

Profile-based linguistic uniformity as a generic method for comparing language varieties

DIRK SPEELMAN, STEFAN GRONDELAERS, DIRK GEERAERTS

University of Leuven – RU Quantitative lexicology and variational linguistics

Dirk.Speelman@arts.kuleuven.ac.be

Stefan.Grondeelaers@arts.kuleuven.ac.be

Dirk.Geeraerts@arts.kuleuven.ac.be

<http://wwwling.arts.kuleuven.ac.be/qlvl>

Abstract. In this text we present “profile-based linguistic uniformity”, a method designed to compare language varieties on the basis of a wide range of potentially heterogeneous linguistic variables. In many respects a parallel can be drawn with current methods in dialectometry (for an overview, see Nerbonne and Heeringa, 2001; Heeringa, Nerbonne and Kleiweg, 2002): in both cases dissimilarities between varieties on the basis of individual variables are summarized in global dissimilarities, and a series of language varieties are subsequently clustered or charted using multivariate techniques such as cluster analysis or multidimensional scaling. This global similarity between the methods makes it possible to compare them and to investigate the implications of notable differences. In this text we specifically focus on, and defend one characteristic of our methodology, its profile-based nature.

Keywords. Aggregate methods, association measures, multidimensional scaling, profile-based analysis, variational linguistics

1. Introduction

The method discussed in this text, viz. profile-based linguistics uniformity, originated in a linguistic context that is not directly related to dialectology or dialectometry. It was first introduced in Geeraerts, Grondelaers & Speelman (1999), where it was used to aid the study of register variation and regional variation in Dutch. The regional varieties that were compared in that publication, as well as in more recent work by the same group, are the national varieties of Dutch, viz. Netherlandic Dutch and Belgian Dutch.

In this text we focus on the similarities between this method and a series of current approaches in dialectometry. Our claim is that there are enough commonalities to make a direct comparison possible, and that, consequently, criteria for evaluating alternative methods within the one field could, and perhaps should also be taken into consideration in other fields. We subsequently present a case study that was designed to evaluate one prominent feature of our method: its profile-based nature.

The structure of this text is as follows. There are two main parts. The first, smaller, part of the text is dedicated to the comparison of profile-based linguistic uniformity measurements and a series of current dialectometrical methods. After an elementary introduction to profile-based uniformity measurements (section 2) we compare this technique to a series of existing methods in dialectometry (section 3). The second, central part of the text focuses on the prominent feature of the methodology, its profile-based nature. First (section 4) we motivate this feature on theoretical grounds. Section 5 then focuses on a case study that was designed for the empirical verification of our profile-based point of departure: we compared a multivariate analysis of language varieties on the basis of a profile-based distance measure with two multidimensional analyses of the same data that were non-profile-based. In the conclusion (section 6) we summarize the findings of the case study and discuss the possible importance of these findings for dialectometry.

2. Presentation of the method

The main characteristics of the method at issue are perhaps best understood in the context of their initial purpose. In Geeraerts, Grondelaers & Speelman (1999), in which the method was introduced, the main research question tested was whether Belgian Dutch and Netherlandic Dutch converged in the period from 1950 to 1990, and whether that converging movement is due – as is generally thought - to changes on the Belgian side. In order to answer these questions, a measure based on one type of variation, viz. formal onomasiological variation (cf. Geeraerts, Grondelaers & Bakema 1994), was used as a basis for the comparison of the different synchronic and diachronic variants of Dutch for which data were collected.

2.1 PROFILES

In Geeraerts, Grondelaers & Speelman (1999) (and earlier also in Geeraerts, Grondelaers & Bakema. 1994), in which the focus was on the lexicon, *onomasiological variation* was said to occur when different terms are used to refer to the same entity (or to the same property, relation, action, state of affairs, etc.). *Formal onomasiological variation* was defined as onomasiological variation in which the use of different terms is not due to a different conceptual classification of the thing referred to, but rather to the use of different synonymous terms for referring to the same concept. An example of conceptual onomasiological variation – i.e. variation that is *not* formal - is the situation whereby the same entity is referred to once by means of a specific term (e.g. “car”) and once by means of the taxonomically hyperonymous term “vehicle”. An example of onomasiological variation that is formal is the situation whereby the same entity is sometimes called “car” and sometimes “automobile” (i.e. where alternative terms for the same concept CAR) are used.

The reason for using formal onomasiological variation as the basis for our comparison is not that it is considered to be the sole relevant type of information for looking at the convergence hypothesis mentioned above, but rather that it is seen as a convenient starting point with the methodological advantage that a distinction can be made between the frequency of terms as such and the frequency of concepts. The claim is not that the latter is of no importance for comparing language varieties, but rather that making the distinction results in a more clear-cut picture of the different levels of variation, and that it makes sense to first look at the former type of variation, because it is a simpler case. Of course, the downside is that formal variation is only one aspect of a much broader reality, but it is an aspect we claim is worth isolating.

Another point that should be made is that we do not claim that the distinction between formal and conceptual onomasiological variation is a matter of easy dichotomous classification. Conceptual differences may be subtle and determining when terms can be accepted to be formal variants of each other is often hard and rather like choosing a cut off point in a continuum. Still, we believe it is a workable procedure, be it a laborious one.

Geeraerts, Grondelaers & Speelman (1999) was followed by subsequent studies (notably Grondelaers, Van Aken, Speelman & Geeraerts, 2001) in which the exclusively lexical focus was given up in favour of inclusion of morphological and syntactical types of variation, approached in very much the same way as the original lexical materials. The definition of formal onomasiological variation was relaxed to cover all situations where alternative linguistic means, e.g. terms or constructions, are used to designate the same *concept* or *linguistic function*¹, without there being a clear semantic difference between the alternatives. A non-lexical example of formal onomasiological variation would be, for instance, the use of the genitive-“s” versus

the use of the preposition “of” to express a relation of possession; e.g. “my father’s house” versus “the house of my father”.

Let us, now that it is clear that formal onomasiological variation is used as the basis for the comparison of language varieties, go somewhat more extensively into the calculations. The basis for the calculations are individual **formal onomasiological profiles**, or **profiles** in short. A profile for a particular concept or linguistic function in a particular language variety is the set of alternative linguistic means used to designate that concept or linguistic function in that language variety, together with their frequencies (expressed as relative frequencies, absolute frequencies or both). Let us, by way of illustration, turn to a real-life example of an onomasiological profile (from Geeraerts, Grondelaers & Speelman, 1999, p.30 and p.157), viz. the profile for the linguistic function of “referring to an entity by using a term for the concept JEANS” in our sample of Netherlandic Dutch recorded in 1990 (the number of alternatives differs from profile to profile. In this example, there are only two alternative terms: “jeans” and “spijkerbroek”).

jeans	81 (70%)
spijkerbroek	34 (30%)

Table 1 - profile of JEANS in Netherlandic Dutch in 1990 (dataset N90)

Of these alternatives “jeans” occurs 81 times and “spijkerbroek” 34 times. In relative frequencies this is 70% versus 30%. As was already stated, the frequencies in this profile are based on the dataset N90 used in Geeraerts, Grondelaers & Speelman (1999). In other words, they are frequencies based on a sample².

2.2 DISSIMILARITY OF PROFILES

The next step is to compare profiles in different language varieties. For instance, the JEANS-profile from N90 (Table 1) can be compared to the JEANS-profile from B90, which is the Geeraerts, Grondelaers & Speelman (1999) dataset for Belgian Dutch in 1990. This profile is given in Table 2.

jeans	64 (97%)
spijkerbroek	2 (3%)

Table 2 - profile of JEANS in Belgian Dutch in 1990 (dataset B90)

So we need a measure for similarity, or, as we call it, **uniformity**, between profiles. Several association measures exist by means of which one could quantify the similarity or dissimilarity between the profiles in the different language varieties, and in fact, several are used and compared in our research. Here we present the two measures we believe to have been most useful in our studies so far, and that also were used in the case study that will be presented in section 5. For the purpose of this text we will present the measures as dissimilarity measures.

The first measure is that of **city block distance** D_{CB} . In order to normalize the two profiles being compared, the measure is calculated on the basis of the relative frequencies. The formula of city block distance is as follows. Given two language varieties V_1 and V_2 , given a linguistic function L , and given x_1 to x_n the exhaustive set of linguistics means to perform or express the linguistic function L , then we refer to the absolute frequency F of the usage of x_i for L in V_j with :

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

For instance, given table 2 above, we would have :

$$F_{B90,JEANS}(spijkerbroek) = 2 \quad (2)$$

Subsequently we introduce the relative frequency R :

$$R_{V_j,L}(x_i) = \frac{F_{V_j,L}(x_i)}{\sum_{k=1}^n (F_{V_j,L}(x_k))} \quad (3)$$

For instance, given table 2 above we would have :

$$R_{B90,JEANS}(spijkerbroek) = 0.03 \quad (4)$$

Now we can define the city block distance D_{CB} between V_1 and V_2 on the basis of their profiles for L as follows^{3 4}:

$$D_{CB,L}(V_1, V_2) = \frac{1}{2} \sum_{i=1}^n |R_{V_1,L}(x_i) - R_{V_2,L}(x_i)| \quad (5)$$

The division by two is for normalization. It maps the results to the interval [0,1].

City block distance is a straightforward descriptive dissimilarity measure that assumes the absolute frequencies in the sample-based profiles are 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 an alternative, or supplementary approach is needed. For this we use a measure that takes as its basis the confidence of there being an actual difference between two profiles, something different from chance: the **log likelihood ratio based dissimilarity measure D_{LLR}** . This time, unlike with D_{CB} , we look at the absolute

(rather than the relative) frequencies in the profiles we compare. More precisely, we consider the frequencies in a profile to be the sample of a random variable that has a multinomial distribution. Then, when we compare a profile in one language variety to the profile for the same concept or linguistic function in a second language variety, we use a log likelihood ratio test (as described in Dunning, 1993) to test the hypothesis that both samples are drawn from the same population. The test yields a value for the log likelihood test statistic $-2 \log \lambda^5$, which is known to have a χ^2 distribution, with $n-1$ degrees of freedom. On the basis of this log likelihood statistic a **p**-value can be calculated for the chance that the underlying distribution is the same for both profiles, in spite of the observed sample differences. Finally, we use $(1 - \mathbf{p})$ as our dissimilarity measure \mathbf{D}_{LLR}^6 .

The strength of this second dissimilarity measure is that it is sensitive to how much evidence there is for the assumption that there actually is an underlying difference between the two profiles, so that we can avoid overrating relative differences that are not based on significant absolute differences. This strength, however, is at the same time a weakness, since by merely manipulating the power⁷ of our design (by increasing the sample sizes) we can change the dissimilarities we measure. So this dissimilarity too has an implicit assumption: that we keep the power of our design under control.

Since both D_{CB} and D_{LLR} have their pros and cons, and since they have a somewhat different meaning (D_{CB} is a measure for how much the ‘structure’ of profiles differs, and D_{LLR} is a measure for how confident we are that there is a structural difference), we use both, as a supplement and an additional test for one another. One simple way of combining both measures is to use D_{CB} , filtered by D_{LLR} . What we mean by this is that the dissimilarity we use is D_{CB} if $D_{LLR} > 0.95$, and zero otherwise⁸.

2.3 DISSIMILARITY BASED ON SEVERAL PROFILES

After having compared the dissimilarities between language varieties on the basis of individual profiles, i.e. on the basis of individual concepts or linguistic functions, we can then calculate a summary dissimilarity measure on the basis of a whole set of profiles (concepts or linguistic functions), by taking the sum (or the average) of the dissimilarities for the individual profiles⁹.

In other words, given a set of linguistic functions L_1 to L_m , then the global dissimilarity D (whichever dissimilarity calculation is used) between two language varieties V_1 and V_2 on the basis of L_1 up to L_m can be calculated as :

$$D(V_1, V_2) = \sum_{i=1}^m (D_{L_i}(V_1, V_2)W(L_i)) \quad (6)$$

The W in the formula is a weighting factor. In the simplest case, this weighting factor is 1 for all profiles L_1 up to L_m (or $1/m$ if we want the scale of the differences to be comparable to that of the calculations for individual profiles). An alternative, which we call the (actual) weighted calculation, is to use weights to ensure that concepts which have a relatively higher frequency (summed over the different language samples) also have a greater impact on the uniformity measurement. In other words, in the case of a weighted calculation concepts (and linguistic functions) that are more common in everyday life and everyday language, are treated as more important¹⁰.

Note that we do not necessarily sum over all profiles. The set of linguistic functions taken together in a summary dissimilarity measure can either comprise the whole set of profiles being investigated, or it can be any subset. The latter option is useful, for instance, to test whether different types of linguistic functions render a different picture (e.g. the set of lexical profiles versus the set of non-lexical profiles).

2.4 CLUSTERING THE LANGUAGE VARIETIES

Once we have global dissimilarities between the language varieties, the next and final step then is to feed the dissimilarities into multivariate methods such as multidimensional scaling or cluster analysis, to investigate how the language varieties cluster. Of course, for this step to be informative, more than two language varieties must be compared, and all two-by-two dissimilarities between the varieties must be calculated first. An example of this technique, using multidimensional scaling, will be given in section 5.

2.5 MAIN CHARACTERISTICS OF THE METHOD

To conclude this section, we want to highlight the three features that are most typical of the introduced method. First, there has been a **lexical focus** in the application of the method. However, we consider this to be a non-essential feature of the method. In fact, one of the main questions we currently try to answer with the method is what are the differences (in terms of how language varieties cluster) between lexical variation and non-lexical variation. Second, and this is an essential feature, the method is **usage-based**. Rather than looking at structural information such as “which terms and constructions are part of a language system (in the sense that using them would be in agreement with the rules of the system), and which are not”, we look at usage-based information such as “which terms and constructions are actually being used, and what is the frequency of their use”. The reason for this particular choice is (a) that the latter information is more fine-grained, and (b) that we accept the added information (the frequencies) as a potentially important aspect of the characteristics of language varieties. In other words, we accept differences that merely consist of different frequencies to be potentially substantial.

The third feature of the method is its **profile-based** nature. Profile-based implies usage-based, but adds another criterion. The additional criterion is that the frequency of a word or a construction is not treated as an autonomous piece of information, but is always investigated in the context of a profile. The reasons for this criterion will be the topic of section 4.

3. Resemblance to dialectometrical approaches

In a.o. Nerbonne and Heeringa (2001) and Heeringa, Nerbonne and Kleiweg (2002) an overview is given of a whole range of methods for measuring the phonetic distance between dialects. The merit of these papers is that the methods being summed up are presented in a way that enables a comparison between them.

This is done by isolating the different steps in the methods, and by acknowledging that, in spite of the differences, all methods share the basic two steps of comparison and classification. The points on which methods differ can then be seen as ‘slots’ that can be filled in alternatively in an otherwise identical overall schema.

3.1 COMPARISON OF VARIETIES IN DIALECTOMETRY

In the first step of all approaches individual dialects are compared. Individual methods differ with respect to the **basic unit of measurement** they use in the comparison, i.e. the most elementary unit on the basis of which dissimilarities are being calculated. Sometimes these are *individual words*, sometimes they are *lists of words* and sometimes they are *written texts*.

Second, the exact **formula or algorithm** for calculating the dissimilarity may differ. Examples are *string edit distance*, *city block distance of feature bundles*, *Euclidean distance of feature bundles*, *Pearson correlation coefficient of feature bundles*, etc. It must be added though that some of the different algorithms listed here require

different representations of what in our (somewhat oversimplifying) schema is treated as one and the same unit of measurement. For instance, if we consider the case of individual words as units of measurement, then we see (at least) the following differences. In the case of classical ‘edit distance’ based calculations, words are represented as sequences of symbols (i.e. as character strings). But in the case of classical ‘feature bundle’ based calculations, words would be represented as sets of phonetic segments (and therefore as units in which certain phonetic features occur a certain number of times). And in more recent variants of ‘edit distance’ based calculations, words are represented as sequences of smaller units (the characters) that themselves are represented as sets of phonetic features.

A third slot that may be alternatively filled in concerns the question whether all dialects are compared directly, or whether each is compared to one particular point of reference, such as a standard dialect. This is **the distinction between direct and indirect comparison**. Fourth, if there are several units being measured in the comparison of two dialects, a choice may have to be made about **deriving a summary dissimilarity** from a set of dissimilarities. In this case, the obvious options are *summing* or *averaging*, and optionally *weighting the items* that are being summed or averaged over.

3.2 CLASSIFICATION IN DIALECTOMETRY

In the second step the dialects are classified, on the basis of the dissimilarities obtained in the first step. Here too different **classification methods** are available. Examples are *multidimensional scaling*, *cluster analysis* and *Kohonen maps*.

3.3 COMPARISON AND CLASSIFICATION IN GENERAL

The way methods are compared in Nerbonne and Heeringa (2001) and Heeringa, Nerbonne and Kleiweg (2002) can be applied to other fields too. Whatever the

varieties being compared - be it dialects, national varieties or register varieties - and whatever the type of linguistic variables being looked at - be it phonetic data, lexical data, syntactic data, or other types of information - the same steps of comparison and classification will have to be taken, and the same ‘slots’ will have to be filled. Our own research too can be fitted into the same schema. The basic units of measurements are profiles. The formulae for comparison are D_{CB} , D_{LLR} or a combination of both (cf. section 2.2). Comparison is direct: all two by two couples of varieties are compared. Summary dissimilarities are calculated as explained in section 2.3. And finally the classification methods being used are (several types of) cluster analysis and multidimensional scaling.

In the remainder of this paper we want to zoom in on the ‘slot’ *basic unit of measurement*, and argue for a perspective that has received little attention in dialectometry so far, viz. the perspective that there can be, and often is variation within individual language varieties, and that acknowledging this perspective has important implications for the design of the basic units of measurement.

4. Motivation for profile-based approach

If one accepts a usage-based approach, in the sense that one decides to look at what actually occurs in e.g. a corpus, and one moreover accepts that mere frequency differences may be important to detect differences between language varieties, then profiles are a straightforward choice for data representation.

For instance, if one accepts that the different relative frequencies in Table 1 and Table 2 are a sufficient ground for talking about a difference between Belgian Dutch and Netherlandic Dutch in 1990, even if they use the same words for the same concept, then Table 1 and Table 2 are a good starting point for detecting this difference.

However, incorporating frequencies does not need to take the form of a profile-based representation. One could also consider each *type*¹¹ in a corpus, together with its token frequency, to be an individual ‘basic unit of measurement’. And indeed this is often done in corpus linguistics.

4.1 ACTUAL MOTIVATION: AVOIDANCE OF THEMATIC BIAS

We believe that, at least for our purposes, there are two good arguments in favour of profiles and against the ‘each type in isolation’-alternative. The first is the **avoidance of thematic bias**. Token frequencies in a corpus could correlate with a formal onomasiological preference in the corpus, but they could also correlate with the thematic specificity of the corpus.

Consider the example where we have two equal-sized corpora V_1 and V_2 , each representing a language variety. Let us assume that type A occurs 200 times in V_1 and 500 times in V_2 , and that type B also occurs 200 times in V_1 and 500 times in V_2 . In an ‘each type in isolation’-analysis, we would conclude that both A and B are more typical of V_2 than of V_1 and that they both indicate a difference between V_1 and V_2 . However, suppose that we now add the information that A and B are two alternative terms for naming some concept X. For the sake of simplicity we assume that there are no other terms to refer to the concept X and that A and B always refer to this concept. So what differs between V_1 and in V_2 in the example is not a different formal onomasiological preference (within the profile, A and B have the same relative frequencies in V_1 and in V_2), but rather a different frequency of references to concept X. A profile-based calculation would signal identity here, and this is the desired result. Not only did we choose in advance to measure only formal onomasiological variation (cf. section 3.1), if we had not done so, interpreting the difference would be difficult, because what is being measured could well be a bias of the corpus design. Complete thematic control over a corpus is a nearly impossible enterprise.

4.2 POSITIVE SIDE EFFECT: AVOIDANCE OF REFERENTIAL AMBIGUITY

A second advantage of profile-based calculations is that they help **avoid referential ambiguity**. Once again, let us consider an example. Again we assume that we have two equal-sized corpora V_1 and V_2 , each representing a language variety. This time we assume that type C occurs 700 times in V_1 and 700 times in V_2 , so an ‘each type in isolation’-analysis would signal no difference. However, if we now add the information that C is ambiguous between two different linguistic functions, and that for instance in V_1 it is used 500 times to refer to concept Y, and 200 times to refer to concept Z, and that in V_2 it is used 300 times to refer to concept Y, and 400 times to refer to concept Z. In a profile-based calculation the different meanings or applications of C would be separated, and, depending of the complete profiles Y and Z, a difference would, or would not be signalled, which is the desired effect. We believe that this necessity for semantic disambiguating is a positive feature of the profile-based approach, especially in light of the fact that many existing approaches do not disambiguate semantically.

But of course, you do not need onomasiological profiles to perform semantic disambiguation. It can also be achieved in calculations that are based on individual variables (e.g. *types*), by splitting them up in such away that no variables are ambiguous (e.g. variable would be *types, restricted to one meaning*). And yet another way would be to work with semasiological profiles (in which *types* would be the basic units of measurement, and the specific meanings or uses of a *type* (if any) would be seen in relation to the overall usage of the *type*), but that approach would be vulnerable to the aforementioned problem of thematic bias in the corpora, since the overall frequencies of meanings or applications would not be taken into account.

5. Consequences of profile-based approach: a case study

The conclusion of section 3 is that problems are to be expected when non-profile-based calculations are used for the purpose of studying formal variation in isolation from conceptual variation. In this section we put this expectation to the test in a case study where we compare our profile-based calculations to two non-profile-based calculations. In this case study we compare 20 varieties, listed in the following table.

name	nr tokens	register	topic
chatRE	205560	chat-material	regional chat channels
chatVA	1182849	chat-material	varia
chatLE	1784084	chat-material	chat channel “Leuven”
chatVL	2736111	chat-material	chat channel “Flanders”
chatBE	1686571	chat-material	chat channel “Belgium”
useTE	2486797	usenet	technical topics
useSP	117195	usenet	sports
useSR	2376788	usenet	supra-regional topics
regL1	1561362	regional popular newspaper	(no differentiation)
regL2	1450968	regional popular newspaper	(no differentiation)
regL3	1666916	regional popular newspaper	(no differentiation)
regA1	1563799	regional popular newspaper	(no differentiation)
regA2	1504606	regional popular newspaper	(no differentiation)

regA3	1810548	regional popular newspaper	(no differentiation)
natRE	1945461	national popular newspaper	regional
natSP	427280	national popular newspaper	sports
natSR	518670	national popular newspaper	supra-regional interest
quaSP	994867	national quality newspaper	sports
quaTE	1431786	national quality newspaper	technical topics
quaSR	3607513	national quality newspaper	supra-regional topics

Table 3 - the subcorpora that are used in the case study

The first column shows the names of the sub-corpora, which are all parts of the ConDiv-corpus (Grondelaers e.a. 2000). The second column shows their size, in number of tokens¹². The next column describes the text types, and the last column specifies a further subdivision by topic, if any. All material is Belgian Dutch. We can do without Netherlandic material here, because in this case study we do not go into our underlying linguistic research questions (which are about the comparison of (a) the most ‘formal’, standard language in Belgium and in the Netherlands, and of (b) the level of divergence from that standard in less ‘formal’ text types in Belgium and in the Netherlands). Instead we focus on the method and we try to observe which differences in the data our method is sensitive to (and which it is not sensitive to). So in this case study the basic task for our method, and for the two methods we compare it to, will be to establish how much difference in language use there is in the 20 corpora that were listed above.

5.1 METHOD 1 : PROFILE-BASED

In the first method, which is our profile-based method, we base our calculations on the combined effect of the 10 lexical and 5 non-lexical profiles. The selection of these profiles, i.e. both the choice of a linguistic function and the generation of an exhaustive list of alternative designations, was based on dictionaries and on the literature¹³. The general procedure was to use the literature, but then again to be careful not to over-represent the “known classical examples of variation” in the selected profiles (in order not to overrate the differences between the text types). Once the profiles were chosen, the actual retrieval of the observations from the corpora was automated, in such a way that recall was maximized, sometimes sacrificing precision heavily. Finally the results were manually verified¹⁴.

The description of all profiles, as well as all frequency information, is made available on [<http://wwwling.arts.kuleuven.ac.be/cluster>]. Here we restrict ourselves to presenting two examples. An example of a lexical profile is the set of the terms “oom” and “nonkel”, which both refer to UNCLE. Of these two terms “oom” is assumed to be the “standard”-variant, whereas “nonkel” sounds more colloquial (the “oom”/“nonkel” profile is an example of register-related variation that is well known and cited in the literature). Retrieval was fairly easy for this example, and so was the manual verification (because neither of the words is ambiguous between different meanings). Nevertheless even here manual verification proved necessary. For instance, in newspapers “nonkel” sometimes occurred in a literal quote (for instance, when a journalist literally quoted an eye witness). Or sometimes “nonkel” occurred as part of a proper name (so that variation was in fact impossible). Such examples were excluded from the dataset. An example of a non-lexical profile is the set “moeilijk te”, “moeilijk om te”, “moeilijk van te” and “moeilijk voor te”, which are all ways to

express HARD TO + inf. (“moelijk” means HARD, and “te” means TO; the other words, “om”, “van” en “voor”, are optional, alternative complementizers).

For the calculations of the dissimilarities we used D_{CB} , filtered by D_{LLR} (cf. section 2.2), and we used a weighted sum to obtain a global dissimilarity (cf. section 2.3). Next, we used multidimensional scaling to plot the varieties in a two-dimensional space and in a three-dimensional space in such a way that distances are as close as possible to the dissimilarities we calculated beforehand. The result is shown below in Figure 1.

The top right plot shows the *stress*, a measure for “unaccounted variance”, of the best MDS-solutions for 1 up to 10 dimensions. It shows that the solution for 2 dimensions is reasonable (11.534%), and the solution for 3 dimensions is very good (4.446%). We have plotted both. The top left plot shows the solution for 2 dimensions. The other plots show the solution for 3 dimensions: the middle plot shows the three dimensions at ones. The bottom row shows reduced versions that retain two of the three dimensions. From left to right, we first see a plot with dimensions 1 and 2 (in other words, the middle plot seen from above), then a plot with dimensions 1 and 3 (in other words, the middle plot seen from the front), and finally a plot with dimensions 2 and 3 (in other words, the middle plot seen from the right).

In what follows we will primarily be looking at the two-dimensional solution (the top left plot). The three-dimensional solution has been added for completeness mainly.

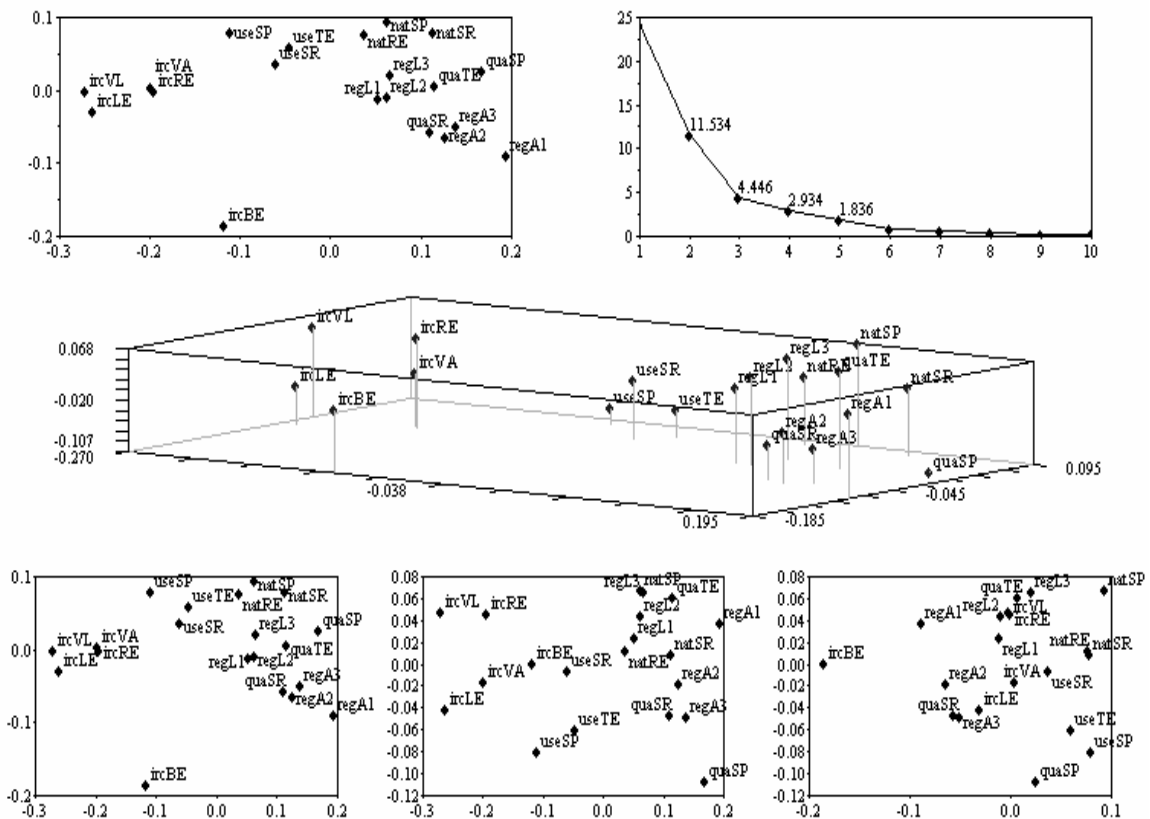


Figure 1- Results for method 1

We see that three main clusters can be discerned: irc material at the left (with ircBE either as an outlier or as a separate cluster), usenet material in the center, and newspaper material at the right. Within the big newspaper cluster we see that the members of the individual newspapers (regLX, regAX, natXX and quaXX) generally also cluster together well in smaller groups, albeit that quaSR is somewhat misplaced, by being so close to regA2 and regA3. Moreover, if we draw an axis from the lower lefthandside to the upper righthandside, we see a dimension that corresponds well with what we would expect to see on an axis that runs from “less formal” to “more

formal”: first of all, and most clearly, there is the general separation of Computer-Mediated Communication on the left, and newspaper materials on the right. Next, it makes sense to find the usenet materials right in the middle, being intermediate in formality between online chat-messages and editorially controlled newspapers. The redaction (editorial control) of a usenet message could indeed be seen as somewhere between that of a chat-message and that of a newspaper article. And finally, if we look carefully at the newspapers (following that axis that runs from bottom left to top right), we notice that in general the national papers are more at the top right end than the regional paper (however, there is the clear exception of quaSR).

5.2 METHOD 2 : PROFILE-LESS CALCULATIONS

In the second calculation we retain the same data of the first calculation, i.e. the 15 profiles, but rather than first calculating dissimilarities for the individual profiles, and then calculation the weighted sum, as is done in section 5.1, we now perform the calculation as if the data consisted of one large profile (which of course is not the case). The result of this reduction is that we loose accuracy at the level of thematic bias control. In theory the problem of referential ambiguity could also show up, but here this is not the case: given the profiles we have, no ambiguities emerge by giving up isolated profiles. For the calculations D_{CB} was used (cf. section 2.2). The result is shown below in Figure 2. The different plots in this figure are completely analogous to those in Figure 1 (with the addition of one graphical element: if the labels are too far away from the dot they belong to, which sometimes is necessary for reasons of space, we use a connection line between the label and the dot).

The top right plot shows that in terms of unaccounted variation the 2-dimensional solution is slightly worse than in method 1, but is still acceptable (stress is 12.875%), and that the 3 dimensional solution is very good, even better than in method 1 (stress

is 3.651%). As in the previous section, we will primarily discuss the 2-dimensional solution.

We see 3 clear clusters: irc at the left, usenet at the top (but in the company of natSP, which is in the ‘wrong’ cluster), and newspapers at the right. Within the newspaper cluster, the grouping of the individual papers is a bit worse than in method 1 (with the exception of regLX, which clusters remarkably well, none of the papers cluster perfectly), but then again it is not that much worse either. The axis we saw in method one, finally, also is much less clear. We rather see a ‘triangle’ of irc versus usenet versus newspapers.

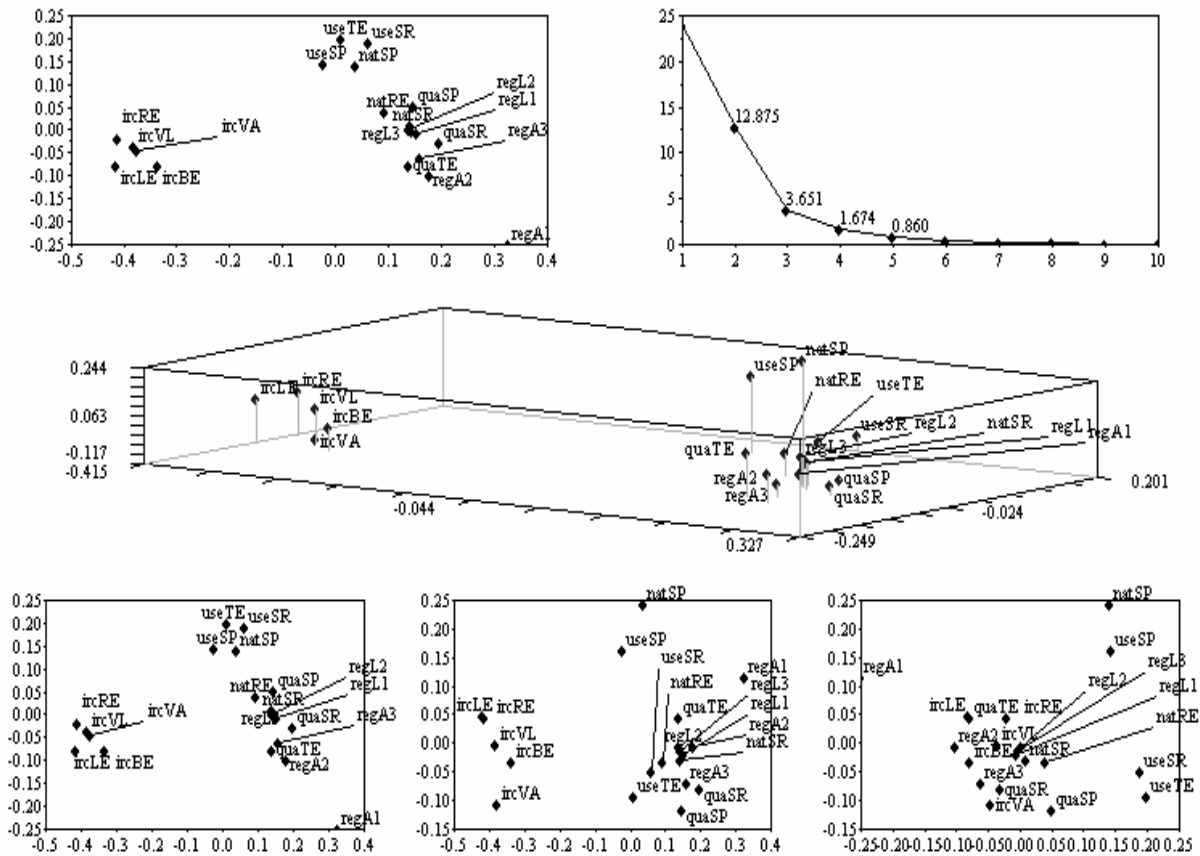


Figure 2- Results for method 2

In sum, the plot may be a bit less ‘elegant’ or ‘promising’, given the loss of any clear axis and given the less clear classification of the newspapers, but technically the clustering potential of the method, illustrated by the stress and by the three clear main clusters, is not worse than in method 1. In fact, this calculation even seems to generate clearer overall clusters by giving less weight to the smaller differences within these large clusters. These smaller differences seem to add up more easily in the profile-based calculations.

5.3 METHOD 3 : KEYWORDS

The calculation in 5.2 could be seen as no more than a variant of version of 5.1. Its closeness to 5.1 was useful to pinpoint the effect of grouping into profiles, but on the other hand 5.2 is not a typical case of ‘each type in isolation’-calculations. Therefore we include another calculation, which is more typical of the ‘each type in isolation’-approach. The strength of this approach lies in automation. The cumbersome process of manually selecting profiles and determining the corpus frequency of each synonymous designation contrasts sharply with the ease with which ‘each type in isolation’-calculations can be automatically applied to a whole corpus.

In this section we proceed as follows. When comparing two corpora, we go over the list of all (wordform) types in the corpora, and test for each type whether its frequency is significantly different between the two corpora. For this, we use the table given below, which is calculated for each type x . We calculate $F_1(x)$, which is the token frequency of x in corpus 1, and $F_2(x)$, its token frequency in corpus 2. We also use N_1 , the total number of tokens in corpus 1, and N_2 , the total number of tokens in corpus 2. The table is:

$F_1(x)$	$F_2(x)$
$N_1 - F_1(x)$	$N_2 - F_2(x)$

Table 4 - The contingency table for the keywords method

To find out whether the frequency of type x is significantly different in corpora 1 and 2, we treat both columns in that table as two samples of a random variable with binomial distribution, and test if the samples indicate a different underlying distribution, using a log likelihood ratio test (see Dunning 1993 for the exact calculations). If the difference is significant (at an error level of 0.05), we call type x a **keyword** (because its use is typical of one of the corpora, in comparison to the

other; in this particular context we don't care for which of the two it is a positive keyword; we just call it a keyword). The next step in the reasoning is that the more keywords show up in the comparison of two corpora, the more important the difference is between their lexicon. Following this reasoning we use this **number of keywords as the dissimilarity measure**. Note that, in contrast, to 5.1 and 5.2, this method only uses lexical information¹⁵ (although in the other methods too the information was predominantly lexical).

The results are given in Figure 3 below. This time the stress curve tells a different story. The 2-dimensional solution is unacceptable (26.302%). The 3-dimensional solution is no more than acceptable (10.668%). And it takes 5 dimensions to have a really good solution.

For the sake of easy reference we still stick to the 2-dimensional plot to discuss the results. However, the reader should also check with the other plots, since the 2-dimensional solution is simply too imprecise.

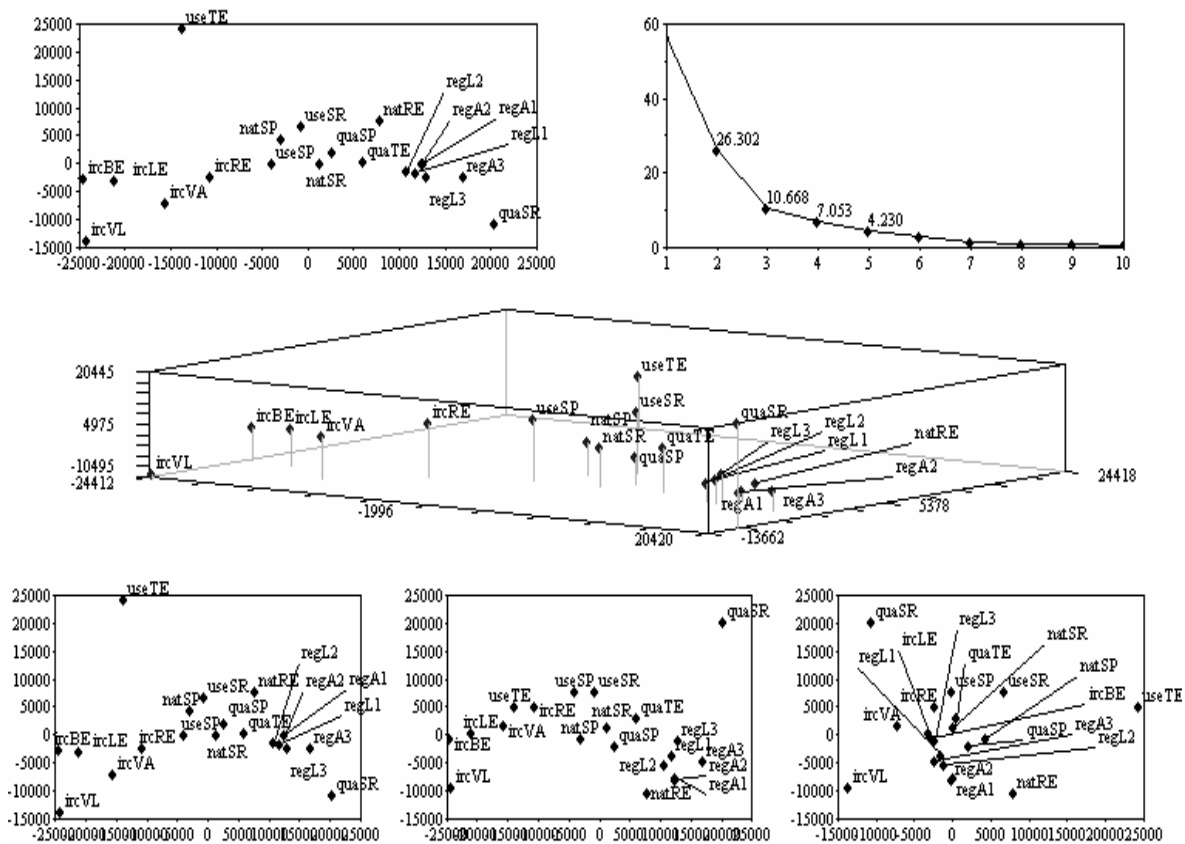


Figure 3 – Results for method 3

What we see here is a plot with less clear bigger clusters. We do have a horizontal axis with Computer-Mediated Communication on the left and “newspapers” on the right (and Usenet in the middle again), but the boundaries are rather fuzzy and at least one item is ‘out of place’ (natSP). Within the newspaper group regional papers cluster closely together. National newspapers, on the other hand, do not. The most remarkable result, however, is the proximity of natSP, quaSP and useSP in the middle of the plot. These are the sub-corpora with the topic “sports”. They cluster together as a result of the keywords method, because they have a common vocabulary, which is

not shared by the other corpora. In other words, the keywords-methods is very good at detecting thematic bias in the corpora. This is a downside for our purposes, because if we want to investigate if language use is different in sports material, then we want to know if language use is different apart from the obvious fact that sports-related topics are used more often, and it is clearly the latter that is (also) detected here. Therefore we conclude that this method is less suitable to our needs. On the other hand, the generation of keyword lists does yield a lot of information about the relation between the corpora, and does so with little effort. Which leads us to the following conclusions.

6. Conclusions

To sum up, we can come to the following conclusion.

First, the case study indicates that indeed the results from the two non-profile-based calculations are at least somewhat different from those of the profile-based method. Especially in the third method, the keywords method, traces of a problematic influence of topical bias were found. In general, the comparison seems to be positive for the profile-based approach. However, there are some comments to be made :

- The keywords method tested in this paper is typical of a fairly wide range of applications, and in that sense was a good starting point for a comparison such as the one presented here; however, there may be tougher competitors around; for instance the methods by Burrows (1992), that only uses function words, probably will be less vulnerable to the topical bias problem¹⁶.
- In general it is hard to compare the merits of the methods in a very thorough way, and the main reason is the lack of a Gold Standard in the field. For instance, there is no clear, detailed knowing in advance of precisely how many clusters should emerge in the MDS-plots. How should the newspapers

cluster exactly ? Therefore there is no clear procedure for evaluating the methods, and learning their merits becomes a slower, more incremental process.

- The labour-intensive nature of the profile-based approach is an obvious drawback. On the other hand, it should be acknowledged that non-profile-based and profile-based methods can complement each other: the latter approach can benefit from the ease with which an exhaustive list of significant frequency differences can be generated with the keyword method explained in section 5.3. More particularly, this complementarity can even be exploited by using the output of the keywords method as the input to the profile-selection step. But for this method to be effective, methods will have to be applied¹⁷ to prune the overly extensive output of the keywords method.

The second conclusion, which is more of a suggestion than a conclusion, and which also links up with the more theoretical part of the paper, is that the current dialectometric methods could profit greatly from opening up towards usage-based methods in general and profile-based methods in particular. In all fairness, it should be though added that due to its high demands at the level of the data collection the applicability of the profile-based method decreases as the number of varieties one compares increases.

Acknowledgements

The research reported on in this paper was supported by VNC-grant 205.41.07 3 as well as by OT-project OT 01/05.

Notes

¹ The terms *concept* and *linguistic function* are used in this paper with no theoretical connotation whatsoever. They respectively refer to the semantics of a term and of a construction, in a very general way. Sometimes, for brevity, *linguistic function* is used as a hyperonym for both the semantics of terms and of constructions.

² One can also use the term ‘profile’ to refer to population-based (relative) frequencies. However, we will not introduce a notational difference between sample-based and population-based profiles, and assume that it is clear from the context which level is intended.

³ The city block distance D_{CB} dissimilarity measure is the complement of the basic similarity measure U that is used in Geeraerts, Grondelaers & Speelman (1999). U , which equals $(1 - D_{CB})$, is defined as :

$$U_L(V_1, V_2) = \sum_{i=1}^n \min(R_{V_1, L}(x_i) - R_{V_2, L}(x_i))$$

⁴ One obvious alternative to city block distance would be Euclidean distance. Euclidean distance and city block distance gave comparable results in the study reported here.

⁵ We refer the reader to Dunning (1993) for the exact calculation.

⁶ One possible alternative to the LLR-based dissimilarity measure would be a dissimilarity measure based on the Fisher Exact test. We did not test that option yet.

⁷ We use the term power as it is used in *inferential statistics*.

⁸ We want to make two remarks, one technical and one methodological. The technical remark is that for the sake of further computation (the MDS-techniques) we avoid zero-dissimilarities and instead use a very small constant, close to, but different from zero, if $D_{LLR} < 0.95$. The methodological remark is that an alternative approach for deriving a combined measure would be to weight D_{CB} with D_{LLR} in a more continuous way, instead of using a cut off point.

⁹ Of course there are several alternatives. One straightforward alternative to taking the sum or the (arithmetic) average of the differences for the individual profiles, would be to use the product of these differences, or the geometric average (i.e. the m -th root of the product, m being the number of profiles).

¹⁰ An example of using weighted calculations can be found in Grondelaers e.a. (2001), p. 183-184.

¹¹ *Type* could either refer to *type of wordform* or to *type of lemma*, depending on the context. Both wordforms and lemmata can be used for what we describe. Therefore we will in general not make the distinction in this text, unless it is necessary.

¹² For the extraction of the corpus data the tool *Abundantia Verborum* was used [<http://www.ling.arts.kuleuven.ac.be/genling/abundant>]. For the log likelihood ratio tests, the frequency list tool in *Abundantia Verborum* was used. For the multidimensional scaling analysis, the functions *cmdscale* and *isoMDS* in the statistical environment R were used [<http://www.r-project.org/>]. The plots were generated with *SCILAB* [<http://www.rocq.inria.fr/scilab/>].

¹³ The selection of the profiles of course is a crucial step, and it is clear that the choice of the profiles, in combination with the decision of how to weight the profiles in the calculations of global dissimilarities, has important implications for the results of the classification. For instance, one can either overstress, or flatten out, the idiosyncrasies of certain corpora. Anyhow, it is clear that one should somehow try to find profiles that are representative of the range of variation one wants to study. Basically, we adopt three different strategies. In earlier studies we selected a limited number of semantic fields (clothing terms and football terms), and covered those in depth. For the moment we are preparing a large scale collection of variables, based on whatever can be found on the literature. Thirdly, we also explore the application of methods for automating the process of profile selection (cf. the next note).

¹⁴ One may wonder if an objective list of rules can be constructed to assign terms to profiles, or to subdivide polysemous terms into monosemous usage of these terms. The answer must be a practical one: the level of accuracy and consistency we eventually obtain is that of a good dictionary; that is, our profiles reflect the heuristic and analytic competence of experienced lexicographers. One might also wonder to which extent the method of profile selection can be automated. We are currently experimenting with a cyclic search procedure that combines (a) keywords methods (cf. section 5.3), as a bootstrap mechanism that finds new potentially interesting terms, and (b) synonym detection methods for instance by means of latent semantic analysis, (Landauer and Dumain, 1997). No matter, unfortunately, how helpful in determining an initial selection of relevant profiles, fully automated processes will probably remain error prone: the final decision, hence, will continue to be the linguist's. Of course, yet another perspective is that it could also be interesting to investigate how such automatically generated profiles would cluster the data, even if they are not perfect from a human point of view.

¹⁵ This bias towards lexical data is related to our choice to base the calculations on the frequency list of wordform types. We could, however, do a similar calculation on the basis of e.g. frequency lists of letter bigrams, trigrams, etc., or word bigrams, trigrams, etc., which would make other levels of information accessible to the technique.

¹⁶ On the other hand, such a method may be less compatible with the ambition of the profile-based method to cover a wide range of potentially heterogeneous variables, which makes a direct comparison of methods less straightforward.

¹⁷ The problem is that the comparison of large corpora generates very long lists of keywords. The solution we envisage is not to build keyword list on the basis of the comparison of two corpora, but rather to build lists of items that are ‘stable lexical markers’ in the comparison of two sets of corpora. For instance, a stable lexical marker for the set A of all irc corpora, when compare to the set B of all other corpora, would be a term that consistently is a positive keyword in (almost) any comparison of an element of A with an element of B. This method seems to efficiently point to interesting terms.

References

- Burrows, J. F. (1992). Not unless you ask nicely: The interpretative nexus between analysis and information. *Literary and Linguistic Computing* 7(2), pp. 91-109.
- Dunning T. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics* 19(1), pp. 61-74.
- Geeraerts D., Grondelaers S., Bakema P. (1994). *The structure of lexical variation. Meaning, naming and context*. Berlin: Mouton de Gruyter. 270 p.
- Geeraerts D., Grondelaers S., Speelman D. (1999) *Convergentie en divergentie in de Nederlandse woordenschat. Een onderzoek naar kleding- en voetbaltermen*. Amsterdam: Meertensinstituut. 172 p.
- Grondelaers S., Deygers K., van Aken H., van den Heede V., Speelman D. (2000). Het ConDiv-corpus geschreven Nederlands. *Nederlandse Taalkunde* 5, 356-363.
- Grondelaers S., van Aken H., Speelman D., Geeraerts D. (2001). Inhoudswoorden en preposities als standaardiseringsindicatoren. De diachrone en synchrone status van het Belgische Nederlands. *Nederlandse Taalkunde* 6, 179-202.

- Heeringa, W., Nerbonne J., Kleiweg P. (2002) Validating Dialect Comparison Methods. In W. Gaul and G. Ritter (eds.): *Classification, Automation and New Media. Proceedings of the 24th Annual Conference of the Gesellschaft für Klassifikation*. Heidelberg : University of Passau, Springer, pp. 445-452.
- Landauer, T.K., Dumais, S.T. (1997). A solution to Plato's problem : the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104, 211-240.
- Nerbonne J., Heeringa W. (2001) Computational Comparison and Classification of Dialects. *Dialectologia et Geolinguistica* 9, pp. 69-83.