Frequency effects in lexical sociolectometry are insubstantial

Tom Ruette (KU Leuven), Katharina Ehret (University of Freiburg) and Benedikt Szmrecsanyi (KU Leuven)

Abstract

This contribution investigates frequency effects in lexical sociolectometry, and explores by way of a case study variation in written English as sampled in the well-known Brown family of corpora. Lexical sociolectometry is a productive research paradigm that is concerned with studying aggregate lexical distances between varieties of a language. Lexical distance quantifies the extent to which different varieties use different labels to describe the same concept. If different labels are used in different varieties, then this will increase the lexical distance between the varieties We aggregate over many different concepts, in order to make generalizable claims about the distance between varieties, independently of a specific concept. Our central question is, "When generalizing across concepts, does concept frequency play a role in the aggregation?" To answer this question, we examine three types of frequency weighting (i) boosting low-frequency concepts, (ii) boosting high-frequency concepts, and (iii) no frequency weighting at all, and investigate whether they have an effect on the aggregation. We find no such frequency effect, and discuss reasons for this absence in lexical sociolectometry.

1. Introduction

Lexical sociolectometry is a productive research paradigm that is concerned with studying aggregated lexical distances between varieties of a language, such as varieties of Dutch (Geeraerts *et al.* 1999), Portuguese (da Silva 2010) or English (Ruette 2012). Lexical distances are measured by determining how often varieties use different labels for the same concept. These labels form together a "lexical variable". For example, *subway* and *underground* both express (among others) the concept "subterranean public transport" in respectively American and British English. Thus, these two variants of the variable {*subway, underground*} generate lexical distance between American and British English. The idea of lexical distance can be understood as a quantitative measure of the degree to which language varieties have different "preferences" for lexical realization of a certain concept.

Lexical sociolectometry is firmly rooted in the Cognitive Linguistic tradition (see Geeraerts and Cuyckens 2007) as a usage-based model of language, because it endeavors to align variational linguistics and Cognitive Linguistics (cf. Cognitive Sociolinguistics, Kristiansen *et al*. 2008 and Geeraerts *et al.* 2010). Moreover, lexical sociolectometry is an aggregation technique, in that its primary interest is not in the behavior of individual lexical shibboleths, but in generalizing over large numbers of lexical phenomena. Such an approach is important, because it finally allows linguists to make claims about language varieties as a whole (cf. type B studies, Biber *et al.* 2009), in contrast to claims that can only focus on the variability of a single linguistic variable with regard to a certain extra-linguistic predictor (cf. type A studies, Biber *et al.* 2009). As an example, consider Labov's (1966) famous study on the pronunciation of postvocalic /r/, showing that different social classes produce different pronunciations. His study, however, does not allow the conclusion that different social classes employ different language varieties, because language varieties are characterized by a larger set of variables (cf. Hudson 1996: 22). Only when several linguistic variables are considered simultaneously – as is done in sociolectometry – is it possible to make claims about varieties in their own right.

Despite its Cognitive Linguistic grounding, research in the lexical sociolectometry tradition has so far remained fairly agnostic when it comes to frequency effects and the question of whether it makes a difference to focus on very frequent lexical variables rather than infrequent ones. As we shall see, fundamental work in lexical sociolectometry (Geeraerts *et al.* 1999, Speelman *et al.* 2003) has tended to give more weight to frequent lexical variables than to infrequent ones, assuming that usage frequency is (linearly) proportional to cognitive weight. Although this sounds straightforward enough, it is not entirely clear that this sort of weighting is optimal upon closer inspection. For example, we know from dialectological research that particularly infrequent phenomena can sometimes be particularly salient. Take, for example, double modal verbs (as in *Tom might could do it*) in dialects of English. Double modal constructions are highly salient to both dialect speakers and linguists, but are so infrequent in actual dialect corpora that corpus analyses of double modal verbs are not feasible (see Anderwald and Szmrecsanyi 2009). In other words, usage frequency does not always predict salience, which in fact could even correlate negatively with frequency (cf. Racz, this volume).

Against this backdrop, the present study sets out to answer the question of whether different types of frequency weighting actually make a difference in lexical sociolectometry. If they do, we would then have to determine which weighting procedure is cognitively most accurate (for example via validation through perception experiments). As we shall see, however, frequency weighting does not yield substantial descriptive differences regarding the aggregated lexical distances computed on the basis of corpus data. This is interesting from a theoretical point of view, given the discussion of how salience and frequency are related.

Our empirical case study investigates lexical distances in a multi-lectal dataset, the Brown family of corpora, which samples two varieties of English (British versus American), two written registers (informative versus imaginative), and two time periods (1960s versus 1990s). Using bottom-up methodologies (specifically, Semantic Vector Space models with subsequent manual screening) to avoid thematic bias or overrepresentation of specific variational dimensions, we calculate lexical distances by aggregating usage frequencies of 337 lexical variables. Some of the variables are considered fairly infrequent in the data: for example, the variable {*organization (of), planning (of)*}, as in (1), occurs less than 15 times per million words in the corpora.

(1) a. The answers derived by these means may determine not only the temporal *organization* of the dance […]. (Brown G)

b. *Planning* of vocational education programs and courses is oriented to local employer needs for trained workers. (Brown J)

Other variables feeding into our calculation of lexical distance are of average frequency in the corpus material, such as the variable {*film, movie*}, as in (2).

(2) a. This is a very individual *film*, mannered, subtle, literary [...]. (LOB C)

b. Maybe he was imagining a life with the woman in the *movie*, who was so different from his thin, nervous, and beautiful wife. (Frown R)

Finally, there are high-frequency variables in our dataset, such as {*issue*, *problem*} in (3), which has a frequency of 1,500 per million words or more in the Brown family.

(3) a. […] that can happen only over an *issue* of fundamental importance. (LOB G)

b. And just as the New Testament is time-conditioned, so is tradition, and so is our modern response to the *problem* of the ordination of women. (F-LOB D)

Given this frequency variance in the variables and our research question on how frequency influences the descriptive yield of a sociolectometric analysis, our task is now to experiment with three different sociolectometric implementations in order to model the multi-lectal nature of the dataset: one that prioritizes low-frequency concepts, one that prioritizes high-frequency concepts, and one that is agnostic about variable frequencies. We then compare the outcomes of the three implementations to test whether frequency weighting is indeed a factor in determining lexical distance between varieties.

There are five steps that are necessary in a sociolectometric analysis:

1. A large set of lexical variables needs to be compiled as the foundation on which the lexical distances will be measured.
2. The frequencies of the variables and their variants need to be observed in a stratified corpus, as the strata can represent language varieties. This is where the usage-based and quantitative characteristic of sociolectometry comes into play.
3. For every lexical variable, the lexical distance between all pairs of language varieties, i.e. the strata from the corpus, need to be calculated using the sociolectometric methodology described below.
4. The lexical distances for the individual variables are to be aggregated, and during aggregation a specific weighting scheme can be applied.
5. The matrix containing the aggregated and weighted lexical distances between all possible pairs of language varieties is transformed into a low-dimensional visualization so that high-level patterns in the structure of the language varieties can be discovered in an intuitive way.

This paper performs experiments at the fourth stage of this process, i.e. aggregating the lexical distances for the individual variables and applying a specific weighting scheme. The contribution is structured as follows: In Section 2 we present our dataset, Section 3 explains our method in detail, in Section 4 we present the actual comparisons in terms of goodness-of-fit (Section 4.1), Multidimensional Scaling visualizations (Section 4.2), and correlation statistics (Section 4.3). Section 5 offers some concluding remarks.

2. Data

As outlined above, our case study draws on corpus data from the well-known Brown family (see Hinrichs *et al.* 2010), which is composed of four parallel corpora covering three lectal dimensions. In Figure 1, we see how the national dimension (American versus British English) and historical dimension (1960s versus 1990s) are represented by the four corpora. Each of these corpora also contains a register dimension (imaginative versus informative), which yields the third lectal dimension in our study.

	1960s		1990s
American English	Brown (1961)	imaginative	Frown (1992)
		informative	
British English	LOB (1961)	imaginative	F-LOB (1991)
		informative	

Figure 1: An overview of the Brown family.

The original Brown corpus was compiled in the 1960s and samples American English imaginative texts (e.g. detective fiction, science fiction) and informative texts (e.g. press texts, learned writing). Its British counterpart is the Lancaster-Oslo/Bergen (LOB) corpus. The matching follow-up corpora, the Freiburg-Brown (Frown) and Freiburg-LOB (F-LOB) corpora, sample data from the 1990s of American and British English respectively. Each of the four corpora comprises 500 texts of about 2,000 words, which amounts to one million words of running text per corpus, or four million words in total (cf. Francis and Kucera 1964; Johansson *et al.* 1978; Hundt *et al.* 1999a, 1999b). We re-annotated these four Brown corpora using the Stanford CoreNLP part-of-speech tagger and lemmatizer (Toutanova *et al.* 2003) to make the part-of-speech tagsets across the individual components uniform.

The dataset under investigation consists of *n* = 337 variables, i.e. 337 groups of near-synonymous nouns that were semi-automatically generated (cf. below) on the basis of a large reference corpus combining the *British National Corpus* (BNC Consortium 2007) and the *American National Corpus* (Reppen *et al.* 2005), as well as the *Blog Authorship Attribution Corpus* (Koppel *et al.* 2003).

The lexical variables were automatically generated in a bottom-up fashion using Semantic Vector Space models (Turney and Pantel 2010) in combination with subsequent manual screening, yielding a supervised semi-automatic methodology. Semantic Vector Space models explore (a function of) the meaning of words by examining the syntagmatic context in which the words appear. As an example, the Semantic Vector Space meaning of the noun *lorry* may be captured through the observation that *lorry* frequently co-occurs with words such as *road*, *driver*, *traffic* or *transport*. In order to construct lexical variables we need to find nouns that are semantically very similar and Semantic Vector Space modelling can help with this: The nouns *lorry* and *truck*, for instance, are very similar because they co-occur with the same context words. By contrast, *lorry* and *sugar* are not similar as they do not tend to occur with the same context words. By keeping track of the syntagmatic qualities of words it is possible to gauge their paradigmatic interchangeability, which is an important characteristic of the variants in a lexical variable. However, we cannot blindly rely on the judgments of a Semantic Vector Space model, as these models have difficulties with polysemy and homonymy (Peirsman *et al.* 2008; Heylen *et al.* 2012). Moreover, a strong similarity between words in a Semantic Vector Space model does not necessarily entail a synonymy relation, as quite often it merely points to a high degree of association between the words. Therefore, we manually verified the similarity judgments of the Semantic Vector Space model and made corrections where necessary (hence the pre-modification "semi-automatically generated"). With some confidence, we can say that the groups of very similar words retrieved indeed constitute lexical variables of single concepts (for a detailed description see Ruette 2012).

In the following, we investigate the influence of the frequency of these variables, i.e. the sum of occurrences of all lexical variants expressing the same concept, by means of three different aggregation methodologies. The histogram in Figure 2 shows the range of frequencies of the lexical variables: as expected, there is a large number of low-frequency variables such as *organization (of)* versus *planning (of)* (indicated by the left-hand bar in the plot), while high-frequency variables such as *problem* versus *issue* are rare. This is in accordance with Zipf's law (Zipf 1949), and confirms that our collection of lexical variables is (at least to a certain extent) a naturalistic sample.

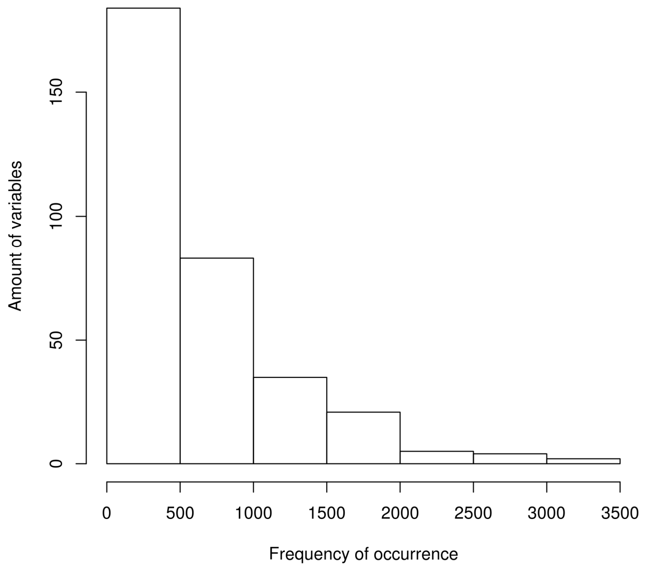

Figure 2: Histogram of the frequencies of the lexical variables, indicating how many variables (*y*-axis) occur how frequently (absolute frequencies, *x*-axis).

3. Method

In this paper we measure the lexical distances between the eight subcorpora that the three lectal dimensions in the Brown family yield:

1. 1960s American English imaginative prose
2. 1960s American English informative prose
3. 1960s British English imaginative prose
4. 1960s British English informative prose
5. 1990s American English imaginative prose
6. 1990s American English informative prose
7. 1990s British English imaginative prose
8. 1990s British English informative prose

Taking into account the frequency variation of the 337 lexical variables under consideration ({*truck* vs. *lorry*}, {*film* vs. *movie*}, {*issue* vs. *problem*}, and so on), we employ three variations of the state-of-the-art aggregation methodology in lexical sociolectometry. Each variation expresses a different approach to the implementation of the frequency effect on the aggregated similarity assessment. This represents the fourth step of a lexical sociolectometric analysis, as outlined at the end of Section 1. In the subsequent sections, we will explore this feature and compare different hypotheses about frequency effects by implementing the three different weighting mechanisms outlined in Section 1. We consider this weighting mechanism to be an implementation of the cognitive frequency effect that we want to investigate.

Our point of departure is the "profile-based linguistic uniformity" method, which was first introduced by Geeraerts *et al.* (1999) to measure the lexical distance between language varieties. It calculates the (dis)similarities between varieties based on so called "formal onomasiological profiles" (Geeraerts 2009), which are the alternating variants in a lexical variable (such as our *subway* and *underground* example) used to express a concept ("subterranean public transport") together with their frequencies. These profiles can be compared by establishing their dissimilarity through the combination of two measurements: the City-Block distance D_{CB} and the Log Likelihood Ratio based dissimilarity metric D_{LLR}. While both metrics establish a kind of distance between two given varieties, the City-Block distance uses the relative frequencies of the profiles and measures the "amount" of difference between them. The Log Likelihood Ratio based measure, on the other hand, relies on the absolute frequencies of the profiles and returns a value for the statistical confidence of there being a difference between the profiles (Speelman *et al.* 2003: 6–8). By combining both measurements, we ensure that the observed frequencies capture the underlying and statistically significant difference between the profiles being compared. Every variable then concretely defines an *n*-dimensional space, with *n* equal to the amount of variants in the variable. Subcorpora are compared in a pairwise fashion, and each pair of subcorpora under comparison is positioned in this space according to the frequency of the variants. We define the relative frequency as the frequency of the variant relative to the frequency of the variable, which is the sum of all variant frequencies. Thus, the distance on the basis of one lexical variable between two subcorpora is the City-Block distance between the positions that these two subcorpora have in the space of the lexical variable. If the D_{LLR} then returns a p-value smaller than 0.05, the two subcorpora are statistically significantly different for the lexical variable under scrutiny, and we then consider the D_{CB} to be a valid measurement. Conversely, if the D_{LLR} returns a p-value greater than 0.05, then the two subcorpora are not statistically significantly different from each other for that particular lexical variable, and we override the D_{CB} between the subcorpora and set it to zero.

In order to obtain the overall lexical distance between two subcorpora, we average the measured distances of all the variables. The calculation of this average can be weighted. The different hypotheses about frequency effects are implemented as three different weighting mechanisms: (i) high-frequency variables boosted, (ii) low-frequency variables boosted, and (iii) no variables boosted.

In a first step, we will apply the profile-based approach with a weighting that boosts high-frequency variables. In Figure 3 we observe the linear relation between the frequency of the variable and the assigned "conceptual" weight in the calculation of the average: clearly, high-frequency variables receive a much higher conceptual weight than low-frequency variables.

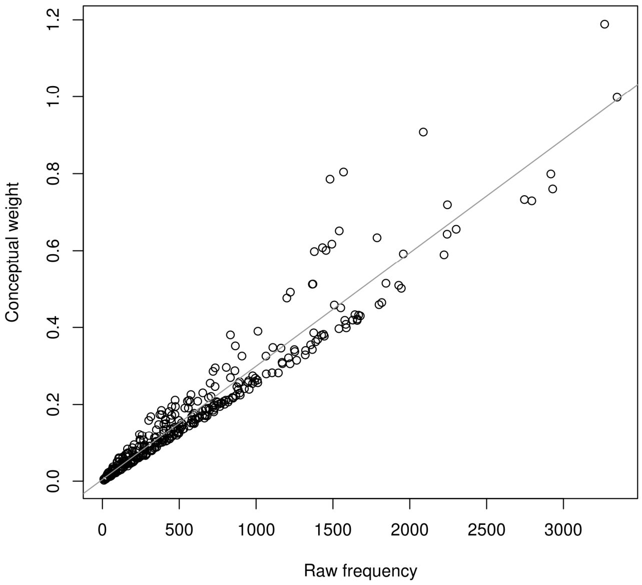

Figure 3: Prioritizing high-frequency variables. Linear relationship between variable frequency and the assigned conceptual weight in the calculation of the average: the higher variable frequency the higher the influence on the average.

Yet, we want to explore all directions of the frequency effect, i.e. the impact of both high-frequency and low-frequency variables on the lexical distance between the Brown subcorpora. Therefore, in a second step, we incorporate an opposite effect of frequency in the profile-based approach. Inspired by Goebl's *Gewichteter Identitätswert* ("weighted identity value", Goebl 1984: 83) – a similarity measurement counting infrequent words more heavily – we use a weighting mechanism which boosts low-frequency variables. The effect of this weighting mechanism can be observed in Figure 4, where there is again a linear relation between the frequency of the variables and the assigned conceptual weight in the calculation of the average. Notice that we only implemented linear relations between variable frequency and the conceptual weights. We will discuss in the conclusion that a non-linear relation might be more appropriate.

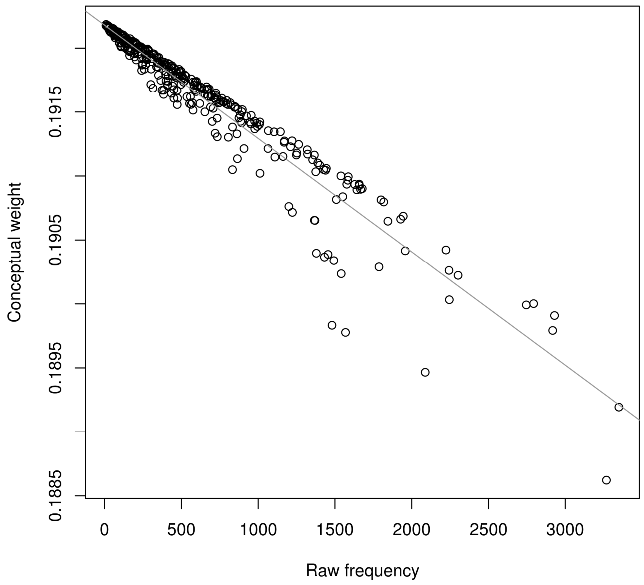

Figure 4: Prioritizing low-frequency variables. Linear relationship between the variable frequency and the assigned conceptual weight in the calculation of the average: the lower the variable frequency the higher the influence on the average.

Lastly, as a baseline to which we can compare the two weighted implementations of the frequency effect, an aggregation with no weighting of the variables at all is performed in the third step, as well.

4. Results

The three variations of the profile-based aggregation method applied on the dataset yield three distance matrices. Each distance matrix holds the aggregated lexical distances between all possible pairs of language varieties. The task is to determine how well these distance matrices capture the variational structure in the dataset. Previous research (Ruette 2012:167–170) on the same dataset indicates that the three variational dimensions in the Brown family – that is, the national distinction between British and American texts, the register distinction between informative and imaginative texts, and the diachronic distinction between texts from the 1960s and the 1990s – translates well into the multi-lectal variability as calculated by the lexical sociolectometry method. Therefore, the three distance matrices should contain these three variational dimensions. However, due to the different implementations of the frequency effect, the structure of these variational dimensions could differ. It is precisely the nature of these differences – or, in fact, the lack thereof – that takes center stage in the following discussion.

4.1. Gauging goodness-of-fit: scree plots

A first possibility for gauging the structure of the variational dimensions across the three aggregation methodologies is to create scree plots. A scree plot generally tells us the fraction of total variance in the data as explained by each individual dimension using a dimension reduction technique. The goal is to explain as much cumulative variation in as few dimensions as possible. If the input is a (non-parametric) distance matrix, the advised dimension reduction technique is (non-metric) Multidimensional Scaling (MDS)[1], which is used to reduce the dimensionality of a distance matrix to two or three dimensions for the purpose of visualization. While a common scree plot shows the explained variation of each added dimension, scree plots for a (non-metric) Multidimensional Scaling approach show the decreasing total remaining error, called "stress", with every added dimension. This is expressed as a percentage for a one, two, three, etc. dimensional solution. Usually, a solution with two or three dimensions is chosen to facilitate visualization. The stress value should always be below 15%, as a low stress value indicates a small amount of error on the estimated low-dimensional configuration.

We calculated the scree plots for the three aggregation methods and present them in Figures 5, 6 and 7. Figure 5 gives the scree plot for the profile-based approach (with high-frequency variables considered more influential in the similarity assessment between language varieties). As outlined above, a one-dimensional solution is unacceptable. However, a two-dimensional solution already has a very acceptable stress below 5%. A three-dimensional solution decreases this by a further 2.5 percent points. The considerable improvement by adding a third dimension, in comparison to the very small improvement in adding a fourth, indicates that a three-dimensional solution fits the data very well.

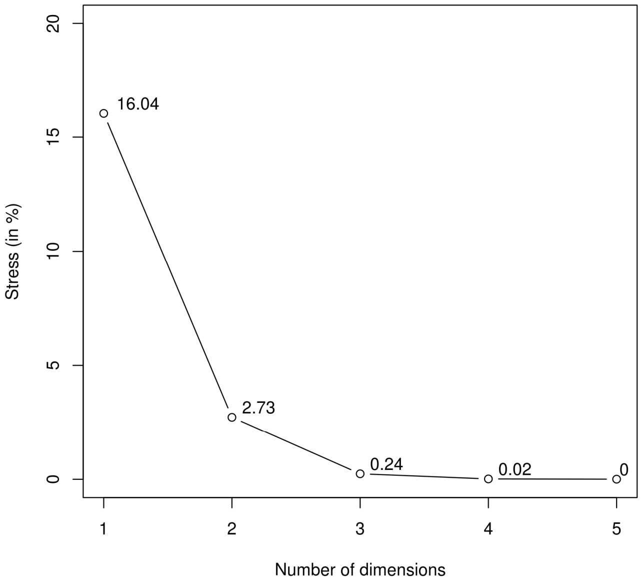

Figure 5: Scree plot of the MDS solutions for the profile-based distance (high-frequency variables boosted) measurements.

In Figure 6, we present the scree plot for the non-metric Multidimensional Scaling solutions of the profile-based distance matrix in which low-frequency variables are given more influence in the similarity assessment. As before, a one-dimensional solution is unacceptable. A two dimensional solution again has a very good stress level below 5%, but adding a third dimension does not cause a substantial further drop. If anything, an added fourth dimension would actually improve the dimension reduction solution.

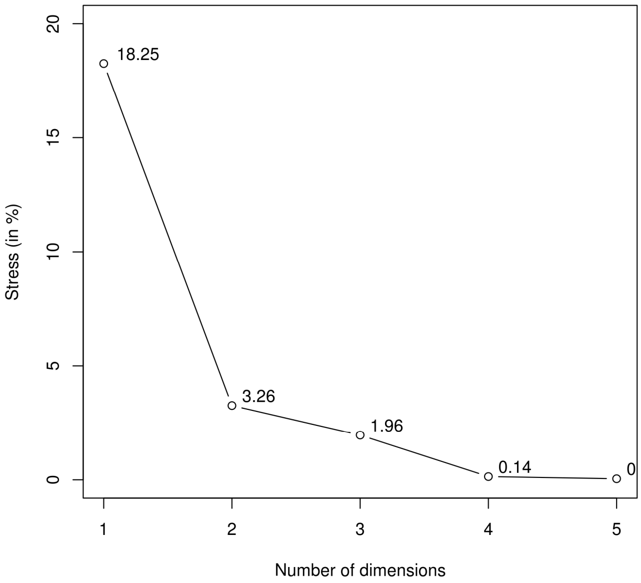

Figure 6: Scree plot of the MDS solutions for the profile-based distance (low-frequency variables boosted) measurements.

Finally, we consider Figure 7, in which the scree plot for the distance matrix is based on a completely unweighted approach. The interpretation of this scree plot is exactly the same as the interpretation of the scree plot in Figure 6.[2]

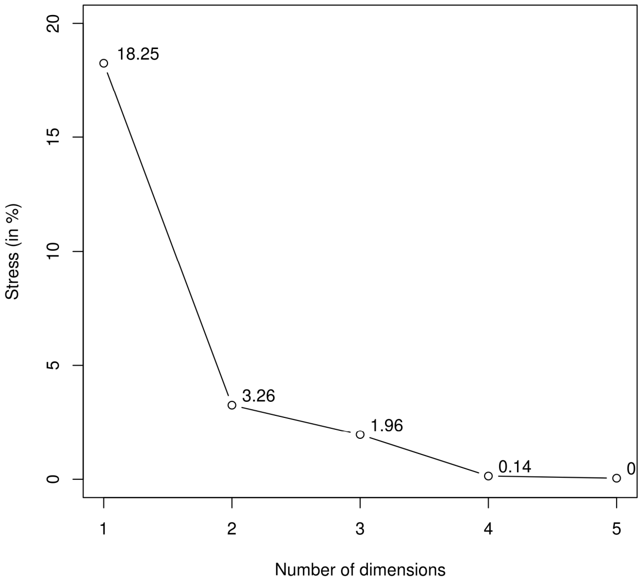

Figure 7: Scree plot of the MDS solutions for the profile-based distance (no frequency effect) measurements.

From the scree plots presented above, we can observe that there are at least two, but probably three latent dimensions present in the distance matrices. All of the approaches pick up two dimensions strongly, and the additional third dimension does not substantially improve the dimension reduction in the unweighted approach and the profile-based approach with boosted low-frequency variables. However, as follows from the scree plot in Figure 5, which shows the (original) profile-based approach prioritizing high-frequency variables, a three-dimensional MDS solution captures an additional latent structure in the distance matrix that decreases the stress value considerably. In fact, the regular profile-based approach reaches a stress level of almost zero when three dimensions are extracted. In contrast, the other profile-based approaches would need at least one additional (i.e. a fourth) dimension to reach such a low stress value.

4.2. Three-dimensional Multidimensional Scaling solutions

The scree plots presented and discussed in Section 4.1 suggest that three-dimensional MDS solutions represent the latent structure that is present in the dataset well. This comes as no surprise, since there are also by design three variational dimensions in the Brown corpora; we hope that the three MDS dimensions overlap with these dimensions. Given this, and the fact that the goal of this paper is to investigate the influence of different interpretations of the effect of variable frequency on aggregate lexical variation, we therefore calculated a separate three-dimensional MDS solution for each aggregation approach (see Figures 8, 9 and 10). These graphs locate the subcorpora studied in sets of two-dimensional planes; the *x*-axis plots scores on the first MDS dimension, while the *y*-axes plot scores on the second and third MDS dimensions. In the diagrams, proximity between subcorpora indicates aggregate lexical similarity, while distance indicates aggregate lexical dissimilarity.

In Figure 8, we find the three-dimensional MDS solution of the profile-based distance matrix with boosted influence of high-frequency variables. Interpretation of the three dimensions is very straightforward: along dimension 1 (horizontally) we find distinction between informative and imaginative subcorpora, as the imaginative subcorpora are clustered in the left half of the diagram, and informative subcorpora in the right half. Along dimension 2 (vertically, upper figure) we find a clear distinction between the British (top) and American (bottom) subcorpora. Dimension 3 (vertically, bottom figure) separates the 1960s subcorpora (top) from the 1990s subcorpora (bottom).

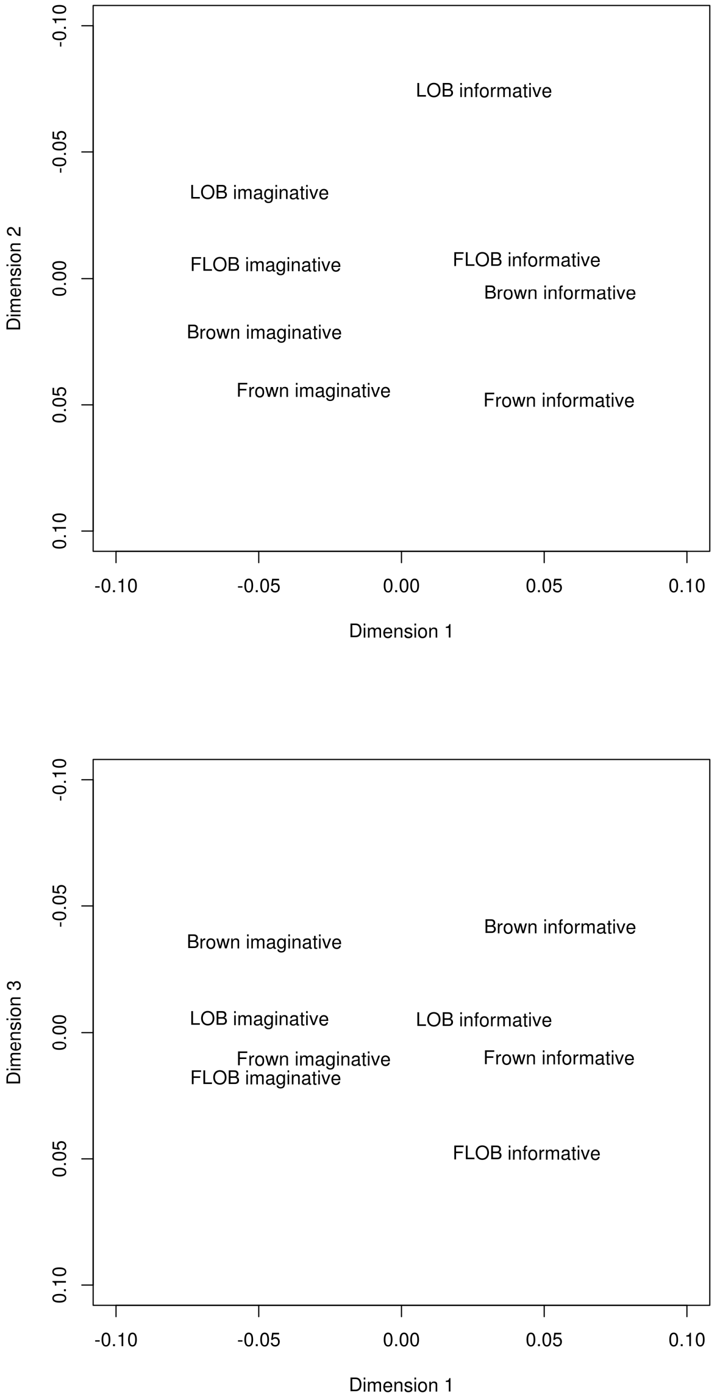

Figure 8: Three-dimensional MDS solution for the profile-based (high-frequency variables boosted) distance measurements.

The distances calculated by means of the profile-based distance metric with extra emphasis on the low-frequency variables are depicted in Figure 9. When it comes to dimensions 1, 2 and 3, it appears that the reversed weighting scheme has not influenced the interpretation of the dimensions at all: Dimension 1 still separates the informative from the imaginative subcorpora, dimension 2 nicely puts the British subcorpora at the top and the

American subcorpora at the bottom, and dimension 3 again distinguishes the 1960s subcorpora from the 1990s subcorpora. If anything, the grouping of the subcorpora is tighter and thus more delineated than in the (standard) profile-based measurements.

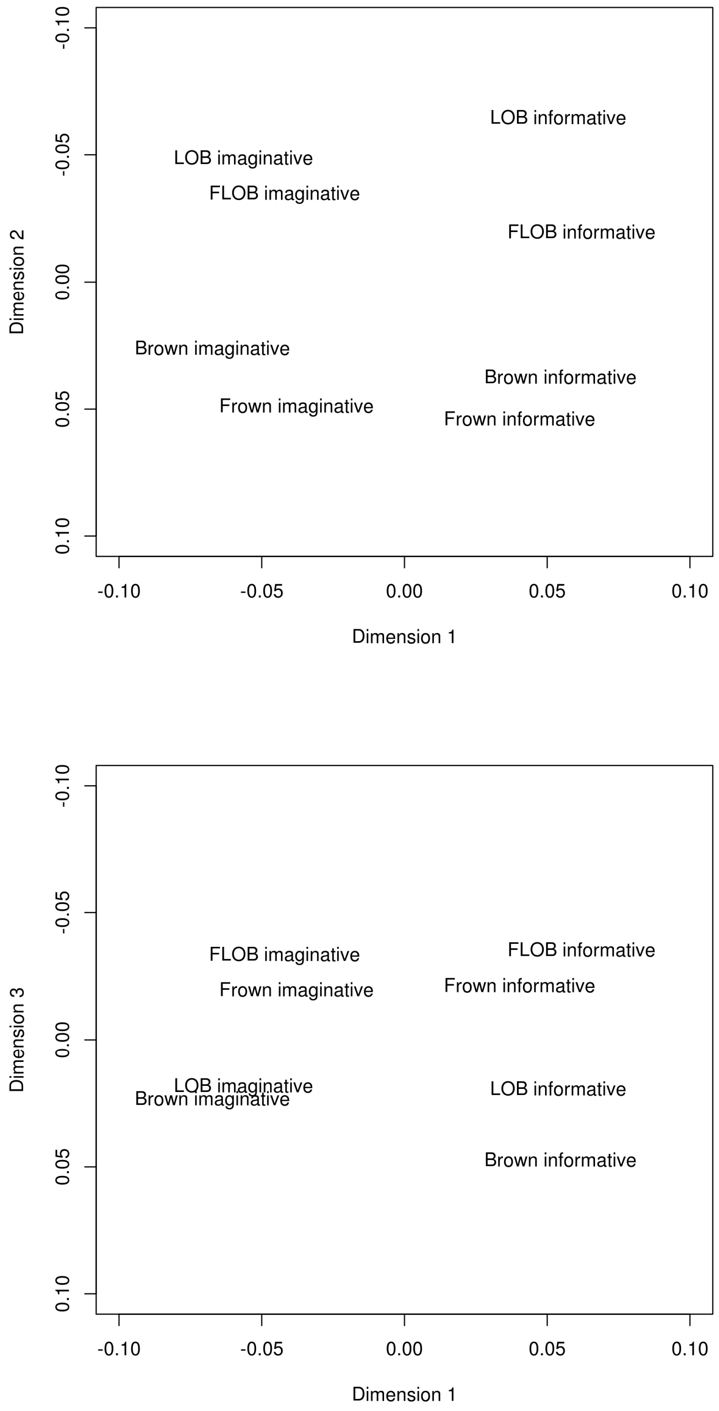

Figure 9: Three-dimensional MDS solution for the profile-based (low-frequency variables boosted) distance measurements.

Finally, Figure 10 depicts the MDS outcome when the distances between the subcorpora are measured by means of a completely unweighted profile-based distance metric.

Just as in the previous profile-based approaches, the three dimensions of the MDS solution are unambiguously interpretable, with dimension 1 linking up with the register difference between the subcorpora, dimension 2 relating to the national difference between the subcorpora, and dimension 3 indexing the temporal difference between the subcorpora. In fact, the level of similarity between the unweighted measurements (Figure 10) and the measurements in which a frequency effect emphasizing high or low frequency variables is built in (Figures 8 and 9) is remarkable. The only difference between these solutions is a mirroring of the third dimension in the low-frequency weighted condition (Figure 9), and this is an artefact of the MDS algorithm rather than a meaningful difference.

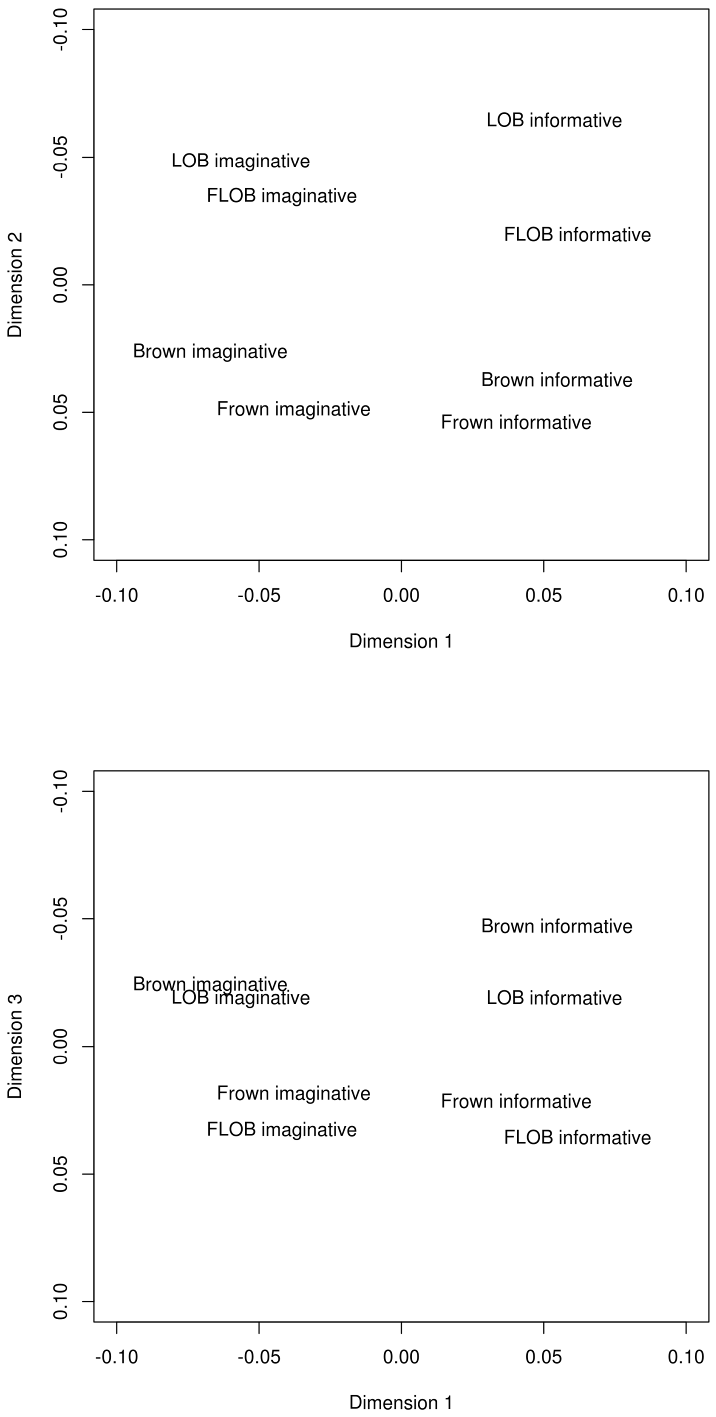

Figure 10: Three-dimensional MDS solution for the profile-based (no frequency effect) distance measurements.

All three weighting methods thus find the same theoretically expected variational dimensions. Overall there are only minor differences in the MDS output, with perhaps a tighter grouping of the subcorpora in the unweighted approach as well as in the approach that emphasizes low-frequency concepts.

4.3. Correlations

The results presented in Section 4.2 indicate that weighting does not greatly alter MDS outcomes. To rule out that MDS itself is somehow responsible for this null finding, we “cut out the middleman” (MDS), so to speak, and quantitatively assess the similarity of the three underlying distance matrices (high-frequency-concept boosted, low-frequency-concept boosted and no weighting). To do this, we perform the Mantel test of correlation (Mantel 1967). Each pair of full-blown distance matrices is considered in Figures 11, 12 and 13.

In Figure 11, the distance matrix of the regular profile-based approach (which prioritizes high-frequency variables) is compared to the distance matrix obtained from the unweighted approach. The correlation value is high at 0.8568, and yields a significant *p*-value at the 0.01 level. The scatterplot also shows a strong linear relation between the distances, with a tighter scattering when the distances are higher. Analogously, Figure 12 compares the profile-based approach in which low-frequency concepts are emphasized with an unweighted approach, and again the results are practically identical; the correlation measure is just short of 1, and the significant *p*-value is at the 0.01 level. These scatterplots also reveal that both distance measurements yield almost exactly the same values, which is why the scree plots in Figures 6 and 7 are also near-identical. Finally, Figure 13 compares the two weighted profile-based aggregations, one with an emphasis on high-frequency concepts, and one with an emphasis on low-frequency concepts. Given the two comparisons just described, it is not surprising that these matrices are also highly correlated, with a significant correlation measure of 0.8559.

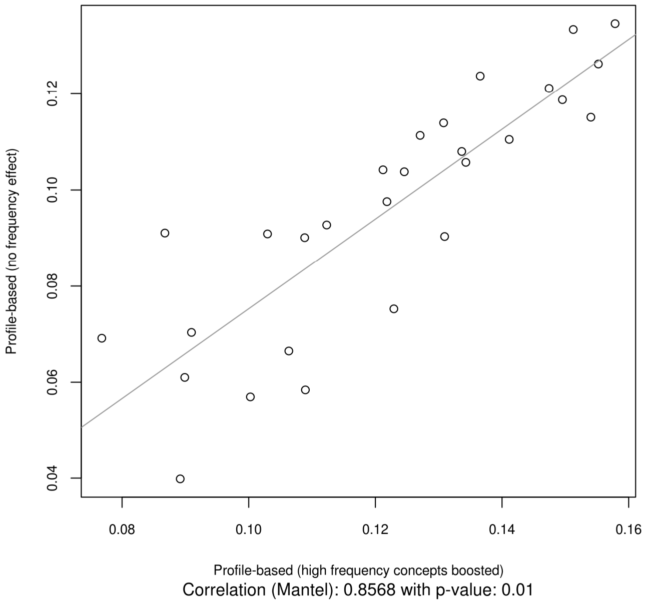

Figure 11: Mantel Correlation between profile-based aggregation with no weighting scheme and the aggregation prioritizing for high-frequency variables.

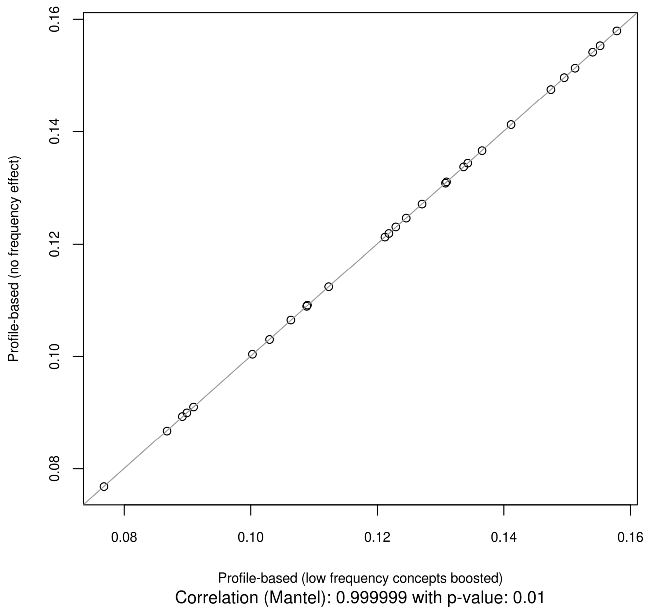

Figure 12: Mantel Correlation between profile-based aggregation with no weighting scheme and the aggregation prioritizing low-frequency variables.

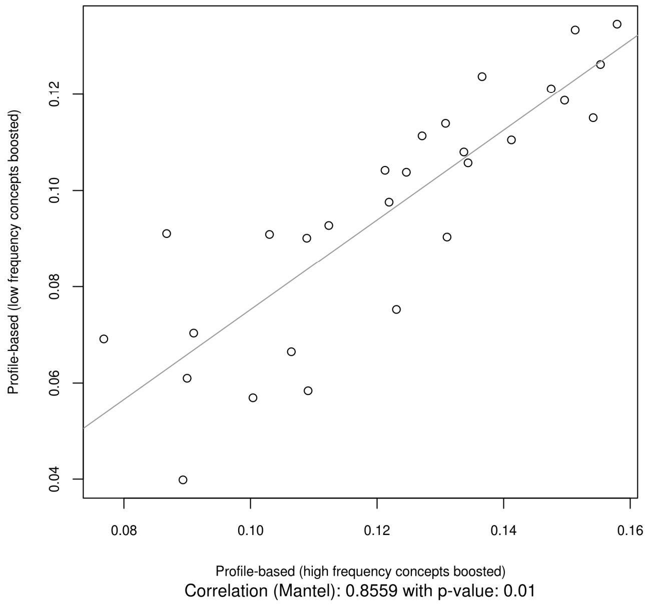

Figure 13: Mantel Correlation between profile-based aggregation prioritizing high-frequency variables and the aggregation prioritizing low-frequency variables.

5. Conclusions

Drawing on a dataset representing multi-lectal variation in written Standard English, we have experimented with different frequency weighting schemes in lexical sociolectometry.

Specifically, we investigated the extent to which different frequency-based weighting schemes alter the descriptive yield in lexical sociolectometry.

We found that different ways to weigh the aggregation of variables does not markedly change the way in which lexical sociolectometry describes the lectal structure in the dataset. While the original profile-based implementation (Speelman *et al.* 2003), which boosts high-frequency variables, works best in terms of goodness-of-fit to the data (see Section 4.1), the qualitative configuration of aggregate lexical distances is very similar regardless of which weighting scheme is used (see Section 4.2); this is because the underlying distance matrices are very highly correlated (see Section 4.3). The fact that lectal structure appears to be rather unaffected by different frequency weighting methods is puzzling, and has two possible explanations; one relating to methodology, the other to substance.

As for methodology, it is quite a recurrent theme in the literature that aggregated linguistic structure is surprisingly robust in the face of different measurement methods. Take, for example, Shackleton's (2007) dialectometric study of phonetic variability in traditional English dialects: Shackleton uses two different measurement methods, one feature-based and one variant-based, to calculate aggregate dialectal relationships. While there are interesting differences in the details, the two methods nevertheless yield a fairly similar macro structure. Therefore, it seems that aggregation research is typically able to uncover the “big picture”, no matter which specific measurement method is used. This is precisely why we do aggregation studies: to not be “overwhelmed by noise” (Nerbonne 2009: 193).

On the other hand, we should emphasize that the literature on frequency effects typically centers on *grammar* (including phonology) as the locus of such effects in usage-based models of language:

> A conceptualization of grammar as pure structure fails to provide us with explanations for the nature of grammar. A theory based on usage, by contrast, which takes grammar to be the cognitive organization of language experience, can refer to general cognitive abilities: the importance of repetition in the entrenchment of neuromotor patterns, the use of similarity in categorization, and the construction of generalizations across similar patterns. These processes, combined with the functions of language in context, such as establishing reference, maintaining coherence, and signaling turn-taking, explain grammar as the ritualization of oft-repeated routines […]. (Bybee 2006: 730; references omitted)

Regarding the extent that lexical sociolectometry can and should take into account processing (in the spirit of the above quote – and this issue still awaits discussion) we note that lexis is probably rather different from grammar. Few analysts would claim that entrenchment, neuromotor patterns, and similar patterns play a decisive role in the structure of the lexicon (except possibly in collocation preferences), and therefore it is ultimately doubtful that we should expect to see major frequency effects in the lexical domain at all. As an aside, we also note that concept and/or synonym frequencies exhibit an extreme sensitivity to lectal dimensions such as register (for example, *film* and *movie* will be fairly absent from text types such as legalese), and hence it may be that the lectal structure uncovered in our dataset is simply so powerful that it overrides any frequency effects.

Be that as it may, we hasten to add that the findings reported here are preliminary, and therefore future research should attempt to replicate our findings on the basis of different datasets exhibiting different lectal dimensionality (e.g. the spoken/written dichotomy) and describing different languages. We would also like to stress that the semi-automatically generated collection of lexical variables is, due to the use of Semantic Vector Space models, inherently somewhat biased towards high-frequency concepts, and that truly low-frequency variables are left out of the statistical analysis. Last but not least, it should be noted that the weighting mechanisms we experimented with are linear, and it would be interesting to see if the results hold when non-linear weighting functions are used. With a non-linear weighting function, the graphs in Figures 3 and 4 would no longer show a straight line, and the frequency effect would be more pronounced. Furthermore, the very low frequency concepts would be considerably more prioritized. Thus, it may very well be that such non-linear weighting functions would uncover frequency effects after all.

Notes

1. Given the non-parametric nature of our data, a non-metric version of Multidimensional Scaling was used. We used *sammon* in the statistical software package R.

2. It may seem suspicious that both scree plots are practically identical, but this identity is not by error, i.e. this is not a printing mistake.

References

Anderwald, Lieselotte and Benedikt Szmrecsanyi 2009 Corpus Linguistics and Dialectology. In: Anke Lüdeling and Merja Kytö (Eds.), *Corpus Linguistics. An International Handbook.* (Series: Handbücher zur Sprache und Kommunikationswissenschaft/Handbooks of Linguistics and Communication Science), 1126–1139. Berlin, New York: Mouton de Gruyter.

Biber, Douglas and Jones, James K. 2009 Quantitative Methods in Corpus Linguistics. In: Anke Lüdeling and Merja Kytö (Eds.), *Corpus Linguistics. An International Handbook. (Series: Handbücher zur Sprache und Kommunikationswissenschaft/Handbooks of Linguistics and Communication Science)*, 1287–1304. Berlin, New York: Mouton de Gruyter.

BNC Consortium 2007 *The British National Corpus (version 3, BNC xml edition).* Distributed by Oxford University in Computing Services on behalf of the BNC Consortium.

Bybee, Joan 2006 From Usage to Grammar. The Mind’s Response to Repetition. *Language* 82 (4): 711–733.

da Silva, Augusto 2010 Measuring and Parameterizing Lexical Convergence and Divergence between European and Brazilian Portuguese. In: Dirk Geeraerts, Gitte Kristiansen and Yves Peirsman (Eds.), *Advances in Cognitive Sociolinguistics*, 41–84. Berlin/New York: De Gruyter Mouton.

Francis, W. Nelson and Henry Kucera 1964 *Manual of Information to accompany a Standard Corpus of Present-Day Edited American English, for use with Digital Computers.* Providence, Rhode Island: Department of Linguistics, Brown University. Revised 1971. Revised and amplified 1979.

Geeraerts, Dirk 2009 Lexical Variation in Space. In: Jürgen E. Schmidt, and Peter Auer (Eds.), *Language and Space I: Theories and Methods, HSK Handbook*, 821–837. Berlin: Mouton de Gruyter.

Geeraerts, Dirk, and Hubert Cuyckens (Eds.)2007 *The Oxford Handbook of Cognitive*

Linguistics. Oxford, New York: Oxford University Press.

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

Geeraerts, Dirk, Gitte Kristiansen and Yves Peirsman 2010 *Advances in Cognitive Linguistics*. Berlin: de Gruyter.

Goebl, Hans 1984 *Dialektometrische Studien: Anhand italoromanischer, rätoromanischer und galloromanischer Sprachmaterialien aus AIS und ALF*. Tübingen: Max Niemeyer.

Heylen, Kris, Dirk Speelman and Dirk Geeraerts 2012 Looking at word meaning. An interactive visualization of Semantic Vector Spaces for Dutch synsets. *Proceedings of the 15th Conference of the European Chapter of the ACL (EACL 2012).*

Hinrichs, Lars, Nicholas Smith and Birgit Waibel 2010 A manual of information for the part-of-speech-tagged 'Brown' corpora. *ICAME Journal* 34: 189–230.

Hudson, Richard. 1996 *Sociolinguistics.* Cambridge Textbooks in Linguistics.

Hundt, Marianne, Andrea Sand and Rainer Siemund 1999a *Manual of information to accompany the Freiburg-LOB corpus of British English (“FLOB”).* Freiburg: Department of English. Albert-Ludwigs-Universität Freiburg.

Hundt, Marianne, Andrea Sand and Paul Skandera 1999b *Manual of Information to accompany the Freiburg-Brown Corpus of American English (“Frown”).* Freiburg: Department of English. Albert-Ludwigs-Universität Freiburg.

Johansson, Stig, Geoffrey Leech and Helen Goodluck 1978 *Manual of Information to accompany the Lancaster-Oslo/Bergen Corpus of British English, for use with digital computers*. Oslo: Department of English. University of Oslo.

Kristiansen, Gitte and René Dirven 2008 *Cognitive Sociolinguistics: language variation, cultural models, social systems.* Berlin: de Gruyter.

Koppel, Moshe, Shlomo Argamon and Anat R. Shimoni 2003 Automatically categorizing written texts by author gender. *Literary and Linguistic Computing* 17 (4): 401–412.

Labov, William. 1966 *The social stratification of English in New York City*. Center for Applied Linguistics: 63-89

Mantel, Nathan 1967 *The detection of disease clustering and a generalized regression approach.* Cancer Research 27: 209–220.

Nerbonne, John. 2009. Data-driven dialectology. *Language and Linguistics Compass* 3(1): 175–198.

Peirsman, Yves. 2008 Word Space Models of Semantic Similarity and Relatedness *Proceedings of the ESSLLI-2008 Student Session, Hamburg, Germany.*

R Development Core Team 2011 *R: A Language and Environment for Statistical Computing.* Vienna: R Foundation for Statistical Computing.

Reppen, Randi, Nancy Ide and Keith Suderman 2005 *American National Corpus (ANC).* Linguistic Data Consortium, Philadelphia. Second release.

Ruette, Tom 2012 *Aggregating Lexical Variation: towards large-scale lexical lectometry.* PhD thesis, University of Leuven.

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

Shackleton, Robert G. Jr. 2007 Phonetic variation in the traditional English dialects: a computational analysis. *Journal of English Linguistics* 35 (1): 30–102.

Toutanova, Kristina, Dan Klein and Christopher Manning 2003 Feature-rich part-of-speech tagging with a cyclic dependency network. In: *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics 2003*, 252–259.

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

Zipf, George K. 1949 *Human behaviour and the Principle of least effort*. Cambridge, MA: Addison-Wesley Press.