies of the variants relative to the frequency of the underlying concept. Sometimes, however, these relative frequencies might give a wrong impression of the actual linguistic distance, e.g. when the samples are rather small and the relative frequencies are unreliable. Therefore, supplementary control is needed. To verify that there actually is a difference between the two profiles, we use the Log Likelihood Ratio Test (LLR) (Dunning 1993). This time, we look at the absolute frequencies in the profiles that are compared. When we compare a profile in one language variety to the profile for the same concept in a second language variety, we use LLR to test the hypothesis that both samples are drawn from the same population. On the basis of this log likelihood statistic a *p*-value can be calculated. If this *p*-value is larger than 0.05 (no significant difference between the samples), we set the dissimilarity between subcorpora at zero. If the *p*-value is smaller than 0.05, we use D_{CB}.

To calculate the dissimilarity between subcorpora on the basis of many profiles, we just sum the dissimilarities for the individual profiles. In other words, given a set of profiles L_1 to L_m, then the global dissimilarity D between two subcorpora V_1 and V_2 on the basis of L_1 up to L_m can be calculated as:

$$D_{CB}(V_1, V_2) = \sum_{i=1}^{m} \left(D_{L_i}(V_1, V_2) W(L_i) \right) \quad (4)$$

The W in the formula is a weighting factor. We use weights to ensure that concepts which have a relatively higher frequency (the sum of the frequencies of all variants, relative to the amount of words in the two subcorpora that are being compared) also have a greater impact on the distance measurement. In other words, in the case of a weighted calculation, concepts that are more common in everyday life and language (as represented in the corpus) are treated as more important. As such, this distance metric takes the conceptual dimension of aggregating lexical variables into account.

On a terminological note, because the distance metric captures the preference for naming a concept with a certain variant, we call it the *onomasiological* metric. The onomasiological metric has two levels: the first level takes onomasiological variation *within the variable* into account (Formula 3), and the second level takes onomasiological variation *across the variables* into account (Formula 4). This distinction will recur below.

3. Automatic variable set generation

A good variable set contains “a large number of variables [… with …] a great deal

of variation irrelevant to questions of geographic or social conditioning [and] will […] provide the most accurate picture of the relations among the varieties examined" (Nerbonne 2006: 464). There have been attempts to generate such variable sets in a "top-down" fashion, by drawing from the relevant literature, cf. Biber (1988) or Szmrecsanyi (this volume). However, the resulting variables are arguably somewhat biased towards the interest of the variationists and dialectologists that created the relevant literature in the first place. Therefore, we propose a "bottom-up" method that — specifically for the lexical focus of our study — generates candidate lexical variables on the basis of a recent methodological advance in Computational Linguistics, called *Clustering by Committee* (Pantel 2003). This rather complex method will be explained in Section 3.1.

The sociolectometric emphasis on a non-biased variable set — in the sense that it does not only contain features with a regional pattern (cf. dialectology) or a stylistic pattern (cf. stylometry) — started in Geeraerts et al. (1999) and is motivated by the research goal of *discovering* lectal differences. This goal is different from typical dialectometric studies, whose goal it is to *show that* there are regional differences, or stylometry, where the goal is to *show that* there are stylistic differences. Therefore, it is acceptable to start from regionally or stylistically patterned variable sets. For sociolectometry, which focuses on interactions between a range of lectal dimensions, a variable set with an a priori distribution is unacceptable.

Generating a truly unbiased variable set is probably impossible. Nonetheless, previous (lexical) sociolectometric studies attempted to make variable sets that are at least as unbiased as possible. With that goal in mind, instead of trying to describe variation in the whole lexicon, Geeraerts et al. (1999) and Soares da Silva (2010) zoom in two lexical fields, and try to come up with all the concepts that are relevant for these specific lexical fields. This is a feasible, yet labor-intensive task which appears to give trustworthy results, yet the results can not simply be extrapolated to the whole of the lexicon – which is in fact another kind of bias. Moreover, this approach is not scalable: to cover the whole lexicon, a very large number of lexical fields would be necessary, and the manual description of variation for concepts relative to these lexical fields would take forever. Therefore, we use the clustering algorithm Clustering by Committee as an automatic and scalable bottom-up approach for generating an unbiased variable set for the lexicon.

The task that we give to Clustering by Committee is basically to identify near-synonyms. This task falls apart in two subtasks. First, there must be a way of measuring semantic similarity between words to assess the degree of synonymy, and second, there must be a way of generating clusters that contain near-synonyms. Both subtasks are covered by Clustering by Committee. We will deal with two aspects of Clustering by Committee. On the one hand, it generates clusters of highly similar words, so-called committees. On the other hand, the returned committees do not necessarily contain acceptable near-synonyms. These two aspects will be dealt with after a non-technical introduction to the algorithm below.

3.1. Clustering by Committee

Before we deal with these two characteristics in more detail, we first explain the Clustering by Committee algorithm in a way that is as intuitive as possible. We sacrifice some technical precision, and trade it for accessibility. However, skipping this section

does not make the remainder of the paper unreadable.

The Clustering by Committee algorithm is rooted in American neo-structuralist Distributional Semantics (for a more elaborate overview, see Geeraerts 2010: 173–178). The belief of (this type of) Distributional Semantics is that the meaning of a word can be indirectly measured by describing the context in which that word appears. Philosophically speaking, the analytic *Meaning is Use* idea of Wittgenstein (1953/2001) is taken to its maximum by the well-known quote of Firth (1957): "you shall know a word by the company it keeps". Practically speaking, Distributional Semantics uses a *Semantic Vector Space* model (Baeza-Yates and Ribeiro-Neto 1999: 25) as a proxy for describing meaning (Turney and Pantel 2010). In what follows, we first describe how Semantic Vector Space models and Distributional Semantics work together. Then, we go through the iterative steps of the Clustering by Committee algorithm.

In Semantic Vector Space models, objects are described by *n* quantifiable characteristics. These characteristics make up an *n*-dimensional space in which the objects can be positioned. Every characteristic is thus a dimension. The position of the objects on these dimensions depends on the value that the characteristics have. In a way, these values can be seen as coordinates of a point in the *n*-dimensional space, made up by the characteristics. The values of a single point are stored in a so-called vector. Every vector then represents the object that is described by its characteristics. The spatial idea that underlies the Semantic Vector Space models does not restrict the objects to tangible items. Indeed, in Distributional Semantics, word meanings are the objects, and the characteristics are contexts in which these words appear.

Let us pay some more attention to these contexts. Words-in-context are to be found in large text corpora. In a so-called *bag-of-words* model, the contexts are merely words that appear left and right from the lemma *w* that we want to describe, as found in all the texts of the corpus. The values of the contextual characteristics are then the frequencies (or derived statistical measures) with which *w* and these contexts co-occur. Of course, these contexts do not need to be limited to neighboring words. In previous work (Peirsman et al. 2007), it was found that syntactic dependency triples are more suited as contexts than the bag-of-words when the goal is to find synonyms. A syntactic dependency triple is e.g. "*w* appears as the subject of verb *v*", or "*w* is a modifier of noun *n*".

If we now take two words, represented by a vector containing the co-occurrence values for many contexts, we can quantify their semantic similarity by applying the spatial idea that underlies the Semantic Vector Space model: if two objects are very close to each other in the *n*-dimensional Semantic Vector Space, then they are bound to have very similar values on a number of dimensions. If two objects behave alike for a large number of characteristics, represented by the dimensions, they must be very similar to each other, with respect to these dimensions. Given that we assume that the dimensions in a Semantic Vector Space with words represent the Distributional Semantics of a lemma, "spatial" closeness of two words stands for semantic similarity between these words. Without going into detail about the metric for Semantic Vector Space closeness (or semantic similarity), we merely mention that spatial closeness of two objects, represented by vectors, is measured by means of the cosine metric (Baeza-Yates and Ribeiro-Neto 1999: 27).

With nothing more than these two building blocks — representation of semantics with the Semantic Vector Space Model, and semantic similarity quantification — the Clustering by Committee algorithm sets out to solve the problem of word sense discovery, which might seem — but isn't — much different from our goal (finding near-synonyms).

The Clustering by Committee algorithm consists of three phases, and the second phase has "finding tight clusters of semantically similar words" as its goal. This is very much what we want to obtain. The tight clusters of semantically similar words are called *committees*. Before we get to that second phase of committee generation, we will explain phase one of the Clustering by Committee algorithm.

In phase one of the Clustering by Committee algorithm, the task is to compute pair-wise similarities between the individual word-types in the corpus. For practical reasons, only a subset of all the individual words is considered. However, Pantel and Lin (2002: Section 4.1) claim that taking a subset does not influence the results too much. The pair-wise similarities between words, measured with the cosine metric, are stored in a similarity matrix *S*, which is the input for the second phase of the Clustering by Committee algorithm.

In phase two of the Clustering by Committee algorithm, the committees are generated through a number of recursive steps. In each recursive step, the algorithm goes through every element *e* in the similarity matrix *S*, and looks for a small set of tight clusters on the basis of the similarity between *e* and the other elements. The retrieved clusters are called committees. The algorithm sorts these committees on the basis of their semantic tightness, measured by means of the average semantic similarity between all the words in the committee. Then, it identifies residue words that are not covered by any committee, by presenting it to each of the already existing committees (tightest committees first). A committee covers a word if the word's similarity to the centroid of the committee exceeds some similarity threshold. The centroid of the committee is (something like) the average of all the vectors from the words that are already in the committee. The algorithm then recursively attempts to find more committees among the residue words. The output of the algorithm is the union of all committees found in each recursive step. This simplified description can be found with more details in Pantel and Lin (2002: Section 4.2 and Figure 1). At the end of phase two, there is a list of committees, which will be used in our study as candidate lexical alternation variables.

Although the Clustering by Committee algorithm goes further with a third phase to discover word senses, we stop our description of the algorithm here, because it is not relevant for our purpose. We refer the interested reader to Pantel and Lin (2002) and Pantel (2003) for further information.

3.2. Committees and synonymy as a gradual category

First, the committees contain words that are semantically speaking very similar to each other, but their lexical relation is not necessarily synonymy. The committees are rated with a "score", which is the average similarity of the words in the committee. The higher the score, the more similar the words in the cluster are, and the bigger the chance that these words are actually near-synonyms. We will use the words in the committees as the variants of a lexical variable.

Second, perhaps the most imminent problem with the Clustering by Committee algorithm is that the results can not (yet) be trusted blindly. The returned clusters contain impurities, in the sense that sometimes – or actually most of the time – the returned committees are not clusters of near-synonymous words, but merely of related words. Therefore, we will regard the outcome of the Clustering by Committee algorithm as a list

of candidate variables, which needs to be filtered out manually.

In addition to this manual filtering and our first onomasiological metric (cf. Section 2, the City-Block distance metric with a conceptual weighting term *W*), we also introduce a second metric. Whereas the first onomasiological metric took care of naming preferences and conceptual weight, the second metric will focus on the *semasiological* structure, i.e. the semantic characteristics of the variables, again within and across the variables.

First, to have semasiological control *within the variable*, we want to weigh variants that are more similar to the meaning of the concept, expressed by the variable, more than variants that are not so similar to the meaning of the concept. More intuitively, if a variable has three variants, and the third variant is less synonymous than the other two, we will weigh down the influence of the third variant in the City-Block distance metric. The meaning of the concept is modeled by means of the centroid of the committee, which is calculated in the Clustering by Committee algorithm, as explained in Section 3.1. The (normalized) similarity $I_L(x_i)$ of a variant in variable L to the underlying concept meaning of L is calculated as in Equation 5, where d is a distance function. We can now plug this semasiological intra-variable weight into Equation 3 by using Equation 6. The division by the maximum distance between V_1 and V_2 for the current variable L maps the $D_{CB,L}$ between [0,1], cf. the division by 2 in Equation 3.

$$I_L(x_i) = \frac{1 - d(x_i, centroid_L)}{\sum_{j=1}^{n} 1 - d(x_j, centroid_L)} \tag{5}$$

$$D_{CB,L}(V_1, V_2) = \frac{1}{maxdist(L, V_1, V_2)} \times \sum_{i=1}^{n} \left| R_{V_1,L}(x_i) - R_{V_2,L}(x_i) \right| \times I_L(x_i) \tag{6}$$

Second, to have semasiological control *across the variables*, we want to weigh variables that are semantically speaking "tighter" more than variables that are "sloppier". The rationale is that tighter variables will contain words that are overall semantically more similar to each other, and might therefore be more true to the classic sociolinguistic alternation variable. Equation 7 normalizes the tightness measure (score) that already comes out of the Clustering by Committee algorithm, and Equation 8 plugs it into Equation 4. The multiplication with the first term maps the final distance to the interval [0,1].

$$S(L_i) = \frac{tightness(L_i)}{\sum_{j=1}^{n} tightness(L_j)} \tag{7}$$

$$D(V_1, V_2) = \sum_{i=1}^{m} (W(L_i) S(L_i)) \times \sum_{i=1}^{m} (D_{L_i}(V_1, V_2) W(L_i) S(L_i)) \tag{8}$$

For the current study, the normalization functions proposed in Equations 5 and 7 are linear functions. There is reason to believe that a linear function might not be adequate here (Kretzschmar, this volume). This issue will be addressed in further research.

4. Case study

The above metrics are operationalized in the following case study on lexical distances between registers of Belgian and Netherlandic Dutch. The case study wants to show two things: (1) the influence and importance of controlling onomasiological and semasiological variation when measuring distances between varieties, and (2) an insight in the structure of Dutch language varieties on the basis of a large bottom-up sample of lexical variation. All in all, studying both register and national variation at the same time shows the importance of a semantically controlled approach and a methodology that grasps the multivariate character of lexical variation. Finally, the use of automatically generated lexical variables (henceforth, *profiles*) ensures generalizable results.

4.1. Research question

The diaglossia idea of Auer (2005), where there are intermediate variants between Standard Variety and (base) dialect, can be applied to a pluricentric language such as Dutch, as used in Belgium and the Netherlands. Just as Auer (2005) notes, the actual situation in the Dutch area is that the Netherlandic Dutch Standard Variety is slightly different from the one used in Belgium, and this Standard Variety has diverged from it over the last decades in phonology and phonetics, but not in the vocabulary (Geeraerts et al. 1999). Moreover, the diaglossia idea predicts that these patterns of convergence or divergence play differently in different registers. Our research question is therefore whether we can observe a multidimensional structure of varieties in actual data by means of the sociolectometric methodology.

The difference at the level of the Standard Variety between Belgian Dutch and Netherlandic Dutch has received quite a bit of attention of mostly Belgian linguists. The following linguists discussed the standard language: Geeraerts (2002), Geerts (1989), Jaspers and Brisard (2006), Schutter (1998), Stroop (1990, 1992), Taeldeman (1991), Willemyns (2007). The following linguists discussed the emergence of an “in-between” language, that sits between regiolects and standard: De Caluwe (2002), Geeraerts (1993), Goossens (2000), Plevoets (2008), Taeldeman (1992), Willemyns (2005). Rather than giving an overview of the individual views and proposed language policies of the aforementioned linguists, we can summarize the converging findings on relevant, objective linguistic matters briefly.

Whereas a Dutch Standard Variety was developed naturally in The Netherlands during the 17th century, Flanders was politically separated from The Netherlands and French was the language of government and high culture. Dutch in Belgium survived in the dialects of the rural Flemish villages, where there was no need to develop a Standard Variety due to limited mobility. After World War II, the upcoming of Dutch in Flanders gained ground during the 50s and 60s, with a climax in 1968. During the 50s and 60s, an official linguistic policy was put in place, so that Belgian Dutch would be normatively dependent on Netherlandic Dutch. Despite large-scale efforts of radio and television, the Belgian Dutch Standard Variety evolved to be somewhat different from Netherlandic Dutch (also because Netherlandic Dutch kept evolving naturally, whereas the norm for

Belgium remained 1950s Netherlandic Dutch). Nowadays, a clear Standard Variety of Belgian Dutch exists in the language of politicians and journalists. In the nineties and before, The Netherlands had a diaglossic (Auer 2005: Section 5) linguistic situation, whereas Flanders was in a diglossic (Auer 2005: Section 4) situation. Only recently, the Flemish situation evolved into a diaglossic situation with the upcoming of Colloquial Belgian Dutch, filling the gap between the standard variety and the dialects.

The diaglossic situation in Flanders is described as being different from the diaglossic situation in The Netherlands. In The Netherlands, the diaglossic spectrum is more limited than in Belgium, and the distance between the highest and lowest variety is smaller, as well. Moreover, in Belgium, the Belgian Standard Variety is not so often applied because there are fewer situations in which the use of that variety is deemed to be appropriate. Whereas Netherlandic speakers might stick to their Standard Variety in slightly less formal situations, Belgian speakers abandon the Belgian Standard Variety much more easily and switch quickly to Colloquial Belgian Dutch. Finally, there is a linguistic gap between the Belgian Standard Variety and Colloquial Belgian Dutch, in contrast to a more gradual differrence between the Netherlandic Standard Variety and Colloqiual Netherlandic Dutch. On the theoretical part of this paper, our goal is to find these characteristics of Dutch in the visualizations of the measured lexical distances.

We now move on to our (methodological) research question: how can we modify the sociolectometric methodology (Geeraerts et al. 1999; Speelman et al. 2003; Soares da Silva 2010) so that it is based on a large variable set, and therefore more generalizable? As explained in detail above (Section 3), Semantic Vector Space models are useful here. The remainder of this case study is structured as follows. First, we will introduce the corpus material in which we are going to look for both national and register patterns (Section 4.2). Next, we present the features that come out of the Clustering by Committee algorithm (Section 4.3). In the third part, an overview of the results of the different weighting approaches is given (Section 4.4). And the final part brings these results together in a discussion (Section 4.5).

4.2. Corpus

The corpus used for the present case study is a sample from a combination of corpora, so that it covers both Belgian and Netherlandic Dutch, as well as several registers. All data in the corpus was recorded or written during the period 1999–2004, which we will regard as a synchronic period. As such, there is no need for a diachronic dimension in the study. The corpora were automatically lemmatized and annotated for part-of-speech with Alpino (Bouma et al. 2001). The corpus consists of five registers.

1. Spontaneous conversations: the *Corpus Gesproken Nederlands* (CGN, Corpus of Spoken Dutch, Taalunie (1998–2004)) contains transcriptions of spontaneous conversations, recorded during telephone calls or during face-to-face interaction. These recordings were made between 1999 and 2004. Abbreviation in the visualizations: *sponcon*.
2. Usenet: we downloaded two Usenet discussion topics from the Google Groups Usenet archive for the period 1999–2004. The two topics are politics and radio. Abbreviation in the visualizations: *usenet*.

3. Popular newspaper articles: both Belgium (University of Leuven) and the Netherlands (University of Twente) have collected all the articles from a number of newspapers during 1999–2004. The University of Leuven collected the Belgian newspapers, and the University of Twente did the same for Dutch newspapers. From these newspapers, we selected the ones that are deemed "popular" newspapers. Abbreviation in the visualizations: *popnp*.
4. Quality newspaper articles: from the same collection of newspapers as in the popular newspaper articles part, we selected the ones that are deemed "quality" newspapers. Abbreviation in the visualizations: *quanp*.
5. Legalese: we downloaded the entire collection of official government announcements in *Staatsblad* for the period 1999–2004 from both the Netherlands and Belgium. Abbreviation in the visualizations: *staatsblad*.

Because we have a Belgian and a Netherlandic part for every register, the abbreviated names are concatenated with *-be* or *-nl*, e.g. *sponcon-be* for spontaneous conversations in Belgium. Also, because a certain part of the methodology (the LLR measure, explained above) requires that the subcorpora are more or less equal in size, we divided all register and country combinations into equally sized fragments. The smallest subcorpus in our set of (2 (countries) x 5 (registers) =) 10 subcorpora contained 2.5 million words. Therefore, we divided all the subcorpora in fragments of 2.5 million words. Since the other subcorpora are much bigger, we randomly sampled Usenet posts, newspaper articles and official announcements so that each subcorpus (except the spontaneous conversations) would consist of 3-5 fragments. To identify the fragments, we concatenated the abbreviations with an index, e.g. *sponcon-be-0* for the single fragment of the subcorpus with spontaneous conversations in Belgium. The corpus consists of 31 fragments of about 2.5 million words each, in sum almost 80 million words.

4.3. Features

The lexical variables or *profiles* for this study are, as mentioned above, automatically generated on the basis of the Clustering by Committee algorithm, applied to data from a large newspaper corpus, i.e. the quality and popular newspapers from Belgium and the Netherlands, described above. For every word type with part-of-speech "noun" in the corpus, a (semantic) vector is construed on the basis of relatively basic syntactic information, as explained in Section 3.1. Our restriction to nouns is a limitation of the underlying semantic representation of words in the Semantic Vector Spaces: although the Distributional Semantic representation works for all part-of-speech classes, its efficiency and accuracy is the highest for nouns. Then, we applied the Clustering by Committee algorithm with the following parameter settings: first, we clustered only the 100 most similar words for every noun; second, instead of searching for large committees, we let the algorithm favor smaller clusters, because this increases the chance of finding committees that only contain synonymous words. This approach yields 2019 committees of usually 2 or 3 words. We provide some successful examples from the output of the Clustering by Committee algorithm in Table 1.

Score	Descriptors (concept)
0.47	Wijze, manier (MANNER)
0.45	Volkerenmoord, genocide (GENOCIDE)
...	...
0.10	Omloop, circuit (CIRCUIT)
0.10	Anarchie, onlust (ANARCHY)

Table 1. Successful examples from the output of CBC

Because the algorithm does not return perfect synonyms, we performed a manual clean-up, by removing committees in which a word occurs that is merely an association, rather than a near-synonym. Note that we did not remove words from a committee, but removed imperfect committees completely. From the 2019 committees, about 600 remained. We employed a number of operational rules of thumb to prune committees:

1. "obvious" near-synonyms as they might appear in a thesaurus are not pruned
2. if the committee consists of the male and female variant of the concept, e.g. *verple(e)g(st)er* "male or female nurse", we do not prune that committee from the automatically generated list
3. Sometimes, variants are clearly very much related (meronymy) and borderline synonymous. Here, we adopt the following rule-of-thumb: if the context in which all the words of the committee are not near-synonymous is extremely constructed, the committee is not pruned
4. all other committees are pruned

Admittedly, this manual selection of the features undoes the completely automatic ambition. Nonetheless, even with this manual selection, the amount of lexical variables that were aggregated in this corpus-based study is still larger than any previous corpus-based study that we know. Moreover, one could see the Clustering by Committee approach as a semi-automatic way of overgenerating a large list of candidate variables, from which a researcher can sample the lexical variables that are appropriate for his goals. A dialectologist would pick regionally distributed variables to zoom in on the regional pattern of the varieties, whereas the sociolectometricist would try to account naturally for as many dimensions of lectal variation as possible in the variable set.

Although the sample of variables that we get via the Clustering by Committee algorithm does not cover by far the complete vocabulary, we strongly feel that a lexical study with so many variables already is more generalizable than a (more in-depth) study of two lexical domains (e.g. Geeraerts et al. 1999; Soares da Silva 2010). To refine this scalable approach, we should make further efforts to improve both recall and precision of the Clustering by Committee algorithm.

4.4. Weighting

Now that we have a large set of lexical variables or *profiles*, we can move on to measuring the actual linguistic distance between the fragments of the subcorpora. The

starting point of our distance metric is minutely described in Speelman et al. (2003) and already overviewed above, in Section 2. Subsequently, the outcome of this distance metric is used as input for Kruskal's non-metric Multidimensional Scaling (Cox and Cox 2001), to generate a two-dimensional approximation of the original distance matrix that can be visualized. In our study, all counting of frequency and measuring distances was implemented in Python; the non-metric Multidimensional Scaling is performed in R by using `isoMDS` in the `MASS` package.

Because we are interested in the influence of semantic control on the lexical distance measurements, we will perform three analyses. First, we only account for the intra-profile onomasiological variation (Section 4.4.1). In other words, all profiles will be deemed equally important, and we restrict attention to the relative frequency of the preference for one word over other synonymous words. Second, we account for intra- and inter-profile onomasiological variation by also incorporating the conceptual weight of the different profiles (Section 4.4.2). More frequent profiles will then weigh in more on the final lexical distance. Third, we add semasiological knowledge to the distance metric by letting more central variants be more influential, and by letting more cohesive profiles weigh more on the distance metric (Section 4.4.3).

4.4.1. Only naming preference

The advantages of the naming preference weight over a non-weighted (keyword-based or profile-less) approach have been laid out in Speelman et al. (2003). In essence, this naming preference only approach is identical to the *U* measure of Geeraerts et al. (1999) and Soares da Silva (2010). We set the *W* term of Equation 4 to the constant 1/*m*, where *m* is the amount of profiles in the study. Basically, this is an average, which will keep the resulting distance within [0,1], which allows for easier comparison.

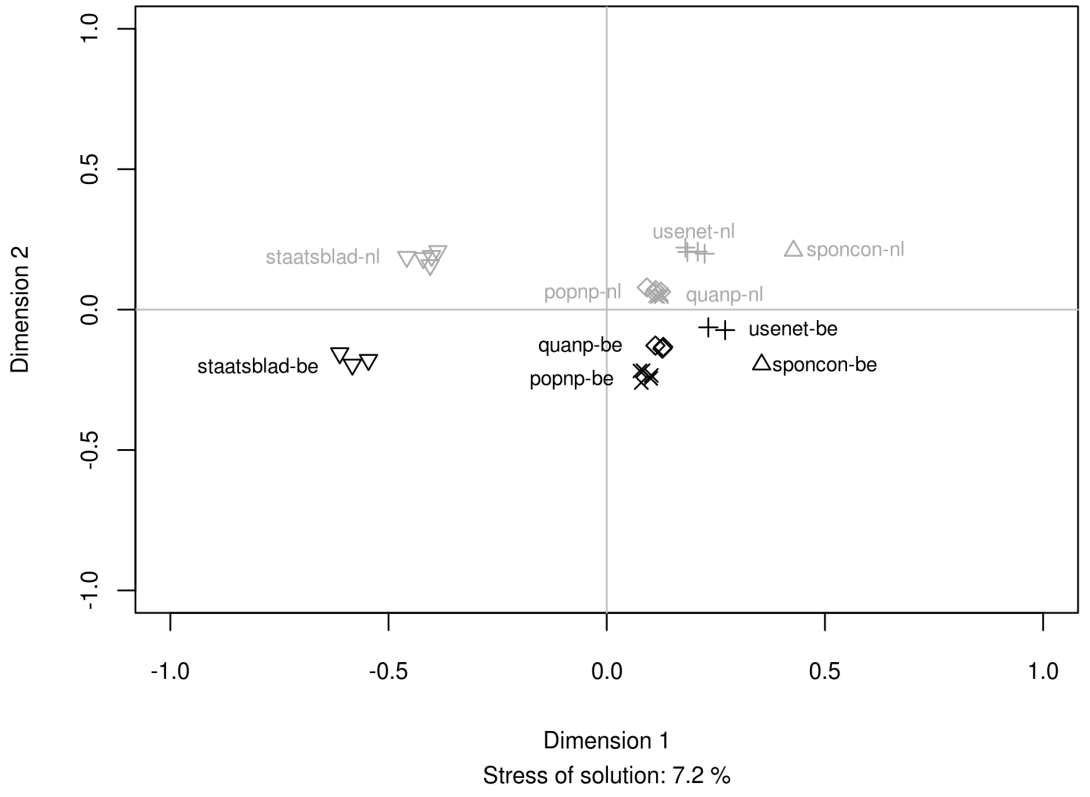

Figure 1. Naming preference weighting. Dimension 1 reveals a register cline, Dimension 2 separates the national subcorpora.

Figure 1 shows the position of the fragments in a two-dimensional space, as calculated with only the naming preference weight. To keep the visualization readable, not every subcorpus has received a label. Instead, we labeled the (obvious) clusters of subcorpora, and assigned symbols and gray values to subcorpora of the same type: a down-pointing triangle stands for legalese (*staatsblad*), a multiplication sign stands for popular newspapers (*popnp*), a diamond stands for quality newspapers (*quanp*), a plus sign stands for Usenet (*usenet*), and an up-pointing triangle stands for spontaneous conversations (*sponcon*); black symbols stand for Belgium (*be*), and gray symbols stand for the Netherlands (*nl*) The gray horizontal and vertical lines indicate the position of zero on the axis, which can be interpreted as a border for categorization. An interpretation of this visualization is appropriate because the stress value — an indication of how much information is lost due to the dimension reduction — of 7.2% is acceptable. Moreover, the two dimensions of the visualization are readily interpretable. The first dimension (left to right) is a clear register cline from legalese, over newspapers and Usenet, to spontaneous conversations. The grey zero line indicates that the distinction between legalese and the other registers is most present in the data, and not the distinction between written and spoken registers, as could be hypothesized. This is not surprising, given the lexical input: legalese stands out for its terminology. The second dimension (bottom to top) distinguishes the two national varieties: Belgian Dutch subcorpora are in the lower part,

and Netherlandic Dutch subcorpora are in the upper part of the visualization. It is noteworthy that the two national varieties are not entirely symmetric. As an example, the two types of newspapers, quality versus popular, are clearly separated in the Belgian part, but lumped together in the Netherlandic part.

4.4.2. Onomasiological weighting

If we now again add the conceptual weight to the formula, by changing back the *W* term to the relative frequency of the profile in the subcorpora, we arrive at the *U'* metric, introduced in Geeraerts et al. (1999) and applied in Soares da Silva (2010). One would expect that for the visualization of the distances between subcorpora, conceptual differences and similarities are emphasized. Practically, we hypothesize that, on the one hand, a cluster of fragments from a single variety will become tighter because of internal conceptual consistency, and that, on the other hand, the distances between the subcorpora will become bigger on the register dimension for the same reason.

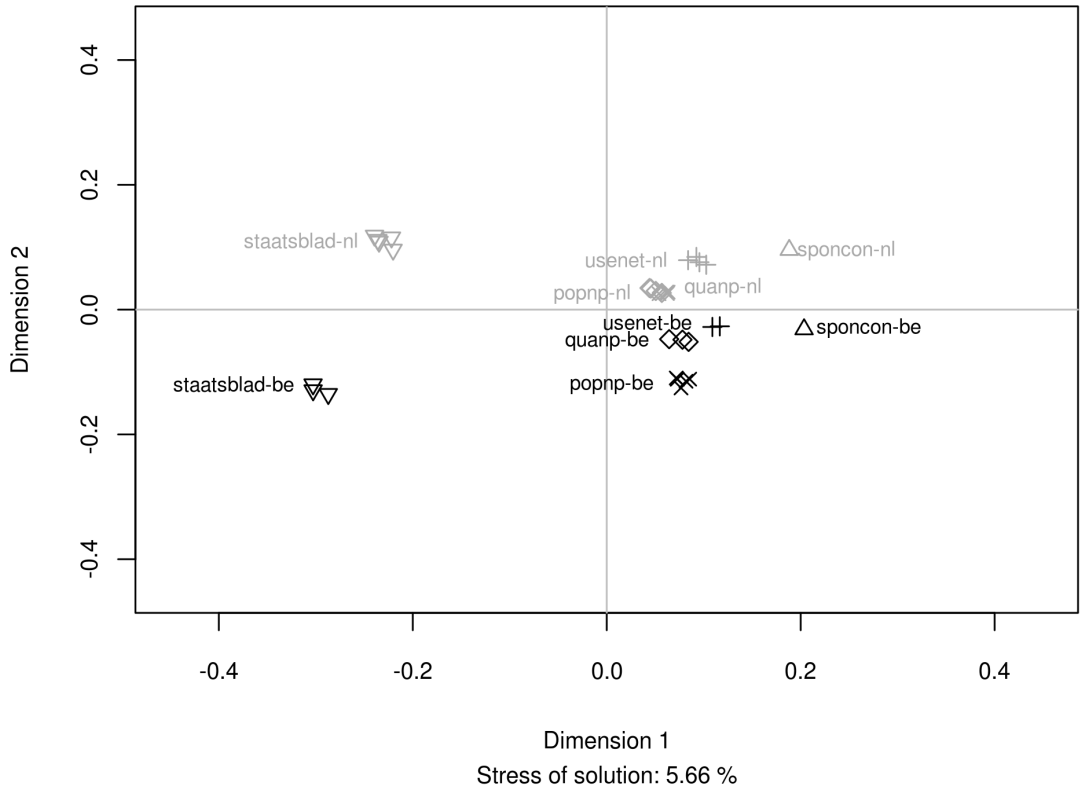

Figure 2. Onomasiological weighting. Dimension 1 reveals a register cline, Dimension 2 separates the national subcorpora. Subcorpus clusters have become tighter.

Figure 2 shows the position of the fragments in a two dimensional space, as calculated with the two onomasiological weights (naming preference and conceptual weight). We notice a decrease in stress value (5.66% versus 7.2% in Figure 1). At first

sight, Figure 2 is very similar to Figure 1. This allows for exactly the same dimensional interpretation of Figure 2 as for Figure 1. Dimension 1 is a register cline, and dimension 2 distinguishes the national varieties. However, small differences can be observed. First, we observe indeed that the register dimension is stretched, in comparison to Figure 1. The spontaneous conversations are pulled out a little bit, away from the newspapers, and the Usenet subcorpora are pushed towards the newspapers, in comparison to Figure 1. Second, it is clear that the clusters of the fragments have become tighter, as expected above.

4.4.3. Onomasiological and semasiological weighting

Finally, we add the semasiological weight for intra- and inter-profile weighting by using the equations 6 and 8.

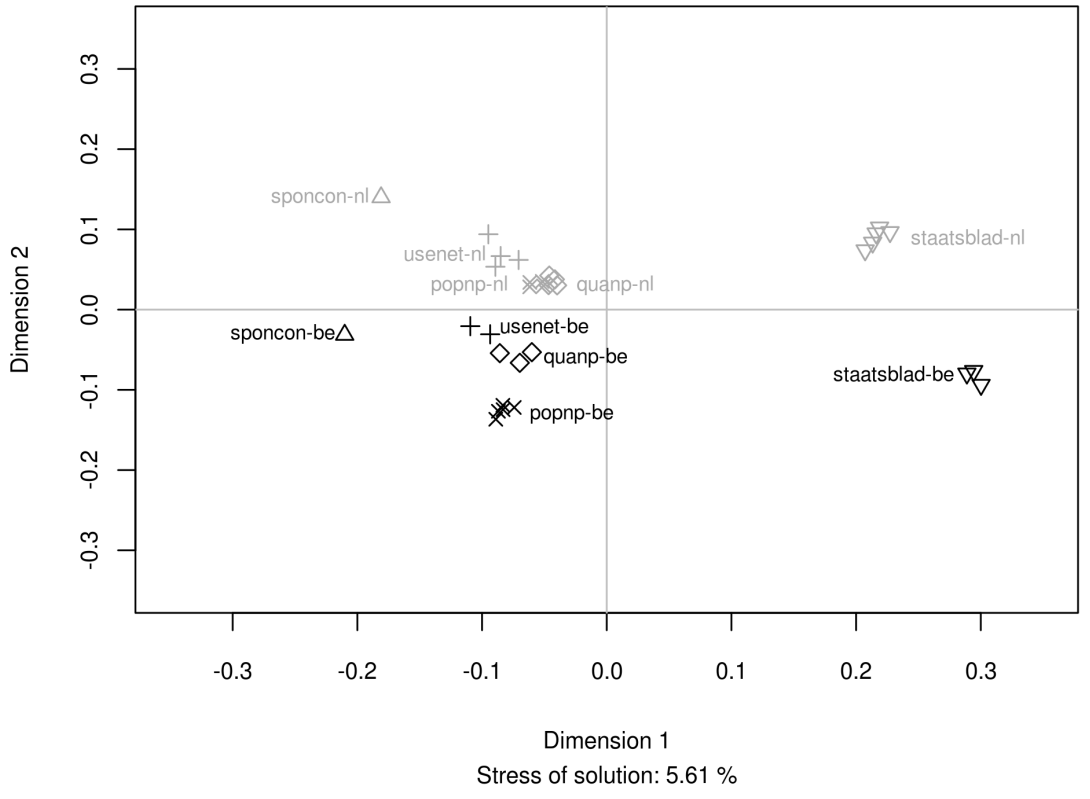

Figure 3. Semasiological and Onomasiological weighting. Dimension 1 reveals a register cline (mirrored in comparison to previous visualizations), Dimension 2 separates the national subcorpora. Subcorpus clusters are tight, and the gradual distinction between spontaneous conversations, Usenet and newspapers has become more clear.

Figure 3 shows the position of the fragments in a two-dimensional space, as calculated with both semasiological and onomasiological weighting. The stress of this solution is again acceptable at 5.61% and the basic interpretation of the dimensions

remains the same: Dimension 1 shows a register cline (mirrored in comparison to previous visualizations, but that is not important in a Multidimensional Scaling solution) and Dimension 2 distinguishes the countries. Comparing Figure 2 and Figure 3, we notice that the spontaneous conversations are pulled out even more, showing a nice three-way distinction between spoken Dutch, written Dutch and legalese. In the middle cluster of written Dutch, a very modest, but interpretable cline can be discovered: whereas the Usenet subcorpora reside on the side of the spontaneous conversations and the quality newspaper subcorpora on the side of legalese, the popular newspapers seem to be in the middle.

4.5. Discussion

In our research question, we have put forward the following three statements on the lectal structure of Dutch:

1. Belgian Dutch is different from Netherlandic Dutch.
2. The diaglossic spectrum in The Netherlands is more limited than the diaglossic spectrum in Belgium.
3. There is a gap between Standard Belgian Dutch and Colloquial Belgian Dutch, whereas there is a more gradual transition between Standard Netherlandic Dutch and Colloquial Netherlandic Dutch.

In this discussion, we verify whether these statements are confirmed in the visualizations presented above. First, we notice that in the three approaches similar solutions were returned when it comes to the interpretation of the dimensions. Dimension 1, which can be considered the most important dimension in a Multidimensional Scaling solution, always produced a register cline. Register differences are thus seemingly more strongly present than the differences on dimension 2 between the national varieties (when it comes to lexical variation). So, indeed, there is (at the level of the lexicon) a general difference between Belgian and Netherlandic Dutch, but it is less important than register differences. This observation shows the importance of visualization techniques for hypothesis generation: a confirmatory statistical anlysis could now be performed on the basis of the hypothesis that the most important predictor of lexical variation in Dutch is not national, but register variation. Of course, one needs to point out that the strong skew caused by the legalese subcorpora might dominate the visualizations. As the input data is lexical, it would not be surprising that the terminologically rich and specific legalese subcorpora are taking up an exceptional position.

Second, throughout the analysis we also find consistently that the quality and popular newspapers in Belgium are separated, whereas they are lumped together in the Netherlands. This seems to link up to the hypothesis that Belgian Dutch has a broader diaglossic spectrum than Netherlandic Dutch.

Third, we do not find visual evidence that there is a gap between Standard Belgian Dutch and Colloquial Belgian Dutch (in the lexicon) versus a gradual transition in Standard Netherlandic Dutch and Colloquial Netherlandic Dutch (in the lexicon).

The solutions are also different on a technical and conceptual level. On the technical level, the stress for the solution that combines onomasiological and

semasiological weighting is the lowest. This means that the Multidimensional Scaling solution for that approach is the most true to its original unreduced distance matrix. However, it would be wrong to conclude that this solution is therefore also the best overall solution. Here, we run into an inherent problem with sociolectometric analyses: there is no gold standard to which solutions can be compared and evaluated.

On the conceptual level, we can observe that the different weighting approaches influence the solution as expected. Adding conceptual weight to the distance metric made the register dimension more concise with stronger subcorpus clusters (compare Figure 1 and Figure 2). Adding semasiological control cleaned up the register distinctions even further and provided more detail so that we could find the expected differences in written Dutch between Usenet, popular newspapers and quality newspapers.

5. Situating lexical sociolectometry and conclusion

The current paper shares a very fundamental point with other contributions in this volume, namely *semantic control*. In the typological paper of Masha Koptjevskaja-Tamm (this volume), the issue of semantic control is explicitly discussed, and also approach with help of the Semantic Vector Space Models. Also, during the workshop, there were concerned remarks on the semantic equivalence of the features that were used in the cross-linguistic study of Stefan Evert et al. Douglas Biber showed in his talk how these remarks can be addressed, cf. Already Biber (1995).

Sociolectometry is also different from many contributions to this volume. Whereas dialectometry and typology focus on one extra-linguistic dimension that should explain the linguistic variation, sociolectometry extends the view to a multidimensional lectal structure. The current paper looks at both a regional (national) and register dimension at the same time, and combines as such the research program of dialectology and the stylistic analyses of Douglas Biber. This broadening of the focus has consequences for the variable set. The features can not be “a priori” selected on the basis of their (regional) distribution, as would be the case in dialect-atlas based dialectometry. Indeed, multiple lectal dimensions should be represented naturally in the variable set. Note that there is no problem in dialectology to focus the variable set towards a regional distribution as the research goal of dialectology is specifically regional variation. In more recent corpus-based dialectometry (e.g. Szmrecsanyi, this volume), however, variable sets are also expanded according to the advice of Nerbonne (2006: 464). An issue of scalability now comes into play. The sheer time-investment needed to collect a variable set that is representative of all variation in the language almost renders the task impossible. Therefore, we proposed in this paper a bottom-up approach to generate a large variable set of lexical alternation variables automatically.

In further research, we set ourselves the task of refining this bottom-up approach to be even more semantically aware. At the moment, the method can not account for polysemy in a word (type). Therefore, the results presented above are not completely based on alternation variables with perfect semantic identity. This is due to the fact that the basic unit of the Clustering by Committee algorithm is not the actual occurrence of a single word (token) in the corpus, but rather an average of all the occurrence of a single lemma (type), as such obscuring the sense differences of a word. To remedy this, we are working to shift from a type-based Vector Model towards a token-based Vector Model.

References

Auer, Peter 2005 Europe's sociolinguistic unity, or: A typology of European dialect/standard constellations. In: Nicole Delbecque, Johan van der Auwera and Dirk Geeraerts (eds.) *Perspectives on Variation,* 7-42. Berlin/New York: Mouton de Gruyter.

Baeza-Yates, Ricardo, and Berthier Ribeiro-Neto 1999 *Modern Information Retrieval*. Addison-Wesley: Association for Computing Machinery Press.

Biber, Douglas 1988 *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, Douglas 1995 *Dimensions of Register Variation: A Cross-linguistic Comparison*. Cambridge: Cambridge University Press.

Bouma, Gosse, Gertjan van Noord and Robert Malouf. 2001 Alpino: wide-coverage computational analysis of Dutch. In: Walter Daelemans, Khalil Sima'an, Jorn Veenstra and Jakub Zavrel (eds.) *Computational Linguistics in the Netherlands 2000. Selected Papers from the 11th CLIN Meeting, Computational Linguistics in the Netherlands 2000,* 45-59 Amsterdam: Rodopi.

Campbell-Kibler, Kathryn. 2011 The sociolinguistic variant as a carrier of social meaning. *Language Variation and Change*, 22, 423-441

Cox, Trevor, and Michael Cox 2001 *Multidimensional Scaling*. London: Chapman and Hall.

Cruse, D. Alan 1986 *Lexical Semantics.* Cambridge: Cambridge University Press.

De Caluwe, Johan 2002 Tien stellingen over functie en status van tussentaal in Vlaanderen. In: De Caluwe, Johan, and Dirk Geeraerts (eds.), *Taalvariatie en Taalbeleid, Bijdragen aan het Taalbeleid in Nederland en Vlaanderen*, 57–69. Antwerpen/Apeldoorn: Garant.

Dunning, Ted 1993 Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1), 61–74.

Firth, John 1957 A synopsis of linguistic theory 1930-1955. In: Frank R. Palmer (ed.) *Selected Papers of J.R. Firth*. London: Longman.

Geeraerts, Dirk 1993 Postmoderne attitudes? *Streven* 60(4), 346–353.

Geeraerts, Dirk 2002 Rationalisme en nationalisme in de Vlaamse taalpolitiek. De Caluwe, Johan, and Dirk Geeraerts (eds.), *Taalvariatie en Taalbeleid, Bijdragen aan het Taalbeleid in Nederland en Vlaanderen*, 87–104. Antwerpen/Apeldoorn: Garant.

Geeraerts, Dirk 2009 Lexical variation in space. In: Juergen Erich Schmidt, and Peter

Auer (eds.), *Language and Space I: Theories and Methods*. HSK Handbook, 821–837. Berlin: Mouton De Gruyter.

Geeraerts, Dirk 2010 *Theories of Lexical Semantics*. Berlin: Mouton De Gruyter.

Geeraerts, Dirk and Hubert Cuyckens 2007 *The Oxford Handbook of Cognitive Linguistics*. Oxford: Oxford University Press.

Geeraerts, Dirk, Stefan Grondelaers and Peter Bakema 1994 *The Structure of Lexical Variation. Meaning, Naming, and Context*. Berlin/New York: Mouton de Gruyter.

Geeraerts, Dirk, Stefan Grondelaers and Dirk Speelman, 1999 *Convergentie en Divergentie in de Nederlandse Woordenschat. Een Onderzoek naar Kleding- en Voetbaltermen*. Amsterdam: Meertens Instituut.

Geeraerts, Dirk, Gitte Kristiansen and Yves Peirsman 2010 *Advances in Cognitive Sociolinguistics*. Berlin/New York: Mouton de Gruyter.

Geerts, Geert 1989 In Vlaanderen Vlaams? *Ons Erfdeel* 32: 525–533.

Goebl, Hans 2006 Recent advances in Salzburg dialectometry. In: Nerbonne, J., Kretzschmar, W. (eds.) *Literary and Linguistic Computing, Special Issue on Progress in Dialectometry: Toward Explanation*, Volume 21(4), 411-435. Oxford; Oxford University Press.

Goossens, Jan 2000 De toekomst van het Nederlands in Vlaanderen. *Ons Erfdeel* 43(1): 2–13.

Hudson, Richard A. 1980 *Sociolinguistics*. Cambridge: Cambridge Textbooks in Linguistics.

Jaspers, Jürgen and Frank Brisard 2006 Verklaringen van substandaardisering: tussentaal als gesitueerd taalgebruik. *Leuvense Bijdragen* 95: 35–70.

Kristiansen, Gitte and René Dirven 2008 *Cognitive Sociolinguistics: Language Variation, Cultural Models, Social Systems*. Berlin: Mouton De Gruyter.

Labov, William 1972 Some principles of linguistic methodology. *Language in Society* 1(1): 97–120.

Nerbonne, John 2006 Identifying linguistic structure in aggregate comparison. *Literary and Linguistic Computing* 21: 463–476.

Pantel, Patrick 2003 *Clustering by committee*. Ph.D. thesis, University of Alberta. Department of Computing Science.

Pantel, Patrick and Dekang Lin 2002 Discovering word senses from text. In: *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and*

Data Mining (KDD 2002), 613–619.

Peirsman, Yves, Kris Heylen and Dirk Speelman 2007 Finding semantically related words in Dutch. Co-occurences versus syntactic contexts. In: *Proceedings of the CoSMO workshop*, 9–16.

Plevoets, Koen2008 *Tussen spreek- en standaardtaal. Een corpusgebaseerd onderzoek naar de situationele, regionale en sociale verspreiding van enkele morfosyntactische verschijnselen uit het gesproken Belgisch-Nederlands*. Ph.D. thesis, University of Leuven.

De Schutter, Georges 1998 Talen, taalgemeenschappen en taalnormen in Vlaams-België. *Verslagen en Mededelingen van de Koninklijke Academie voor Nederlandse Taal- en Letterkunde*, 108 (2-3): 227–251.

Soares da Silva, Augusto 2010 Measuring and parameterizing lexical convergence and divergence between European and Brazilian Portuguese. In: Geeraerts, Dirk, Gitte Kristiansen and Yves Peirsman (eds.), *Advances in Cognitive Sociolinguistics*. Berlin: Mouton de Gruyter.

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

Stroop, Jan 1990 Towards the end of the standard language in the Netherlands. In: J.A. van Leuvensteijn and J.B. Berns (eds.), *Dialect and Standard Language in the English, Dutch, German, Norwegian Language Areas*, 162–177. Proceedings of the Colloquium “Dialect and the Standard Language”, Amsterdam, 15-18 October 1990.

Stroop, Jan 1992 Weg standaardtaal. *Onze Taal* 61(9): 179–182.

Szmrecsanyi, Benedikt 2011 Corpus-based dialectometry: a methodological sketch. *Corpora* 6(1): 45-76.

Nederlandse Taalunie 1998–2004 *Corpus Gesproken Nederlands*. via TST-centrale.

Taeldeman, Johan 1991 Dialect in Vlaanderen. Herman Crompvoets and Ad Dams (eds.), *Kroesels op de Bozzem*, 34–52. Waalre: Stichting Nederlandse Dialecten.

Taeldeman, Johan 1992 Welk Nederlands voor Vlamingen. *Nederlands van nu* 40(2): 33–51.

Taylor, John 1989 *Linguistic Categorization*. Oxford: Oxford University Press.

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

Willemyns, Roland 2005 Verkavelingsbrabants. Werkt het integratiemodel voor tussentalen? *Neerlandica Extra Muros* 3: 27–40.

Willemyns, Roland 2007 De-standardization in the Dutch language. Territory at large. In: Christian Fandrych and Reinier Salverda (eds.), *Standard, Variation und Sprachwandel in germanischen Sprachen / Standard, Variation and language change in Germanic languages*, 265–279. Tübingen: Gunter Narr Verlag.

Wittgenstein, Ludwig 1953/2001 *Philosophical Investigations*. Oxford: Blackwell Publishing.