Lexical variation in aggregate perspective

Tom Ruette, Dirk Speelman, Dirk Geeraerts

Quantitative Lexicology and Variational Linguistics, University of Leuven

1 Introduction

The current paper shows how a sociolectometric approach is needed to disentangle the multidimensional structure of the varieties in a pluricentric language. There are different sociolectometric approaches, i.e. corpus-based methods, perception experiments, or attitude questionaires. Although the focus of a sociolectometric approach is on the varieties, the choice of the variables under analysis is crucial; we focus on lexical variation. Furthermore, in this paper we compare two quantitative corpus-based methods, which differ in their conceptual control of lexical variables: on the one hand, we take a method that ignores the conceptual relationship between the lexemes in the variable set, on the other hand, there is a method that incorporates knowledge about conceptual identity between lexemes. The importance and difficulties of conceptual control when studying variation in the lexicon as a whole is shown by means of a case-study on the pluricentric language Dutch. The pluricentric character of Dutch is now widely accepted: Dutch is used both in Belgium and in the Netherlands, but each nation has its own norm generating center (cf. Clyne, 1992). This is different from the imposed situation in earlier years, especially the sixties, where Dutch in Belgium was supposed to be exogenically modeled on the norms of the Netherlands. Recently, by means of empirical work of e.g. Geeraerts *et al.* (1999) and experimental work of e.g. Impe *et al.* (2008), this historical view had to be adjusted to the current view, as described in Auer (2005).

Rather than providing further empirical proof of the pluricentric character of the Dutch lexicon, the case-study aims to show the pertinence of a sociolectometric methodology that can aggregate patterns of non-categorical lexical variation while incorporating an appropriate amount of conceptual control — in contrast to a methodology that discards any conceptual knowledge. As such, the study touches upon two general issues in the broader field of variationist linguistics: on the level of words, we look at the problematic status of lexical variation and the difficulty of delineating word meaning; on the level of structure, we run into the methodological issue of aggregating the probabilistic variational patterns of many words in order to reach a general view on the lexicon, rather than on individual words.

Let us start, however, more generally with the status of variation in a linguistic system. Attempts of incorporating variational rules in the linguistic system have been criticized (e.g. Bickerton, 1971) on the argument that variation has no place in the search for an abstract and idealized linguistic system of *competence* and *langue*. However, a paradigm-shift in linguistics towards usage-based approaches turned the ubiquity of variation into something that should not be ignored. Nonetheless, even in usage-based Cognitive Linguistics, which studies *parole* by definition and can therefore hardly escape variation, there has been a tendency to overestimate the homogeneity of language communities and consequent non-variability. As of recently, Cognitive Linguistics has taken up the challenge of incorporating variational dimensions in the study of linguistic phenomena. Evidence for this are two collected volumes by Kristiansen & Dirven (2008) and Geeraerts *et al.* (2010) on *Cognitive Sociolinguistics*, which combine theoretical, methodological and empirical studies that incorporate cognitive, semantic and lectal dimensions in their linguistic descriptions. The idea of Cognitive Sociolinguistics is best explained by looking at an exemplar case-study of Szmrecsanyi (2010). In that study, the English genitive alternation between an *of*-construction and an *'s*-construction is approached in the well-known Cognitive Linguistic fashion, with semantic, pragmatic, psycholinguistic, structural and functional predictors. In addition to these typically Cognitive Linguistic predicting factors, however, extra-linguistic factors are included as well: e.g. register (newspaper versus informal), medium (spoken versus written) and geography (British versus American English). Based on many observations of genitive constructions in corpora that are representative of these lectal factors, it appears that "the magnitude of the effect that individual conditioning factors [e.g. semantic and pragmatic factors] may have on genitive choice [...] is demonstrably mediated by language-external [i.e. lectal] factors" (Szmrecsanyi, 2010).

The example given above — representative of a wide-spread trend in Cognitive Linguistics — studies a single linguistic phenomenon very closely. And although the gained insights of these single-feature studies are at the very heart of the linguistic enterprise, they hardly allow for extrapolations and abstractions about the linguistic system in general: it is not because lectal factors have an important mediating influence on the choice of a specific genitive form (in English), that they have the same effect on other linguistic items (in other languages). In order to reach a more general level of that kind, the behavior of many linguistic variables needs to aggregated so that idiosyncratic differences are middled out, structures emerge and systematicity can be induced. This aggregate perspective also appeals to the answer of Geeraerts (2010) on his question on the plausibility of a system when variation is rampant: finding a linguistic system is an empirical question, that can be answered by looking for stastically recurring structural patterns in variational data. Geeraerts' answer to his own question is in that sense very similar to the view of Harder (2010):

> The "system" [...] is the collection of expressive options that are available for speakers to tap in producing actual utterances. Structural differentiation means that these options are linked and subsumable in categories, which entails a degree of abstraction; but because linguistic structures survive by reproduction, linguistic systems tend to have roughly that degree of abstraction which is functional for speakers. (Harder, 2010, p. 271)

This speaker-in-the-community oriented Cognitive Linguistic view on system allows for a degree of variation, because the abstraction is not complete. The abstraction, according to Peter Harder, takes the form of clusters of expressive options — so-called *structural categories* —, but these clusters are fuzzy and do not cover the complete set of expressive options.

Returning to the topic of the current paper (lexical variation in a pluricentric language), how can these theoretical insights be applied? To answer this question, we will address lexical variation in Section 2 and aggregation in Section 3. In Section 4, we will perform a case-study on aggregated lexical variation in the pluricentric language Dutch. Finally, we bring together the theoretical insight and the results of the case-study in the conclusion of this paper.

2 Lexical variation

Harder (2010, p. 270) claims that there are three stages in a socio-dynamic perspective on linguistic system. The first stage consists of mere *fluctuations*, comparable to the brabbling of a toddler. From these fluctuations a structure emerges consisting of categories that contain the fluctuation, but, as we recall from the quote above, this structure is an incomplete abstraction of the fluctuations. The abstraction goes only so far as the language user deems appropriate, c.q. until communication is enabled. This is the second stage of emerging structure. The third stage consists of the initial stage fluctuations that have turned into variation within the emerged structural category. Variationist research zooms in on the third stage, assuming the categories from the second stage. As an example, Harder gives the seminal Labovian study on the structural category "postvocalic -r", with its category-bound variants, which appeared to be related to social classes in New York (Labov, 1966). Scholars of the linguistic system have traditionally removed staged three (variation, or rather variable usage) and focused on the abstract and idealized stage (two) of structural categories. However, an adequate study of the linguistic system must not ignore the stage three variation, as structure and variation can not exist without each other. Structure without variation is ridden of the linguistic reality, and variation without structure is mere fluctuation, incapable of enabling communication.

Although this idea of system is primarily geared towards linguistic categories such as consonants or Germanic strong verbs, it can conveniently be "translated" towards the conceptual categories of the lexicon. There is, however, an important question related to the level of abstraction in stage two. If on the one hand the categories are chosen to be as narrow as a single word (or symbol), the variation within these categories is *semasiological variation*. This means that one studies the different senses or aspects of meaning of a single word. If on the other hand the categories are chosen to be as broad as "concepts", the variation in naming these categories (i.e. that different words may name the same concept) is *onomasiological variation*. This means that one studies the different ways of expressing (with words) the conceptual category. Obviously, this very old distinction between a semasiological or an onomasiological approach is related to the study of polysemy versus the study of synonymy.

In this paper, we restrict ourselves to onomasiological perspective, yet fully aware of the semasiological issues waiting around the corner. We refer to Geeraerts (2009) for an overview of research on lexical variation, and zoom in here briefly on a distinction between *Formal Onomasiological Variation* (FOV) and *Conceptual Onomasiological Variation* (COV). A FOV approach resembles the sociolinguistic variable: FOV grasps a quality of a set of words that express the same concept, and just like in a sociolinguistic variable, each word in the set may have a specific socio-stylistic correlation. COV, on the other hand, links up to the more subtle variation in concepts that are being used in language. Most obviously, at a very high level, and example could be that one can use specific words to talk about "beer" or about "semantics". At a more fine-grained level, one could say that "fiddle" and "violin" are an example of FOV, but because "fiddle" has a slightly more ordinary tone to it than the more prestigious "violin", there is also COV between these words. In the case-study to this paper, we will show that this distinction between FOV in *choosing* a word to express a concept versus COV when *using* words to talk in a certain way crops up in a methodological difference between the two sociolectometric approaches that we compare.

3 Aggregation

As said above, aggregation of many variables is necessary when the goal is to describe general patterns in a system. In order to find underlying dimensions of variation in a large set of (lexical) variables, the individual patterns of the variables thus need to be aggregated. Aggregation of many features is already applied in e.g. dialectometry and text categorization. However, we find problems in both dialectometry and text categorization when it comes to dealing with lexical variation.

In dialectometry (Seguy, 1971; Goebl, 1975; Nerbonne & Kretzschmar, 2003), lexical variation is almost always considered to be categorical per location (except e.g. Grieve *et al.* , 2011): either a certain location — or at best a single interviewee per location — is attributed the use of word *a* or the use of word *b*. This categorical approach is mainly due to the type of input data, i.e. a lexical dialect atlas, used in most dialectometric studies. Dialect atlases have been painstakingly constructed in earlier years by the efforts of dialectologists that visited pertinent locations for their purposes and accumulated data through interviews and questionnaires. Categorical word choices per location were a necessary (but currently not any longer acceptable) methodological decision. Because dialectometric methodology is tailored around the categorical dialect atlas input format, their quantitative aggregation methods can not straightforwardly be applied to corpus-driven input, where lexical variation is a probabilistic matter.

Unlike dialectometry, an aggregation method that incorporates both probabilistic word preferences in an onomasiological approach was introduced in Geeraerts *et al.* (1999) and further formalized in Speelman *et al.* (2003). This so-called *profile-based* approach — where "profile" stands for the (relative frequencies of a) set of words in a conceptual category — is formally introduced below. The rationale of the method is — just like most aggregation methods — to measure the "distance" between pairs of subcorpora on the basis of their probabilistic overlap in onomasiological word preferences for expressing an underlying conceptual category. A small distance between subcorpora implies a general agreement in word choice, whereas a large distance implies a general disagreement in word choice.

Profile-based distances between subcorpora are calculated by means of the following method. Given two subcorpora V_1 and V_2, a conceptual category L (e.g. SUBTERRANEAN PUBLIC TRANSPORT) and x_1 to x_n the exhaustive list of variants (e.g. {`subway`, `underground`, `tube`} as the profile, then we refer to the absolute frequency F of the usage of x_i for L in V_j with[1]:

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

Subsequently, we introduce the relative frequency R:

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

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

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

The City-Block distance is a straightforward descriptive dissimilarity mea-

[1] The following introduction to the City-Block distance method is taken from Speelman *et al.* (2003, Section 2.2).

sure 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. If however the samples are rather small, the relative frequencies become unreliable, and a supplementary control is needed. For this we use a measure that takes as its basis the confidence of there being an actual difference between two profiles: the *Fisher Exact* test. This time, unlike with D_{CB}, we look at the absolute frequencies in the profiles we compare. When we compare a profile in one subcorpus to the profile for the same concept in a second subcorpus, we use a Fisher Exact test to check the hypothesis that both samples are drawn from the same population. We use the p-value from the Fisher Exact test as a filter for D_{CB}. We set the dissimilarity between subcorpora at zero if $p > 0.05$, and we use D_{CB} if $p < 0.05$.[2]

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:

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

The W in the formula is a weighting factor. We use weights to ensure that concepts which have a relatively higher frequency (summed over the size of the two subcorpora that are being compared[3]) 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 are treated as more important.

Now, we put text categorization in contrast with the profile-based approach, which incorporates probabilistic information of word choice. In text categorization, non-categorical (probabilistic) word choice is well accounted for (unlike dialectometric approaches), but text categorization totally ignores the onomasiological perspective on lexical variation. This is primarily due to the fact that text categorization often zooms in on topical categorization, and the onomasiological approach to lexical variation within conceptual categories is exactly a way of downplaying thematic bias in the variational patterns (Speelman *et al.* , 2003). However, other forms of text categorization, e.g. authorship attribution or linguistic profiling — quite the opposite of topic classification —, also ignore onomasiological variation and use mere (relative) occurrence frequencies of the features in the aggregation step. This is problematic, especially given the recent trend in authorship attribution studies to use content words.

Whereas the profile-based approach will be the quantitative method that in-

[2] To employ the Fisher Exact test, the subcorpora need to be more or less equal in size. Also, if the frequency of the profile was lower than 30 in the two varieties that are being compared, that profile was excluded from the comparison.

[3] The size of the two subcorpora is not the actual amount of words in the two subcorpora, but the sum of all profiles in these two subcorpora with a frequency higher than 30.

corporates conceptual control in our comparison of methods, we will use the text-categorization approach as the quantitative method that ignores conceptual similarity between the words in the variable set. Except for the used distance metric, the two approaches are identical. The underlying metaphor of both the profile-based and categorization approach is spatial: subcorpora are represented as points in an n-dimensional space by means of the occurrence frequencies of n words. A made-up example in a two-dimensional space, i.e. with two words, containing two text types might make this rather abstract metaphor more clear. Given two subcorpora representing the text types "academic articles" and "computer mediated communication", and given two words "hence" (a linking word used in academic articles) and "LOL" (an abbreviation of Laughing Out Loud, commonly used in IRC), one might construct the "space" in Figure 1. The position of the academic articles in the bottom right part is due to the high frequency of "hence" and the low frequency of "LOL" in these texts. The position of the computer-mediated communication in the top left part is due to the low frequency of "hence" and the high frequency of "LOL" in these texts. Obviously, these data are made up for the sake of the argument. Now, two lines can be drawn through the origin of the space and the position of the text types (on the basis of the frequencies of the words that make up the dimensions), yielding an angle, for which the cosine can be calculated. A small angle implies high similarity between the text types, and will yield a high cosine value; a large angle implies low similarity, and will yield a low cosine value. More information on the cosine metric can be found in Baeza-Yates & Ribeiro-Neto (1999).

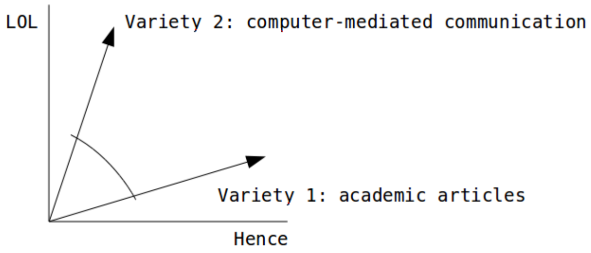
LOL
Variety 2: computer-mediated communication
Variety 1: academic articles
Hence

Figure 1: 2 Dimensional example of Vector Model

Formally, given two subcorpora V_1 and V_2 in which the frequencies of a large number of words were counted and stored in the respective vectors $\vec{x}$ and $\vec{y}$, we calculate the distance between the subcorpora by means of Equation 5.

$$D_{cos}(V_1, V_2) = 1 - cos(\vec{x}, \vec{y}) = 1 - \frac{\vec{x} \cdot \vec{y}}{|\vec{x}||\vec{y}|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2} \tag{5}$$

4 Case-study

The case-study of this paper is an analysis of aggregated lexical variation in the pluricentric language Dutch. It consists of a comparison between the state-of-the-art text categorization distance metric, which ignores conceptual control, and the profile-based distance metric, which includes conceptual control. In order to garantuee an objective comparison, we will apply both methods to the same dataset, which is tailored to contain a specific constitution of variational dimensions. The method that best approaches the expected structure will be considered superior. In what follows, we first introduce the dataset by describing the set of lexical features and the corpus in which these features will be counted. Second, we apply the profile-based method to this dataset. Then, the state-of-the-art text categorization method is also applied to the dataset. Finally, it will be concluded that the profile-based onomasiological approach grasps the a priori constitution of variational dimensions much better than the text categorization method.

The lexical input features are derived from the "Referentiebestand Belgisch Nederlands" (Martin (2005), Eng. *Reference List of Belgian Dutch*, abbreviation "RBBN"). This reference list contains words or expressions that exclusively appear in Belgian Dutch, and have no occurrences in The Netherlands, according to dictionaries, corpora and informants. The list contains about 4000 items, ranging from colloquial items, over culturally linked (e.g. Belgian institutes) to register-specific and freely varying items. As an example, a small selection of items is listed in Table 1, but the whole list can be downloaded freely from the website of the "Instituut voor Nederlandse Lexicologie". For each Belgian Dutch item, the list provides an alternative from general Dutch, or sometimes typically Netherlandic Dutch. From the 4000 items on the list, we only retained 1455 items for which the Belgian Dutch item itself and its alternative consist of one single word. If we restrict the RBBN list to these single word items — and thus excluding multi-word-units and expressions —, these items can be counted accurately in an automatic way by merely keeping track of the occurrence frequency of the words in the subcorpora[4]. Indeed, expressions and multi-word-units may be distributed over the sentence because of syntactic constructions, making automatic counting very hard. All (single) words on the list were analyzed with the Alpino parser, so that accurate countings on the lemmata could be performed, while controlling for the part-of-speech. Linking back to the issue of conceptual categories in Section 2, we accept the conceptual categories of the makers of the RBBN in their equivalence judgement between the Belgian Dutch item and its alternative.

Because we know that this list contains Belgian Dutch words and an alternative, we can predict that the main variation in the list will be due to a na-

[4] We address the issue of possible polysemy issues and the need for word sense disambiguation when doing automatic counting in the conclusions.

Belgian Dutch	*General Dutch*	*Translation of concept*
suikerboon	doopsuiker	candy to honor the birth of a baby
appelsien	sinaasappel	orange (fruit)
unaniem	eenparig	unanimous
ambras	ruzie	a row
confituur	jam	marmalade
binnenkoer	binnenplaats	atrium

Table 1: Selected examples from the RBBN

tional pattern. Indeed, even the non-national variation which is present in the list (e.g. colloquialisms) is still embedded in the Belgian Dutch point-of-view of the RBBN. Or in other words, every variable in the variable set is at least nationally patterned. Therefore, we expect the results of our method to show a clear distinction between the two national varieties, and other variational dimensions will only appear after that.

In our corpus, we incorporate samples from the two national varieties of Dutch, taken from two registers (quality newspapers and Usenet), and from two topics (politics and economy). We collected a total of 6 million words, which were evenly split over the nations, registers and topics. The quality newspaper articles were sampled from two large newspaper corpora that are available for both Netherlandic and Belgian newspapers. From these two corpora, we selected four newspapers that are deemed to be quality newspapers: "De Standaard" and "De Morgen" for Belgium, and "Volkskrant" and "NRC" for The Netherlands. For most of the articles that appeared in the newspapers, there is access to the category in which it was published. This categorization was used to filter out the articles on the topics "politics" and "economy". The Usenet posts were downloaded from a large Usenet archive, available online at Google Groups and automatically stripped from meta-information (headers and html code) and reduplicated content (quotes from previous posts). Only posts from the groups "be.politics", "be.finance", "nl.politiek" and "nl.financieel.*" were downloaded, where the country affiliation of the group was taken to be an indication of the nationality of the author of the post, and where the topical restriction of the group indicates the topic of the post. All texts were lemmatized and tagged with part-of-speech information by the Alpino parser (Bouma *et al.* , 2001).

With these three dimensions (country, register, topic) and two levels for each dimension 8 combinations are possible. These combinations, e.g. Belgian quality newspapers on economy (abbreviated as `qnp.be.e`), will be represented by the subcorpora, for which we will calculate the pairwise distances. However, to increase the number of data points and in order to verify the internal consistency of the subcorpora, we divided every subcorpus into two equally sized groups (abbreviated as e.g. `qnp.be.e.0` and `qnp.be.e.1`). In total then, we counted the frequencies of the linguistic characteristics which we introduce above, in 16 subcorpora.

Given the omnipresent country dimension in the input features, the primary variational dimension that could be exptected to be revealed among the subcorpora is the Belgian Dutch versus Netherlandic Dutch dimension. Or in terms that relate to the distance measurement method: in a pair-wise comparison of subcorpora with a national difference, the distance will be bigger than a comparison of two subcorpora with the same national affiliation. Because the typical Belgian Dutch words are sometimes restricted to a specific register, e.g. colloquialisms, a register distinction should emerge, as well. And as words and their conceptual categories are inevitably sensitive to topic, we would expect the difference between political and economical subcorpora to emerge, too. However, the register and topic dimension should be secondary to the country dimension.

4.1 Results of profile-based method

We first look into the results of the profile-based approach, introduced above. To the selected Belgian Dutch items on the RBBN list, we added the knowledge which alternatives are conceptually equivalent General Dutch words. In other words, we introduce conceptually controlled profile information to the distance metric. A profile thus consists of a Belgian Dutch word from the RBBN list, together with its general Dutch alternative. Remember that the underlying distance metric is basically a City-Block distance measure (see Formula 3). Now, we zoom in on the two- and three-dimensional visualizations of all the pairwise profile-based distances between the subcorpora, made by means of non-metric two-way one-mode Multidimensional Scaling Cox & Cox (2001), as can be seen in Figure 2.[5]

Multidimensional Scaling is a dimension reduction technique which is applied here to a matrix holding all the pairwise profile-based distances between the subcorpora. Because the result of a Multidimensional Scaling analysis is a reduction of the original input, a certain error is introduced. The error-rate is grasped by a "stress" value, with 0% stress equal to no error at all. It is generally acceptable to present Multidimensional Scaling solutions up to a stress level of 10-15%. Usually, Multidimensional Scaling is used to return one-, two-, or three-dimensional reductions, so that visualization is possible. With every added dimension, the error-rate goes down, as the reduction becomes less severe. The fall of error-rate with added dimensions is grasped in a so-called *screeplot*. The screeplot in Figure 3 shows a stress difference of about 7% between a one-dimensional and a two-dimensional Multidimensional Scaling solution. Therefore, we first interpret the horizontal dimension (of an unrotated solution) as it represents the most important variation in Figure 2. In this case,

[5] The coordinates of a Multidimensional Scaling solution can be scaled freely, as long as the same scaling is applied to all dimensions. Therefore, we discarded a scale on the axes, as these numbers would not be insightful. However, we made sure that the x and y (and z for three-dimensional solutions) axes are always equal, so that the distances between the subcorpora can be interpreted.

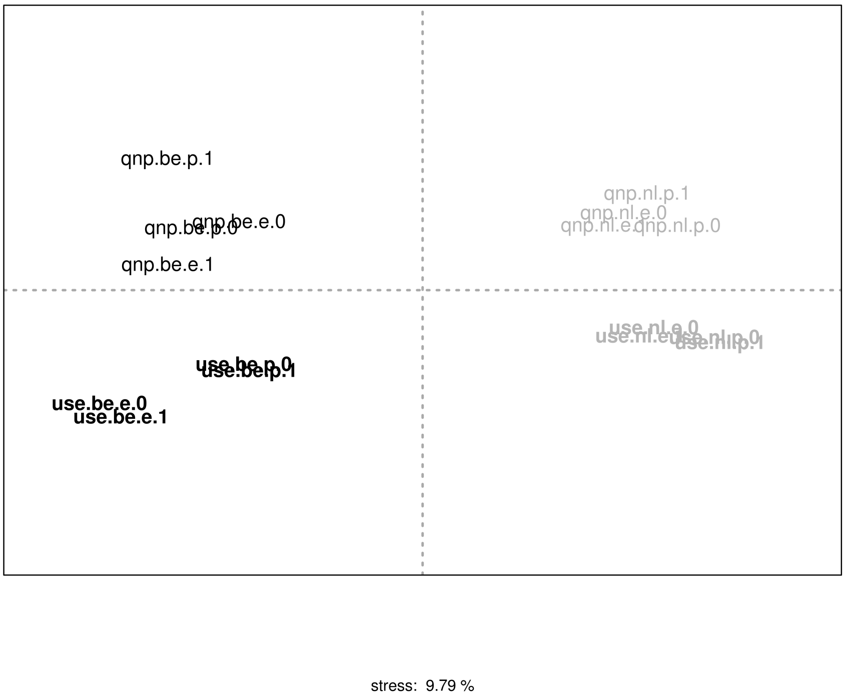

Figure 2: Linguistic distance between subcorpora (profile-based, two-dimensional)

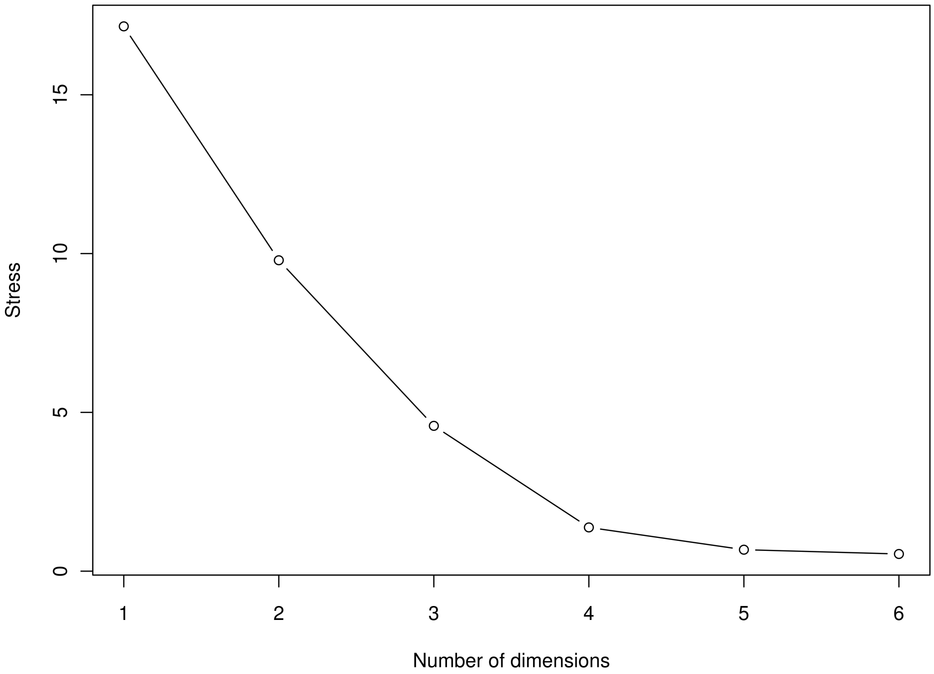

Figure 3: Screeplot for non-metric Multidimensional Scaling solution (profile-based)

the profile-based approach makes a distinction between Belgian subcorpora (black font) and Netherlandic subcorpora (grey font) on the first dimension. The grey zero-line divides the two countries perfectly. The vertical dimension makes a distinction between quality newspapers (normal font) and usenet articles (bold font). Here again, the grey zero-line marks a perfect distinction between the two registers. Overall, there is a very clear grouping of the subcorpora, with only clear separation of the topics in the Belgian Usenet. The range of Belgian register variation is also somewhat larger than the Netherlandic range, but this has probably to do with the focus on Belgian Dutch variation in the input features. Most importantly, however, the profile-based approach yields a visualization that complies with our expectations of finding a national pattern first, followed by register variation on the second dimension.

The screeplot suggest that a three-dimensional solution might even improve the quality of the visualization with another 5 or 6%. Therefore, we calculated a three dimensional solution, which is represented in Figure 4. Instead of rendering a three-dimensional plot, we drew the scatterplot of dimension 1 versus dimension 2, and the scatterplot of dimension 1 versus dimension 3. This shows us how, even in a three-dimensional solution, dimension 1 still divides Belgian and Netherlandic subcorpora, and that dimension 2 divides the quality newspaper articles from Usenet. However, this register division in the three-dimensional solution is not as neat as in the two-dimensional solution, because one of the Netherlandic Usenet fragments crosses over into the quadrant of the Netherlandic quality newspaper fragments. For dimension 3, we can see a split for the topics of the Belgian subcorpora, with on the top left of dimension 3 subcorpora with an e for economy-related subcorpora, and politics fragments at the bottom. On the Netherlandic side, the register (dimension 2) and topic (dimension 3) split is muddled. The register and topic divisions of the Belgian subcorpora, however, are perfect for respectively dimension 2 and dimension 3. The quality of the grouping on the Belgian side is obviously due to the input variables which are specifically sensitive for Belgian Dutch variation. This indicates that the choice for a Belgian Dutch term is not only nationally patterned, but also stylistically.

4.2 Results of categorization method

Now, we present the method and the results of the state-of-the-art categorization approach, which uses the cosine similarity metric, instead of the adapted City-Block distance that is used in the profile-based approach.

In the current case-study, we take the RBBN items (and the alternatives) as individual features and remove the knowledge of conceptual categorization. If we calculate the similarities (and consequent distances) with these input features between the subcorpora in our dataset, and then produce the two-dimensional visualization with Multidimensional Scaling, we get the plot in Figure 5. If we create a screeplot (Figure 6) to show us how much stress difference there is be-

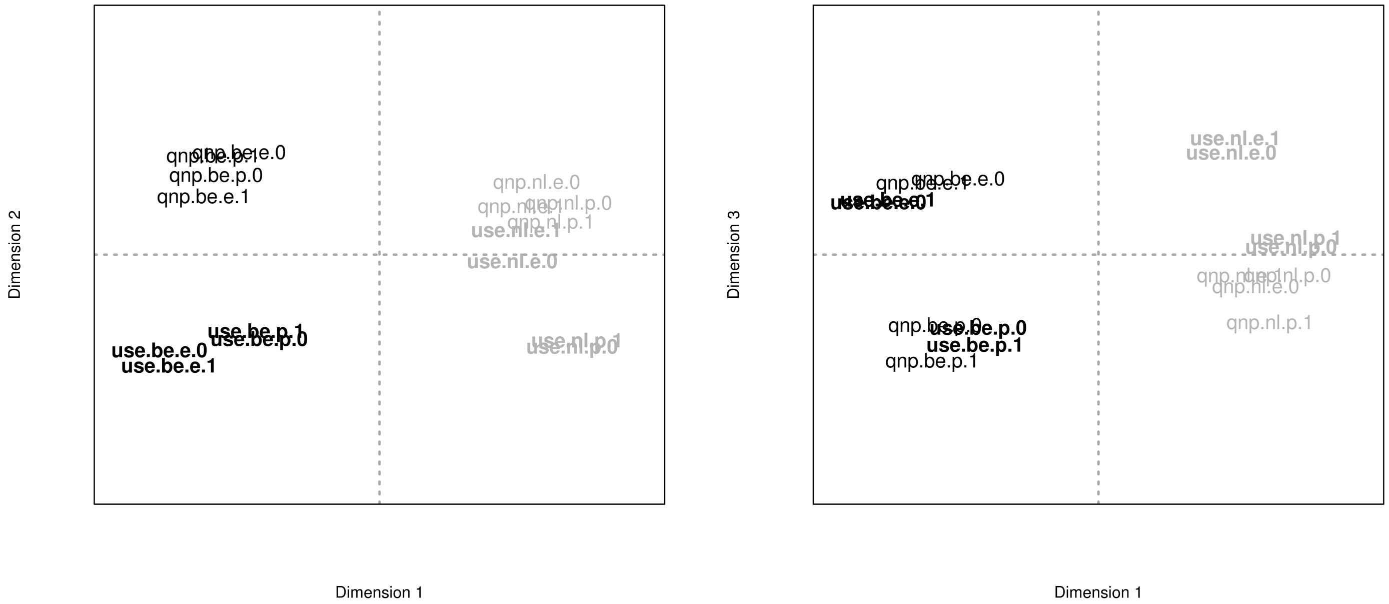

Figure 4: Linguistic distance between subcorpora (profile-based, three-dimensional)

tween the first and the second dimension, we see that the second dimension reduces the stress of a one-dimensional solution with about 8%. Therefore, we will interpret the two dimensions in their own respect, knowing however that the first dimension (of an unrotated solution) represents more “important” variation than the second dimension.

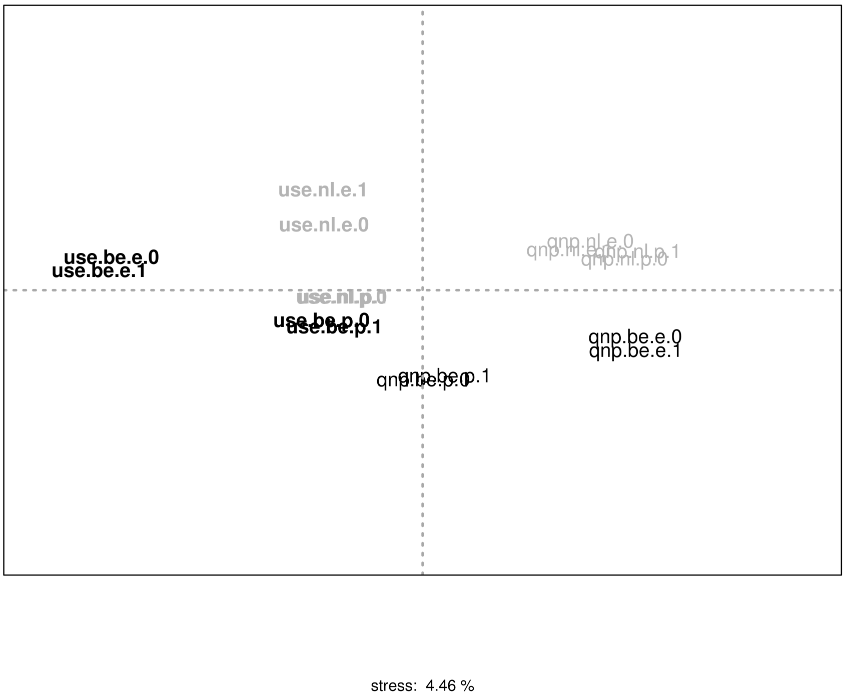

Figure 5: Linguistic distance between subcorpora (cosine, two-dimensional)

In Figure 5 we see on the horizontal axis (from left to right, dimension 1) a distinction between the Usenet articles (bold font) and the quality newspaper articles (regular font). The light grey vertical line indicates the zero-line of the horizontal dimension. Normally, that line demarcates the boundary between two areas. However, in the current approach, we see that the quality newspapers from Belgium on politics are crossing this line slightly. Moreover, whereas we would expect the most important variation (thus, on the horizontal dimension) to be related to country, we encounter a distinction between registers. The vertical dimensions (from bottom to top) tends to divide Belgium (black font) from The Netherlands (grey font), but not very clearly. The (politics) Netherlandic usenet articles sink below the horizontal zero-line, and the (economy) Belgian usenet articles rise above that line. Moreover, we notice that the topics are set apart in groups, as well, except for the quality newspapers from The

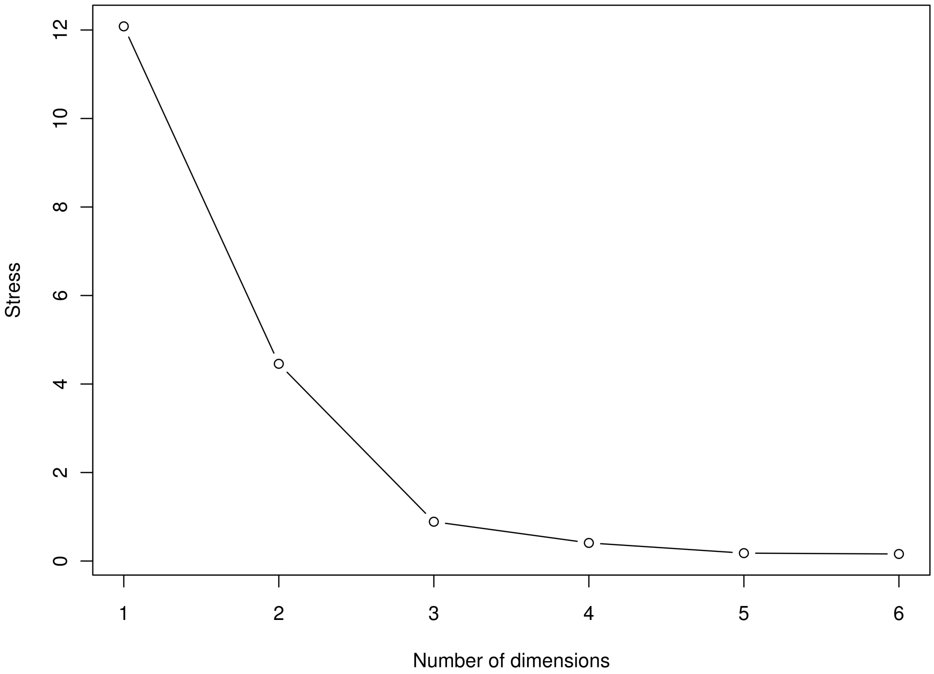

Figure 6: Screeplot for non-metric Multdimensional Scaling solution (cosine)

Netherlands. All in all, the categorization approach yields somewhat unclear grouping of subcorpora and an unexpected promotion of register variation as the most important variation in the input features.

The screeplot shows that a three-dimensional solution would reduce the stress even more up to an almost optimal level. Therefore, we calculated a three-dimensional solution and represent the three dimensions in Figure 7. We apply the same idea as for the profile-based approach to plot dimension 1 and 2, and then dimension 1 and 3. Just like in the two-dimensional solution, we see that dimension 1 tends to divide quality newspaper fragments from Usenet fragments, and that dimension 2 tends to divide the national subcorpora. The three-dimensional solution does a slightly better job than the two-dimensional solution, because the nation division on dimension 2 is now almost correct. Dimension 3 divides largely the topics, with politics-related fragments at the top, and economy-related fragments at the bottom. This division is almost perfect, although the grouping of the subcorpora is not so neat. Overall, though, the categorization method yielded messier output than the profile-based approach.

5 Conclusion

The two main theoretical questions of this paper have been (a) how important is the notion of a conceptual category in an aggregate study of variation in the lexicon and (b) what is the status of conceptual categories for lexical variation? Moreover, we have claimed that sociolectometric methodology, of which the current study is an example, is needed to study a pluricentric language. The link with pluricentric languages, c.q. Dutch, is also made in the case-study, which shows how conceptual categories — and their consequent conceptual control — are necessary to reveal the national dimension in the lexicon. In other words, the national varieties of Dutch do not differ so much in their *use* of words — both Belgium and the Netherlands use different words for different topics and registers —, but they do differ in their *choice* of words — for expressing a conceptual category. This latter point is made clear in the case-study by means of the comparison between a profile-based onomasiological approach and a text categorization approach. The text categorization approach grasped the mere use of individual words and compared the use of words in two subcorpora by means of the cosine similarity metric, which was not informed about the conceptual similarity between words. Consequently, the text categorization showed that there was a pattern of register and topic in the input features, stronger than the anticipated national pattern. The onomasiological approach, on the contrary, revealed a strong national dimension in word choice for naming a conceptual category.

Given the a priori known pattern of national variation in the dataset used in the case-study, one might jump to the conclusion that an onomasiological approach is better suited for finding variational patterns in the lexicon, and the

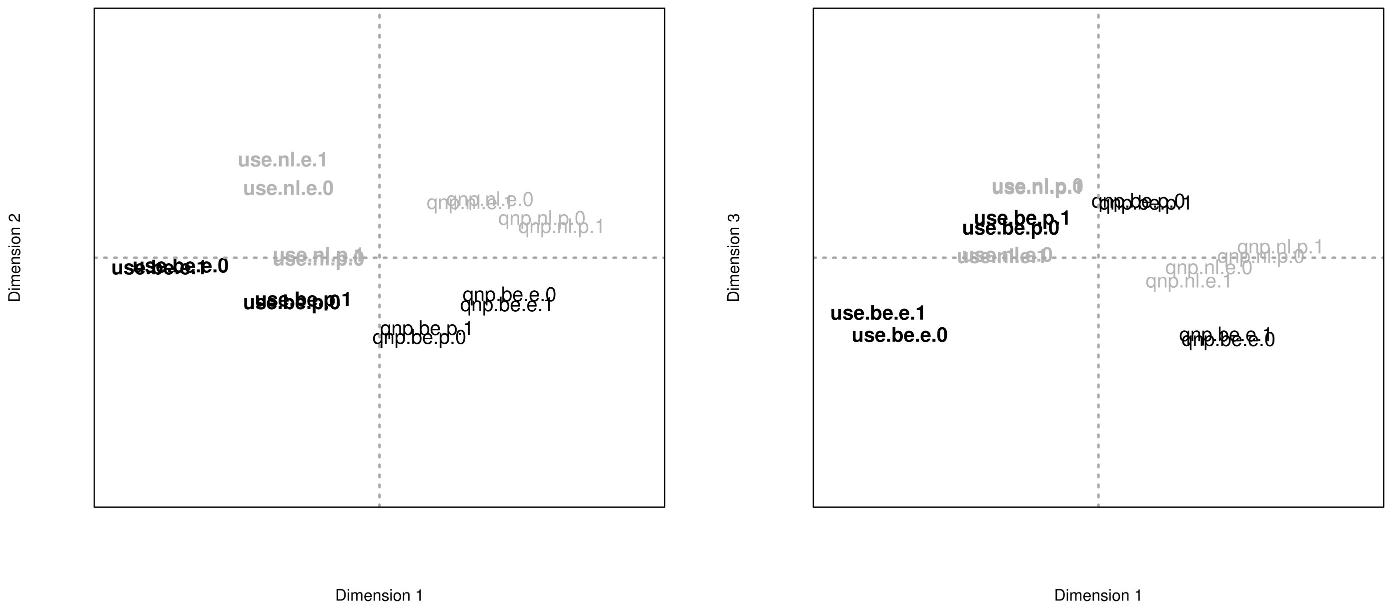

Figure 7: Linguistic distance between subcorpora (cosine, three-dimensional)

preferred method for any sociolectometric study. However, there are a number of problems with this conclusion.

First of all, perhaps we are wrong in the assumption that national variation is the strongest dimension in the lexical variable set and the available subcorpora; it could be well possible that word use — as shown in the categorization approach – is actually more strongly influenced by a register or topic dimension, and that the onomasiological approach artificially weakens these dimensions[6]. In that case, we would have to tone down the conclusion, and say that an onomasiological approach with conceptual control is a methodological means of revealing and boosting specific underlying dimensions of variation. Moreover, we would like to point out that our corpus only sampled two topics and two registers, which is not enough to support strong generalizations. Further research is therefore needed with more topics and registers. All this, of course, does not weaken the strenght of a profile-based approach, but it rather points out the importance of knowing what is being measured. Our claim now is that the profile-based approach allows for much more control over what is measured than the text categorization method, and should therefore be preferred.

Second, the onomasiological approach assumes a relation of identity of (conceptual) meaning between the variants and this is theoretically problematic. Following Edmonds & Hirst (2002), we agree that perfect synonymy — the highest possible level of detail in describing a conceptual category, and still finding multiple words that fit the category — is extremely rare. By admitting this, our notion of semantics or word meaning follows the Cognitive Linguistic view that encyclopedic knowledge is indispensable. Translating the idea of Peter Harder that structural categories need not to be complete, and that the abstraction goes only as far as is functional for language users — here we link up to the prototype theory of word meaning, cf. Rosch & Mervis (1975) —, we can reach near-synonymy by slightly relaxing the level of detail of the conceptual category: not every language user has an identitical representation of a word in his head, but nonetheless two language users can communicate with that word. Idealized Cognitive Models (Lakoff, 1987) or Frames (Fillmore, 1994) are examples of describing meaning, while balancing semasiological detail and operational functionality. In future research, we will operationalize the bottom-up creation of conceptual categories by applying Word Space Models (Turney & Pantel, 2010).

Third, an onomasiological approach requires prior semasiological analysis to exclude contextual nuances or polysemy. In the case-study of this paper, the lemmatized forms of the RBBN words were naively counted in the corpus, without further checking the context of each occurrence. Closer inspection revealed that the RBBN list does not contain many potential polysemous items, so that we can ignore the small error that must be present in the frequencies for the purposes of the current paper. However, as we want to perform in future research

[6] Although the profile-based City-Block distance incorporates a W term that brings the frequency of the conceptual category in play.

the above analyses with a naturalistic sample of lexical variation, instead of an a priori list of national variation, a semasiological study for every occurrence needs to be done in order to establish the conceptual control. As this would be an unfeasible manual task when using a large amount of variables, we will rely further on the advances being made in the field of Word Space Models.

To conclude this paper, we try to answer our initial questions. How important is the notion of a conceptual category in an aggregate study of the lexicon? The case-study has shown that conceptual control is necessary to reveal variational dimensions that are hidden in the overwhelming content (topic) function of words. Without conceptual control, the conclusion of the categorization approach would have been that different words are used to refer to different content, and that they may also signal register and perhaps national differences. This observation, albeit true and undeniable, is not the goal of an aggregation study: it is obvious that an aggregation of many words will be sensitive to content differences among subcorpora. Therefore, conceptual control, in the form of conceptual categories that group together similar words, is needed. And this brings us to the second question: what is the status of conceptual categories for lexical variation? Although practical as a methodological, heuristic device, the conceptual categories remain somewhat artificial because of the flexibility in their definition. In the current case study, the makers of the RBBN clearly had referential equivalence in mind for most categories. However, conceptual categories can be defined more strictly or less strictly at a whimp of the researcher, because there is no consensus over the appropriate level of detail in the definition, especially since the incorporation of encyclopedic knowledge in word-meaning. The level of detail that is operational in the language community can only be retrieved by studying the actual use of words.

And then we are back at variation.

References

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

Baeza-Yates, Ricardo, & Ribeiro-Neto, Berthier. 1999. *Modern Information Retrieval.* ACM Press / Addison-Wesley.

Bickerton, Derek. 1971. Inherent Variability and Variable Rules. *Foundations of Language and Cognitive Processes*, **7**(4), 457–492.

Bouma, Gerlof, van Noord, Gertjan, & Malouf, Rob. 2001. Alpino: wide-coverage computational analysis of Dutch. *Pages 45–59 of:* Daelemans, Walter, Sima'an, K., Veenstra, J., & Zavrel, J. (eds), *Computational Linguistics in the Netherlands 2000. Rodolpi, Amsterdam.*

Clyne, Michael. 1992. *Pluricentric languages: differing norms in different nations.* Mouton de Gruyter.

Cox, Trevor, & Cox, Michael. 2001. *Multidimensional Scaling.* Chapman & Hall.

Edmonds, Philip, & Hirst, Graeme. 2002. Near-synonymy and Lexical choice. *Computational Linguistics*, **28**(2), 105–144.

Fillmore, Charles. 1994. Starting where dictionaries stop: the challenge of corpus lexicography. *Pages 349–393 of:* Atkins, B.T. Sue, & Zampolli, Antonio (eds), *Computational Approaches to the Lexicon.* Oxford: Oxford University Press.

Geeraerts, Dirk. 2009. Lexical variation in space. *Chap. 45, pages 821–837 of:* Schmidt, Juergen Erich, & Auer, Peter (eds), *Language and Space I: Theories and Methods.* HSK Handbook. Mouton De Gruyter, Berlin.

Geeraerts, Dirk. 2010. Schmidt redux: How systematic is the linguistic system if variation is rampant? *Pages 237–262 of:* Boye, Kasper, & Engberg-Pedersen, Elisabeth (eds), *Language Usage and Language Structure.* Berlin/New York, Mouton de Gruyter.

Geeraerts, Dirk, Grondelaers, Stefan, & Speelman, Dirk. 1999. *Convergentie en divergentie in de Nederlandse woordenschat. Een onderzoek naar kleding- en voetbaltermen.* Amsterdam: Meertens Instituut.

Geeraerts, Dirk, Kristiansen, Gitte, & Peirsman, Yves (eds). 2010. *Advances in cognitive sociolinguistics.* Berlin/New York: Mouton de Gruyter.

Goebl, H. 1975. Dialektometrie. *Grazer linguistisch Studien*, 32–38.

Grieve, Jack, Speelman, Dirk, & 2011, Dirk Geeraerts. 2011. A statistical method for the identification and aggregation of regional linguistic variation. *Language Variation and Change*, **23**, 193–221.

Harder, Peter. 2010. *Meaning in Mind and Society: A functional contribution to the social turn in Cognitive Linguistics.* Cognitive Linguistics Research, no. 41. Berlin/New York, Mouton de Gruyter.

Impe, Leen, Geeraerts, Dirk, & Speelman, Dirk. 2008. Mutual intelligibility of standard and regional Dutch language varieties. *International Journal of Humanities and Arts Computing*, **2**, 101–117.

Kristiansen, Gitte, & Dirven, Rene (eds). 2008. *Cognitive Sociolinguistics: Language Variation, Cultural Models, Social Systems.* Cognitive Linguistics Research. Mouton De Gruyter, Berlin.

Labov, William. 1966. *The social stratification of English in New York City.* Center for Applied Linguistics.

Lakoff, George. 1987. *Women, fire and dangerous things: what categories reveal about the mind.* University of Chicago Press, Chicago.

Martin, Willy. 2005. *Het Belgisch-Nederlands anders bekeken: het Referentiebestand Belgisch-Nederlands (RBBN).* Tech. rept. Vrije Universiteit Amsterdam, Amsterdam, the Netherlands.

Nerbonne, John, & Kretzschmar, William. 2003. Introducing Computational Techniques in Dialectometry. *Computers and the Humanities*, **37**, 245–255.

Rosch, E., & Mervis, C.B. 1975. Family Resemblances: Studies in the Internal Structure of Categories. *Cognitive Psychology*, **7**(4), 573–605.

Seguy, J. 1971. La Relation entre la distance spatiale et la distance lexicale. *Revue de Linguistique Romane*, **35**, 335–357.

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

Szmrecsanyi, Benedikt. 2010. The English genitive alternation in a cognitive sociolinguistics perspective. *In:* Geeraerts, Dirk, Kristiansen, Gitte, & Peirsman, Yves (eds), *Advances in Cognitive Sociolinguistics.* Berlin/New York, Mouton de Gruyter.

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