# Measuring the lexical distance between registers in national varieties of Dutch

*Tom Ruette, Dirk Speelman, Dirk Geeraerts*

**Abstract**

From a lectometric point-of-view, distances between language varieties can be quantified by means of aggregating the dissimilarity in the behavior of linguistic characteristics. Given the fact that Dutch has evolved to become a pluricentric language, a sociolectometric approach can be applied to measure the distance between varieties of Dutch. The current paper builds upon Geeraerts *et al.* (1999) and Speelman *et al.* (2003) to measure these distances. Just like Geeraerts *et al.* (1999), we adopt a focus on lexical variation. We compare two ways of measuring this lexical distance, extending the work of Speelman *et al.* (2003).

***Keywords***: lectometry, Dutch as a pluricentric language, lexical distance, feature aggregation

## 1 Introduction

In a lectometric study, the behavior of many linguistic characteristics in a number of varieties or *lects* is aggregated in order to make a general assertion on the structure of the lects under observation. In dialectometry, the characteristics that appear in dialect atlases are aggregated to make a classification of dialects (e.g. Seguy, 1971). Sometimes, dialectometricists also use corpus-based data (e.g. Szmrecsanyi, 2011). In sociolectometry, corpus-based frequencies of lexical (Geeraerts *et al.* , 1999; Soares da Silva, 2010) or syntactic (Speelman *et al.* , 2003) characteristics have been aggregated to study the convergence and divergence of national varieties or register variation. In stylometry, characteristics are aggregated to position registers on functionally interpretable dimensions (e.g. Biber, 1988), and to compare these dimensions across languages (e.g. Biber, 1995).

The sociolectometric perspective of Geeraerts *et al.* (1999) and Soares da Silva (2010) already showed the value of a lectometric approach to pluricentric languages (Clyne, 1992). Diachronically, this perspective quantifies a convergence or divergence between the centers of the pluricentric language. In our study, we focus synchronically on the lexical distances between subcorpora that are representative for registers in two national varieties of Dutch (the pluricentric character of Dutch as used in Belgium and The Netherlands has been discussed in Clyne (1992, p. 71) and Auer (2005, Section 5.2)), and also by Geeraerts in this volume. In our focus

on the lexicon, we follow the *profile-based* approach of Geeraerts *et al.* (1999). This approach aggregates lexical alternation variables (so-called *profiles*), in which the variants are words that are (claimed to be) semantically identical. An example of a profile would be the set of synonymous words {`subway`, `underground`, `tube`}, to which we then refer as SUBTERRANEAN PUBLIC TRANSPORT.

There are a number of metrics to measure distances between subcorpora. In Speelman *et al.* (2003), a comparison was made between a keyword-based distance metric, a profile-less metric and a profile-based metric, while dealing with registers of Belgian Dutch. This study will compare a state-of-the-art document categorization metric to a profile-based metric, while dealing with register and topic variation in both Belgian and Netherlandic Dutch. Indeed, corpus-based lect categorization is in essence not different from text categorization, so it may be hypothesized that state-of-the-art text categorization methods outperform the profile-based method.

Unlike Speelman *et al.* (2003), which focused on both lexical and syntactic characteristics, our focus lies entirely on the lexicon. In order to decide which metric works best — the document categorization metric, or the profile-based metric — , our list of lexical items was constructed in such a way that we could predict the patterning of the subcorpora. It will be shown that the profile-based metric is less sensitive to register and topic than the state-of-the-art categorization metric. This result confirms the findings of Speelman *et al.* (2003).

The remainder of this paper is structured as follows. In Section 2 we show how the two metrics work and how they differ in the calculation of distances. Section 3 introduces the subcorpora that are representative of the national varieties and registers that we are studying. This section also introduces the profiles that will be aggregated to measure the distance between the subcorpora. The results of both methods are presented in Section 4. In the final part (Section 5), we summarize the findings in a conclusion, and we enumerate steps to be taken in further research.

## 2 Method

The two distance metrics that are being compared in this study fit in a larger method that is known as the Vector Model (Baeza-Yates & Ribeiro-Neto, 1999, p. 27), which is visualized in Fig. 1. In a Vector Model, the texts of a corpus are represented by a list of characteristics. The frequency with which these characteristics occur in the texts are stored in a table. A row of this table contains all the characteristic frequencies of a single document. The generation of this table is shown in Fig. 1 by moving from step 1 to step 2. A made-up example of this table can be found in Table 1.

| *Profile* | SUBTERRANEAN PUBLIC TRANSPORT | | | ... |
|---|---|---|---|---|
| *Variants* | `subway` | `underground` | `tube` | ... |
| Am.Eng | 12 | 1 | 1 | ... |
| Br.Eng | 2 | 14 | 8 | ... |

**Table 1:** *Made-up example of observation table*

A row of this table is then conceived as a *vector*. A vector is a list of $n$ numbers that represent the coordinates of a point in an $n$-dimensional space. Usually, one represents the vector as a line through the origin of the space and the point coordinates, as can be seen in Fig. 2. It is assumed that spatial closeness corresponds to conceptual similarity. Sometimes, this similarity is taken to be semantic in *Word*

*Space Models*, or thematic in text categorization. In lectometry, the spatial closeness of the points (which represent the subcorpora) is assumed to be similarity in language use. That is why we can construct a square distance matrix that compares all possible pairs of subcorpora (varieties), as shown in step 4 of Fig. 1.

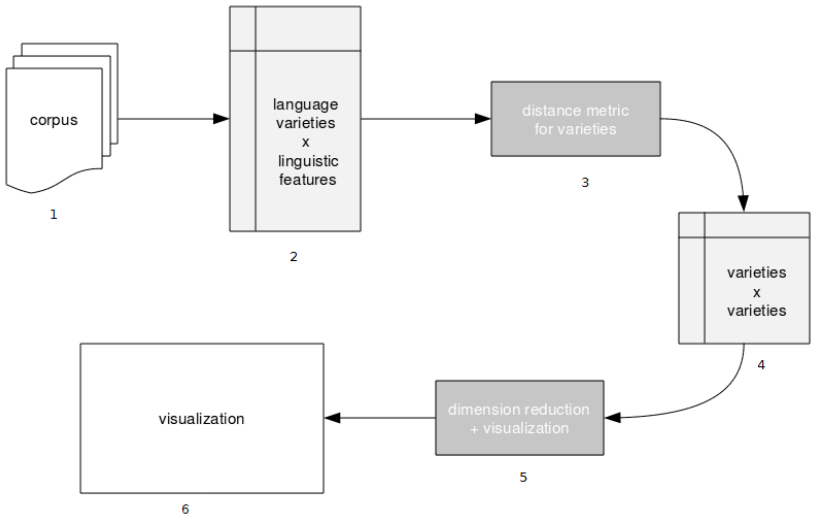


**Figure 1:** *Modular character of corpus-based sociolectometry*

Whereas a Vector Model would take the distance matrix from step 4 as input for a cluster algorithm in order to generate a categorization, we apply a dimension reduction technique (non-metric Multidimensional Scaling, cf. Cox & Cox (2001, Chapter 3)) that allows then for a visualization of the subcorpora.

These six steps make up the blueprint of the lectometric methodology. The current paper experiments with two distance metrics in step 3 to see how they influence the final visualization. Below, we first present the profile-based distance metric (Section 2.1) and then the state-of-the-art document categorization method (Section 2.2).

## 2.1 Profile-based method

The profile-based method was first introduced in Geeraerts *et al.* (1999) and was then used to study the convergence and divergence between Belgian and Netherlandic Dutch in two lexical fields. The *uniformity* metric that was proposed in Geeraerts *et al.* (1999) is equivalent to the slightly adapted City-Block distance presented in Speelman *et al.* (2003). For ease of formalization, we base our introduction of the profile-based distance metric almost entirely on Speelman *et al.* (2003, Section 2.2 and 2.3).

Given two subcorpora $V_1$ and $V_2$ that represent the varieties under scrutiny, a profile $L$ (e.g. SUBTERRANEAN PUBLIC TRANSPORT) and $x_1$ to $x_n$ the exhaustive list of variants (e.g. {`subway`, `underground`, `tube`} in the profile $L$, then we refer to the absolute frequency $F$ of the usage of $x_i$ for $L$ in $V_j$ with[1]:

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

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

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 $L$ 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 measure that assumes the absolute frequencies in the sample-based profile to be large enough for the relative frequencies to be good estimates for the relative frequencies in the underlying population-based profiles. 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: the Fisher Exact based dissimilarity measure $D_{FE}$. This time, unlike with $D_{CB}$, we look at the absolute frequencies in the profiles we compare. When we compare a profile in one language variety to the profile for the same concept in a second language variety, we use a Fisher Exact test to test the hypothesis that both samples are drawn from the same population. We use $(1-p)$, with the $p$-value from the Fisher Exact test, as our dissimilarity measure $D_{FE}$. $D_{FE}$ is then used as a filter for $D_{CB}$. We set the dissimilarity between subcorpora at zero if $D_{FE} < 0.95$, and we use $D_{CB}$ if $D_{FE} > 0.95$.[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.

## 2.2 State-of-the-art categorization method

In text categorization, part of the task is to measure the similarity (or distance) between texts. In a corpus-based study, lects are represented by texts, and therefore

Section 2.2).

[2]To employ the $D_{FE}$, 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.

it is not unthinkable that a state-of-the-art text categorization method is successful in finding structure among lects. However, unlike the profile-based distance metric, the text categorization method does not take the onomasiological variation within one profile into account. In other words, a text categorization approach ignores the *Profile*-line in Table 1. Instead, the mere absolute frequency of every individual variant is used. The similarity metric that is used is the *cosine similarity measure*. This similarity metric can be transformed into a distance metric by subtracting the outcome from 1 (dissimilarity = 1 - similarity).

More information on the cosine metric can be found in Baeza-Yates & Ribeiro-Neto (1999). Basically, the metric interpretes the angle between two vectors (the line through the origin and the point) as a similarity indication: the smaller the angle, the higher the similarity.[4] The cosine of an angle is maximal (= 1) when two vectors are perpendicular (which is the furthest they can be apart in a Vector Model), and minimal (= 0) when two vector coincide.

A made-up example in a two-dimensional space, i.e. with two words as features, for two text types might make this rather abstract introduction more clear. Given two varieties "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 vector space in Fig. 2. 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. Therefore, two vectors can be drawn through the origin and the position of the varieties, yielding an angle, for which the cosine can be calculated. A small angle implies high similarity, and will yield a high cosine value; a large angle implies low similarity, and will yield a low cosine value.

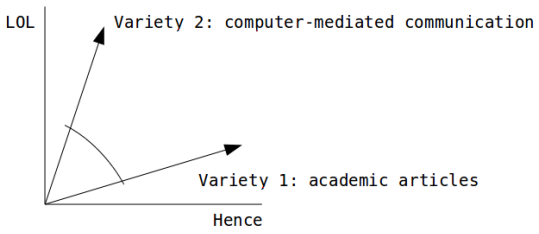


**Figure 2:** *2 Dimensional example of Vector Model*

Formally, given two subcorpora $V_1$ and $V_2$ that represent the varieties under scrutiny, represented by the respective vectors $\vec{x}$ and $\vec{y}$ from the table in step 2 of Fig. 1, we calculate the distance between subcorpora by means of Equation 5.

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

[4] Because the cosine metric works with the angle between the vectors and not the coordinates of the points, the use of absolute frequencies is acceptable.

# 3 Corpus and characteristics

The current study will implement these two distance metrics in the overall methodology of lectometric studies to measure the distance between registers in two national varieties of Dutch (Belgian Dutch and Netherlandic Dutch). As we perform a corpus-based study, the textual material that was collected for analysis is introduced in Section 3.1. Section 3.2 introduces the profiles that are counted in the texts of the corpus for aggregation.

## 3.1 Corpus

In our corpus, we incorporate samples from the two national varieties, taken from two registers (quality newspapers and Usenet), and from two topics (politics and economy). Topical control is important because our linguistic characteristics are lexical, and therefore highly sensitive to topical variation. 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. We assume that Usenet contains lexical examples of Colloquial Belgian Dutch (Geeraerts, this volume). 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 dimensions, 8 combinations are possible. These combinations, e.g. Belgian quality newspapers on economy (abbreviated as `qnp.be.e`), will be seen as the language varieties, for which we will calculate the pairwise distances. However, to increase the number of data points, we divided every variety into two equal parts (abbreviated as `qnp.be.e.0` and `qnp.be.e.1`). In total then, we counted the frequencies of the linguistic characteristics which we introduce below, in 16 subcorpora, bringing all this information together in the variety-by-feature matrix presented above in step 2 of Fig. 1.

## 3.2 Characteristics

Our goal is to compare the two distance metrics in light of a sociolectometric study on a pluricentric language. The input features are lexical items, which will be seen as alternation variables in the case of the profile-based distance metric, and which will be seen as individual features in the case of the state-of-the-art categorization distance metric. The lexical items are derived from the "Referentiebestand Belgisch Nederlands" (Martin (2005), Reference List of Belgian Dutch, abbreviated as

"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 (compare to category 7 of Colloquial Belgian Dutch markers in Geeraerts, this volume). The list contains about 4000 words, ranging from colloquial items, over culturally linked (e.g. Belgian institutes) to register-specific and freely varying words. As an example, a small selection of items is listed in Table 2. For each Belgian Dutch item, the list provides an alternative from general Dutch, or 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 word. If we restrict the list to items that consist of a single word — 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. Indeed, expressions and multi-word-units may be distributed over the sentence because of syntactic constructions. Here, too, 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.

| *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 2:** *Selected examples from the RBBN*

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 national pattern. Indeed, even the non-national variation which is present in the list (colloquialisms, culture-specific items, etc.) is still embedded in the Belgian Dutch point-of-view of the RBBN. Therefore, we expect the results of our method to show a clear distinction between the two national varieties, and subsequently some separation in the registers.

# 4 Results

The two approaches to measuring the distance between varieties are embedded in a methodology that presents these distances in a visualization (see the last step of the method in Fig. 1. The advantages of this method are on the one hand that one has a visual representation of the relative distances between the varieties, and on the other hand that one can easily interpret the clustering of the varieties along the dimensions of the plot, i.e. from left to right, from bottom to top. To aid the interpretation of the visualizations below, we assigned grey values and font characteristics to the 16 varieties (see Section 3.1) in this study. The two national varieties are separated by their grey value: Belgian varieties are in black, Netherlandic varieties are in grey. The distinction between quality newspapers and usenet articles is made by the boldness of the font: Usenet articles are in a bold font, while quality newspaper articles are in regular font. We do not highlight the distinction between political and economical varieties, as this difference is not very outspoken. However, we provided

descriptive label (p for politics, and e for economy) in the visualizations for further scrutiny.

Before we move on to the presentation of the results, a final, rather technical remark has to be made concerning "stress". The "stress" value of a visualization, always reported below the plot, is an indicator of the quality of the visualization. Note that the Multidimensional Scaling procedure is in essence a reduction of the data, which almost always implies that some data is lost and that the lower dimensional solution is merely an estimation of the original data. The stress value grasps this difference between original and estimated data and should not be larger than 10% - 15% to be acceptable. A stress value that is too high can be remediated by finding a solution with more dimensions. Every added dimension reduces the error (and thus the stress value) and approaches the original data more. Therefore, one can often interpret the first dimension as more important than the second. That is, of course, if adding a second dimension reduces the stress of a one-dimensional solution considerably. This reduction can be consulted in a screeplot.

## 4.1 Results of profile-based method

We first look into the results of the profile-based approach. To the selected Belgian Dutch items on the RBBN list, we add the knowledge which alternatives are referentially equivalent General Dutch words. In other words, we introduce 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 visualization of the profile-based generated variety-by-variety distance matrix, as can be seen in Fig. 3.

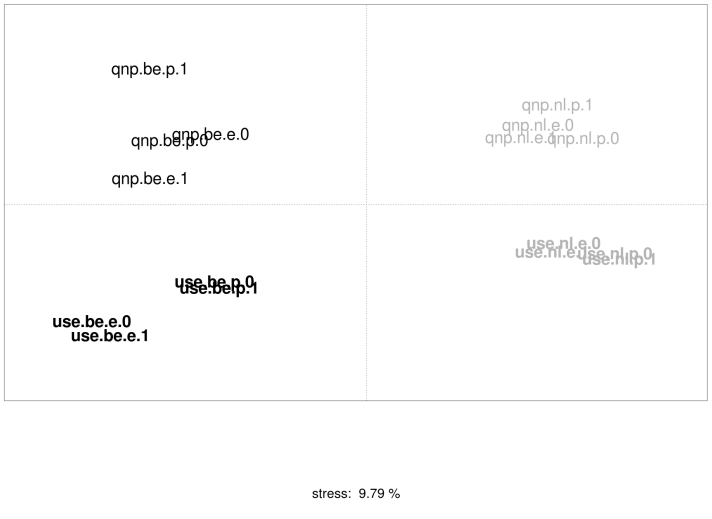


**Figure 3:** *Linguistic distance between subcorpora (profile-based)*

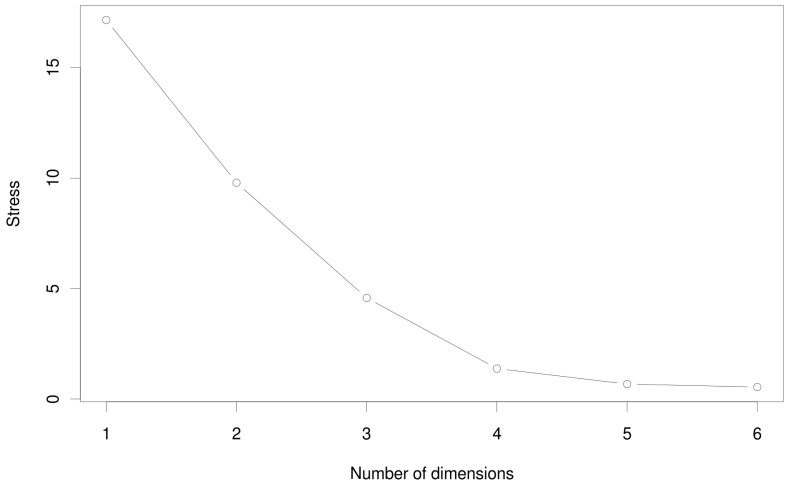


**Figure 4:** *Screeplot for non-metric MDS solution (profile-based)*

The screeplot in Fig. 4 shows a stress difference of about 7% between a one-dimensional and a two-dimensional MDS solution. Therefore, we first interpret the horizontal dimension as it represents the most important variation. In this case, the profile-based approach makes a distinction between Belgian varieties (black font) and Netherlandic varieties (grey font) on the first dimension. The grey zero-line divides the two countries perfectly. The vertical dimension makes a distinction between quality news papers (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 clustering of the varieties, although there is only clear separation of the topics in the Belgian Usenet. The range of Belgian register variation is also somewhat larger than the Netherlandic range. 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.

## 4.2 Results of categorization method

Now, we present the results of the state-of-the-art categorization approach, which uses the cosine similarity metric. We take the RBBN items (and the alternatives) as individual features and remove the knowledge of semantic identity between items. If we calculate the similarities (and consequent distances) with these input features between the varieties in our dataset, and then produce the visualization with MDS, we get the plot in Fig. 5.

If we create a screeplot (Fig. 6) to show us how much stress difference there is between 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

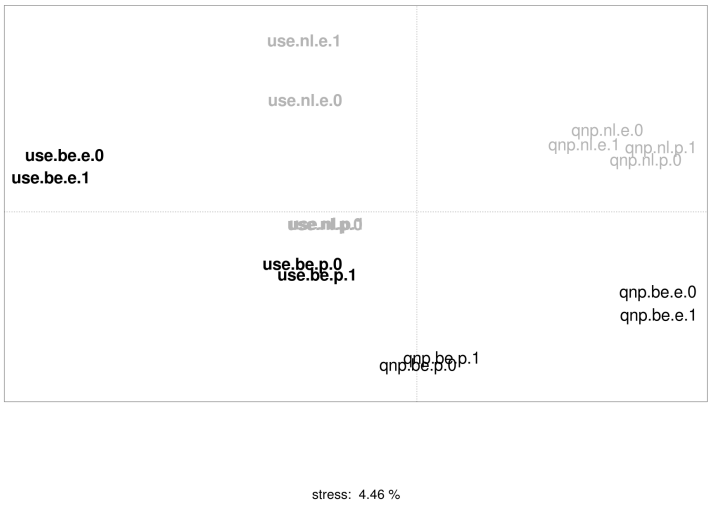


**Figure 5:** *Linguistic distance between subcorpora (cosine)*

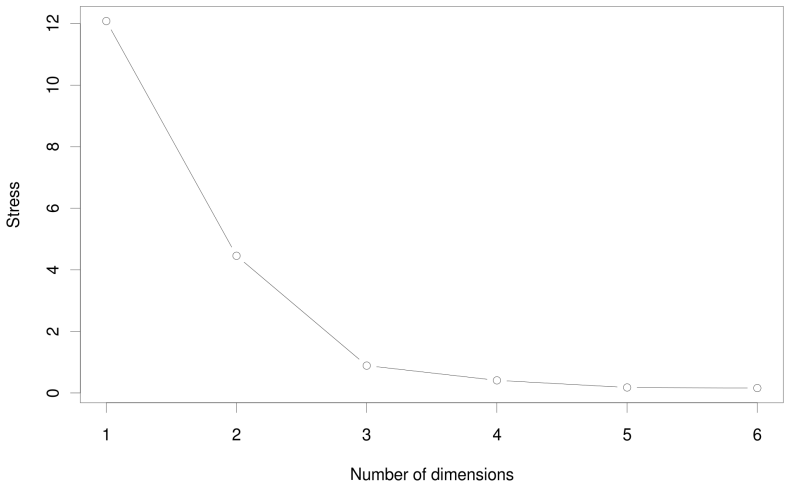


**Figure 6:** *Screeplot for non-metric MDS solution (cosine)*

dimension represents more "important" variation than the second dimension.

In Fig. 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, as well, except for the quality newspapers from The Netherlands. All in all, the categorization approach yields somewhat unclear clusters of varieties and an unexpected promotion of register variation as the most important variation in the input features.

# 5 Conclusions and further research

In the conclusion to this paper, we would like to make two points. The first point deals with the application of lectometric methods in research on pluricentric languages, the second point summarizes the findings on the comparison of the two distance metrics. Finally, we sum up a number of problems that will be addressed in further research.

From the above comparison of methods, but also from the previous socio- and dialectometric studies, it is clear that the quantification of linguistic distances with lectometric methods is insightful. In an objective way, the structure of lects is revealed on the basis of many linguistic characteristics. This approach can easily be extended to the study of pluricentric languages, as shown in the current study, Geeraerts *et al.* (1999) and Soares da Silva (2010). Questions in the area of pluricentricity revolve often around measurements of linguistic distance: e.g. the distance between dominant and non-dominant varieties might be an argument for speaking about separate languages, standardization progress can be measured by quantifying the distance between nation-internal varieties and a prestige variety.

On the methodological level, the current study has replicated and extended the work of Speelman *et al.* (2003). The comparison of two distance metrics has shown, just like Speelman *et al.* (2003), that a profile-based approach reduces thematic bias viz. a state-of-the-art text categorization method. Moverover, the results of the profile-based approach linked up better to the expected pattern (national variation is more important than register variation) of the specifically lexical input features in comparison to the categorization method.

Finally, we would like to point to three problems that are to be tackled in further research. First of all, the bias in the input features needs to be removed. Indeed, the lexical variables were picked from a reference list of Belgian Dutch, which caused our results to be primed for national variation. To overcome this problem, we should generate a list of lexical variation in a bottom-up, preferably automatic fashion. For this, advanced methods are being developed in a branch of Computational Linguistics, which bears the name Distributional Semantics (e.g. Sahlgren, 2006). Second, as pointed out by Lavandera (1977) and Labov (1978), lexical items

are polysemous and this polysemy should be overcome by detailed analysis of the context in which they appear. However, our current approaches do not control the polysemy of the lexical items. In future research, we will fall back on further advances in Distributional Semantics to address this problem. And third, the current lectometric approaches do not allow insight in the behavior of the individual characteristics. Their behavior is obscured by the aggregation step. In future research, the application of a more advance Multidimensional Scaling method (three-way MDS) will help to overcome this.

## 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.

Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge University Press.

Biber, Douglas. 1995. *Dimensions of register variation: A cross-linguistic comparison*. Cambridge University Press.

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.

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

Labov, William. 1978. Where does the linguistic variable stop? A response to Beatriz Lavandera. *Working papers in sociolinguistics*, **44**, 5–22.

Lavandera, Beatriz. 1977. Where does the sociolinguistic variable stop? *Working papers in sociolinguistics*, **40**, 6–24.

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

Sahlgren, Magnus. 2006. *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Ph.D. thesis, Department of Linguistics, Stockholm University.

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

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

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