A lectometric analysis of aggregated lexical variation in written Standard English with Semantic Vector Space models

Tom Ruette[i], Katharina Ehret[ii] and Benedikt Szmrecsanyi[i]
[i]KU Leuven / [ii]University of Freiburg

Lectometry is a corpus-based methodology that explores how multiple language-external dimensions shape language usage in an aggregate perspective. The paper combines this methodology with Semantic Vector Space modeling to investigate lexical variability in written Standard English, as sampled in the original Brown family of corpora (Brown, LOB, Frown and F-LOB). Based on a joint analysis of 303 lexical variables, which are semi-automatically extracted by means of a SVS, we find that lexical variation in the Brown family is systematically related to three lectal dimensions: discourse type (informative versus imaginative), standard variety (British English versus American English), and time period (1960s versus 1990s). It turns out that most lexical variables are sensitive to at least one of these three language-external dimensions, yet not every dimension has dedicated lexical variables: in particular, distinctive lexical variables for the real time dimension fail to emerge.

Keywords: lectometry, lexis, aggregation, Semantic Vector Space models, Standard English

1. Introduction

This paper presents a comprehensive analysis of lexical variation in written Standard English. Drawing on state-of-the-art lectometric methods (Geeraerts et al. 1999, Speelman et al. 2003), we explore the extent to which lexical choices in the Brown family of Standard English corpora (Hinrichs et al. 2010) are systematically structured by three lectal dimensions, i.e. standard variety, discourse function, and real time period. Our goal is to offer a data-driven vision for the study of lexical variation by introducing Semantic Vector Space models as a means for

International Journal of Corpus Linguistics 21:1 (2016), **48–79**. DOI 10.1075/ijcl.21.1.03rue
ISSN 1384–6655 / E-ISSN 1569–9811

identifying lexical variables, and by utilizing Individual Differences Scaling as a tool for transparently aggregating lexical variation.

Our approach in this paper consists of the following five steps: (i) draw on large corpora (the British National Corpus, the American National Corpus, and the Blog Authorship Attribution Corpus) and 'Semantic Vector Space' modeling (Turney & Pantel 2000) to obtain an unbiased lexical variable set; (ii) determine the frequency distribution of lexical variant forms in the Brown corpora; (iii) rely on the 'Profile-based Distance' Metric (Speelman et al. 2003) to transform the distributional information into distances among the cross-categorizations of lectal dimensions that are represented in the Brown corpora; (iv) utilize 'Individual Differences Scaling' (Takane et al. 1977) to analyze the distances; and (v) interpret.

To exemplify the link between lexical variation and lectal dimensions, consider the lexical items that refer to the concept of SUBTERRANEAN PUBLIC TRANSPORT in English: *subway* is the American English variant, *underground* the British English one. This is a clear-cut case of transatlantic variation, yet it is unclear to which extent variation along these lines pervades the lexicon. We generalize in this study a large amount of lexical choices of the type *subway* versus *underground*, and find that variation in the lexicon is both lectally structured and systematic. On more methodological grounds, we advance 'lectometry' by sketching a semi-automatic bottom-up method (Semantic Vector Space modeling) for collecting lexical variables, which moves the field beyond the study of aprioristically defined usual-suspect phenomena (e.g. *subway* versus *underground*, *truck* versus *lorry*). We also highlight the benefits of a transparent aggregation method (Individual Differences Scaling), which enables the lectometrist to marry quantitative measurements with the qualitative discussion of linguistic variation phenomena.

This paper is structured as follows: in Section 2, we review the literature and discuss the theoretical and empirical background guiding our study. Section 3 presents the dataset and explains our methodology in more concrete terms. Section 4 reports and discusses the findings. Section 5 offers concluding remarks.

2. Theoretical and empirical background

In this section, we discuss in more detail the theoretical and empirical background for lectometry. First, we investigate lexical variation from a theoretical perspective, in Section 2.1. Then, lectometry is described as an aggregative variational linguistic method in Section 2.2. Since calculating distances is central to lectometry, Section 2.3 discusses similarity measures. The two final subsections describe respectively the semi-automatic methodology for finding lexical variables

(Section 2.4) and the transparent aggregation method of Individual Differences Scaling (Section 2.5).

2.1 Lexical variation in a theoretical perspective

The current paper considers lexical 'onomasiological' variation and follows Geeraerts et al. (1994: 5): "The onomasiological perspective takes its starting point on the level of semantic values and describes how a particular semantic value […] may be variously expressed by means of different words". Our point of departure is therefore the variations among words which are used to label a certain concept, and, in our approach, lexical variables comprise sets of near-synonymous words. In contrast to an onomasiological approach, the 'semasiological' approach takes a certain word as its starting point and investigates which concept or concepts are expressed by that word. For instance, *subway* is also a semasiological variable as it can not only be used to refer to the concept SUBTERRANEAN PUBLIC TRANSPORT, but also to the concept UNDERGROUND TUNNEL FOR PEDESTRIANS. The term *subway*, however, is considered a typical American English (henceforth: AmE) label for SUBTERRANEAN PUBLIC TRANSPORT, and at the same time a typical British English (henceforth: BrE) label for UNDERGROUND TUNNEL FOR PEDESTRIANS. Thus, if we did not also make a semasiological distinction in the meaning of *subway*, our analysis of the lexical variables *subway* versus *underground* would be confounded. While the two approaches cannot be strictly separated, this paper primarily adopts the onomasiological view on lexical variation. Onomasiological studies of lexical variation may be traced back to the investigation of Zauner (1902) on Romance names for body parts, in which the names for body parts in several Romance languages were compared. The onomasiological perspective is very similar to the concept of 'sociolinguistic alternation variables'. Labov (1972: 271) argues that "social and stylistic variation presuppose the option of saying 'the same thing' in several different ways: that is, the variants are identical in reference or truth value, but opposed in their social and/or stylistic significance". Labov's (1972) construct of saying "the same thing" can be informally translated into the onomasiological definition of Geeraerts, Grondelaers, and Bakema (1994: 5), quoted above.

However, we concur with Lavandera (1978), who argues that truth-conditional interchangeability, i.e. interchangeability *salve veritate* (Quinn 1951) can be problematic for non-phonological alternation variables, including lexical ones. Note, however, that we adopt in this study the idea of paradigmatic (not truth-conditional interchangeability) as is customary in 'distributional semantics'. In distributional semantics, along the lines of the famous quote "you shall know a word by the company it keeps" (Firth 1957: 11), the meaning of a word is shaped by the contexts in which it occurs (see also Sinclair 1991). This brings us to the

corpus-linguistic notion of a collocation, which is defined as "a lexical relation between two or more words which have a tendency to co-occur within a few words of each other in running text" (Stubs 2002: 24). A common measure for the strength of a collocation is the Church & Hanks (1990) 'Pointwise Mutual Information' index. The PMI index is defined in terms of the probability of having the combination *x* and *y* compared to the probabilities of having *x* and *y* separately (Geeraerts 2010: 173). As a matter of fact, this index will resurface in the Semantic Vector Space construction, cf. Section 2.4. With respect to the topic of this paper, Sinclair's (2004: 29) seminal distinction between the open-choice principle (words having a fixed meaning) and the idiom principle ("words tend[ing] to go together and make meanings by their combinations") is central; the present study places an emphasis on idiomaticity to define word meaning. A thorough discussion of distributional semantics in a corpus linguistic context can be found in Geeraerts (2010: 165–178).

In sum, we consider interchangeability an empirical, not a formal semantic problem. Our computational implementation of distributional semantics with Semantic Vector Space models will be explained in below.

2.2 Lectometry

Lectometry is an aggregative variational linguistics method, where claims are based on the analysis of feature aggregates as opposed to single features. As such, lectometry bears similarities to dialectometry (e.g. Seguy 1971, Goebl 1984, Nerbonne & Kretzschmar 2003), multidimensional register analysis (e.g. Biber 1988), and also some methods used in quantitative typology (e.g. Cysouw 2005, Bickel 2007, Borin & Saxena 2013, Wälchli & Szmrecsanyi 2014). All these methodologies consider the big picture by aggregating over a number of features. In dialectometry (see Szmrecsanyi 2011, 2013 for a demonstration of how to conduct corpus-based dialectometry), for instance, multiple linguistic features with a geographical pattern are considered simultaneously to explore the extent to which feature aggregates reveal geo-linguistic patterns. Technically speaking, dialectometrists calculate (dis)similarity measures which can subsequently be correlated, for example, with geographical distances (Nerbonne 2009). Outside dialectometry (and in the realm of American English), good examples of aggregation research include Labov et al.'s (2006) *Atlas of North American English*, or seminal work of Grieve et al. (2011), who uses statistical methods such as spatial autocorrelation for smoothing noise in the geographical signal.

In register analyses, the goal is to identify dimensions of functional opposites, such as narrative versus non-narrative texts, and to establish which linguistic features are associated with particular dimensions using Factor Analysis (as utilized

in Biber 1988) and similar techniques. Because register analysis is primarily interested in usage patterns of linguistic features (rather than the linguistic objects such as dialects or texts in which these features are being observed, as in dialectometry), register analysts typically calculate correlation measures among the linguistic features (rather than dissimilarities among the objects, as is customary in dialectometry). This central interest in linguistic features notwithstanding, register analysis can yield a 'typology', or a functionally and linguistically motivated categorization of texts (Biber 1989). This then resembles output in lectometry.

Lectometry is in a way a super-construct, generalizing and drawing on both dialectometric and register analytic methodologies: it simultaneously aims to (i) measure dissimilarities among linguistic objects, and (ii) establish which linguistic features are implicated in the differences. What sets lectometry apart from both dialectometry and register analysis is that lectometry considers multiple language-external factors simultaneously — register, geography, real time, etc. (while dialectometry tends to be exclusively interested in geography, and register analysis in functional differences). Due to the fact that we take a lectometric approach, it is this multiplicity of language-external factors that will take center stage in the present paper.

2.3 Calculating distances

A lectometric approach aggregates over multiple linguistic variables with the goal of finding patterns in the interaction of language-external factors (on how to find these variables see Section 2.4). Such patterns emerge from measuring the pairwise dissimilarities between (sub)corpora which represent cross-sections of language-external factors. The pairwise dissimilarities, in turn, are calculated on the basis of the usage behavior of the linguistic variables. Consider language-external factors such as standard variety (BrE versus AmE) and discourse function (imaginative versus formative); these yield cross-categorizations such as "BrE informative", "AmE imaginative", and so on. Two cross-categorizations can be considered similar if their usage patterns of linguistic variables are similar. For example, BrE informative prose and BrE imaginative prose may be similar to the extent that they both use a certain lexeme (e.g. *underground*) frequently, and more frequently than AmE prose.

There are many ways to calculate linguistic distances (see Borin & Saxena 2013). In the present study, we rely on the 'profile-based distance metric' (Speelman et al. 2003). The idea of a profile-based distance metric is to take into account the semantic, conceptual, or functional relations between individual features (lexemes). In a lexical variable, two or more lexemes are united by a common meaning or concept. Standard (dia)lectometric aggregation methods (e.g.

Nerbonne & Kretzschmar 2003, Heeringa 2004) ignore such relations between features; Speelmanet al. (2003), Heylen & Ruette (2013) and Ruette et al. (2014) discuss how this neglect can substantially impact results. What is important here is that the profile-based distance metric calculates the distance between pairs of corpora for every single variable (i.e. set of semantically connected features) using a standard lectometric distance metric. Next, these distances are averaged over the variables (see Speelman et al. 2003 for details).

The fact that measuring distances is so crucial in the investigation of variability among cross-categorizations of lectal variables rules out the usage of Principal Components Analysis (PCA) and Factor Analysis (FA), which are used in authorship attribution (Grieve 2007) and register analysis respectively. Both PCA and FA rely on correlation metrics among the linguistic features, and only from there, the component/factor scores can be used to estimate similarity among cross-categorizations of lectal variables. The advantage of PCA and FA, however, is that they provide access to the behavior of the individual sources of variability in the form of variable loadings. Such access has been absent from traditional dialectometry, thanks to its reliance on aggregative distance metrics that abstract away from the behavior of individual features. To remedy this shortcoming, we propose the usage of INDSCAL, which is discussed in Section 2.5.

2.4 Finding variables: Semantic Vector Space models

This study will marshal Semantic Vector Space models to identify an unbiased set of lexical variables in a collection of large corpora (the British National Corpus, the American National Corpus, and the Blog Authorship Attribution Corpus; see Section 3.2 for details). Subsequently, the distributional behavior of these externally generated lexical variables in the Brown corpora will be analyzed and interpreted. The outcome of any aggregate analysis, including register analyses and dialectometry, is highly dependent on the linguistic variables that are fed into the analysis. In register analysis and dialectometry, we often find a bias towards functionally or geographically "interesting" features that have already been investigated in isolation in the literature. Such a bias is not advisable in lectometry proper, because it will not yield a truthful picture of the multivariate and multidimensional structure of language-external intersections. This is another way of saying that in lectometry, the set of input variables needs to represent the totality of available linguistic variation as accurately as possible. The set of linguistic variables should therefore consist of many lexical variables (concepts). Furthermore, the variants (lexemes) of each variable should give an accurate account of the lexical variation of its concept. However, such a high quality sample of lexical variables is hard to come by. A seminal study in lectometry, Geeraerts et al. (1999), draws

on introspection to identify lexical variables in two lexical fields (sports and clothing), but this procedure does not scale to the entire lexicon as it would not be feasible to intuit all possible lexical variables in all possible lexical fields. Another possibility would be to consult a synonym dictionary or a thesaurus, a method that has other problems: dictionaries and reference works are biased towards a somewhat higher register and do not cover all aspects of the lexicon, and often the exact methodology behind dictionaries is not made explicit.

Instead, we adopt a distributional semantics perspective, and utilize a semi-automatic, data-driven methodology for finding semantically similar words, thus creating a representative sample of lexical variables: Semantic Vector Space models. These can discover lexical variables in naturalistic corpus data by considering the distributional patterns of words. Consider in this respect the contribution of Peirsman et al. (2015). The idea is that we can find words that may have an onomasiological relation to each other by applying the criterion of interchangeability. In this approach, the semantic similarity of words is determined by whether or not the words under scrutiny occur in identical or at least comparable contexts in a reference corpus. Semantic Vector Space models yield a sizeable set of potential lexical variables from all corners of the lexicon as long as the data that was used as input is sufficiently large and representative.[1] We refer the reader to Turney & Pantel (2010) for an introduction to and a more detailed discussion of Semantic Vector Spaces. In a last step, we subject the output of the Semantic Vector Space Model to manual verification, as the state-of-the-art Semantic Vector Space modeling does not yet permit a *fully* unsupervised analysis. Recall in this connection that we aim to sketch the potential of the method in this paper, without denying that there is still scope for perfection.

2.5 Visualizing distances: Individual Differences Scaling

Once the lectometrist has at her disposal a set of linguistic variables (see previous section), she is in need of an aggregation method that can link the language-external factors transparently to the aggregate behavior of the individual variables.

1. An anonymous reviewer pointed out correctly that a Semantic Vector Space Model might also produce a certain bias in the sample of lexical variables by underestimating extra-linguistic variation. SVS models rely on the fact that the word forms co-occurring with the nearsynonyms tend to be the same as the concepts in the target words However, near-synonyms like *hiccups* and *singultus* will have far less similar contexts because medical texts not only discuss different aspects of the phenomenon but often use different words to refer to the same co-occurring concepts (e.g. *oesophagus* instead of *gullet*). On the aggregate level, this will result in measuring points (variables) that can (at best) detect differences between overall similar varieties, but not between very dissimilar varieties.

Note that work in dialectometry is often criticized because there is no access to the linguistic variables after the aggregation step (Schneider 1988, Woolhiser 2005). There are a number of proposals to remedy this shortcoming (e.g. Wieling et al. 2011, Wieling & Nerbonne 2011), but the issue is still a matter of current debate in the research community. Grieve et al. (2011) in particular advocate the use of Factor Analysis for transparent aggregation, but this approach does not straightforwardly accommodate the use of distance-metric generating methods such as the Levenshtein algorithm (Heeringa 2004), which are frequently used in dialectometry (Plevoets et al. 2008 explain the amount of mathematical effort that has to go into using a distance metric in a method that is based on Singular Value Decomposition, such as Factor Analysis).

Therefore, we propose the use of Individual Differences Scaling (also known as INDSCAL; see Takane et al. 1977) for transparent aggregation. INDSCAL is a type of Multidimensional Scaling (MDS), which is a customary method in dialectometry. Traditional (two-way) MDS takes a matrix of pairwise distances between 'objects' (for example, cross-categorizations of extra-linguistic factors) and attempts to represent these distances in a lower-dimensional plane. The distances between these objects are measured by so-called 'sources' (in our case linguistic variables, such as word choice for realizing a certain concept). The main problem with two-way MDS is that it ignores the differences between sources, as already pointed out by Horan (1969). A single distance has to represent the distance between the objects, and this single distance is obtained by means of averaging over the distances of the individual sources. The crucial advantage of INDSCAL, on the other hand, is that it can take multiple two-way distance matrices, each representing the (non-averaged) pairwise distances between the objects for a single source. The aggregation of the distance matrices happens within the INDSCAL calculations. Therefore, INDSCAL yields an aggregation output that accounts for the differences between the sources. To employ INDSCAL in a multivariate and multifactorial lectometric framework, it is necessary to re-interpret the sources as linguistic variables, and the objects as language-external intersections (cf. Ruette & Speelman 2013).

The outcome of an INDSCAL analysis consists of two parts. The first part is the 'Group Stimulus Space', which is comparable to the output of a typical two-way MDS analysis. The Group Stimulus Space shows the position of the objects or language-external intersections in a low-dimensional space. The low-dimensional artificial INDSCAL distances (or their ranking) are kept as close as possible to the original distances. The interpretation of the Group Stimulus Space is completely parallel to the interpretation of a traditional MDS solution, insofar as it is restricted to an interpretation of the dimensions. The second part of the INDSCAL output consists of the 'Configurations Weights' for the individual sources. These

Configuration Weights give an impression of the extent to which the individual sources agree with the dimensions, i.e. every source obtains a Configuration Weight. As an example, a Configuration Weight greater than 1 indicates that a particular source considers a particular dimension to be particularly crucial. A Configuration Weight smaller than 1 indicates that a particular source downplays the distinction made by that dimension. Put differently, a lexical variable (e.g. *lorry* versus *truck*) with a Configuration Weight smaller than 1 for a certain dimension (say, 1960s versus 1990s) exhibits a weaker-than-normal variational pattern for that dimension. By contrast, a lexical variable (think of *lorry* versus *truck* again) with a Configuration Weight greater than 1 for a certain dimension (say, BrE versus AmE) exhibits a stronger-than-normal variational pattern for that dimension. Thus, *lorry* is not particularly sensitive to the 1960s/1990s dimension, but reflects the difference between AmE and BrE.

3. Data and methodology

In this section, we flesh out the more theoretical introduction of the previous section. First, we introduce the data collection that our case study will be based on. Then, we minutely detail the parameters of the Semantic Vector Space model that we use for finding lexical variables. Finally, the steps in the lectometric analysis are made explicit.

3.1 The Brown corpora of standard written English

The original Brown family of corpora (see Hinrichs et al. 2010) consists of four matching components and contains published AmE and BrE texts from the 1960s and 1990s. Each of the four corpora counts roughly 1 million words, comprising 500 samples of about 2,000 words of running text from fifteen different genres. We rely on the original Brown corpus compilers' classification of these genres into two broad categories: informative and imaginative. It is this binary distinction which we will use in our analysis as the functional dichotomy. The nine informative genres cover press reportage (A), editorials (B), reviews (C), religion (D), skills and hobbies (E), popular lore (F), belles letters and biographies (G), as well as miscellaneous (H) and learned writing (J). The six imaginative genres sample texts from general fiction (K), mystery and detective fiction (L), science fiction (M), adventure and Western fiction (N), romance and love story (P) as well as humor (R). Table 1 provides an overview of the Brown family according to functional category, variety, and time period.

Table 1. Overview of the Brown family and their language-external parameters

Functional category	Standard variety	Time period	Corpus
Imaginative	American English	1960s	Brown (K-R)
		1990s	Frown (K-R)
	British English	1960s	LOB (K-R)
		1990s	F-LOB (K-R)
Informative	American English	1960s	Brown (A-J)
		1990s	Frown (A-J)
	British English	1960s	LOB (A-J)
		1990s	F-LOB (A-J)

3.2 Semantic Vector Space modeling: Technical details

As noted above, the set of lexical variables used in lectometric analysis should be unbiased. We specifically need to avoid two different kinds of bias: (i) bias within lexical variables and (ii) bias across lexical variables. The former occurs when not all possible realizations (i.e. 'variants') of the concept underlying the lexical variable are accounted for (this is essentially about Labov's (1969) 'Principle of Accountability'); the latter occurs when the set of lexical variables is thematically biased. We can address both levels of bias by employing Semantic Vector Space models, which return semantically highly similar, if not identical, paradigmatically and empirically interchangeable words that serve as lexical variables in our lectometric analysis.

Ruette (2012) offers a discussion of caveats and limitations of Semantic Vector Spaces and their use for finding lexical variables; suffice it to say that we limit the set of lexical variables to fairly frequent nouns since Semantic Vector Space models require a large amount of data to be effective and nouns are the most frequent word class in Standard English corpora (see the frequency matrices in Hinrichs, Smith & Waibel 2010). For this reason, the models perform best for the semantic modeling of nouns. Moreover, we note that the Brown corpora are not sufficiently large for the calculation of a trustworthy Semantic Vector Space model. Hence, we combine the British National Corpus (BNC Consortium 2007), the American National Corpus (Reppen et al. 2005), and the Blog Authorship Attribution Corpus (Schler et al. 2006) to obtain a dataset of about 250 million words of text. Based on this combined dataset, we use Semantic Vector Space modeling to generate a comprehensive set of lexical variables. Subsequently, we determine the distribution of these externally generated variables in the Brown corpora.

Our implementation of the Semantic Vector Space model relies on a bag-of-words model that considers three words to the left and the right of the target word. Additionally, the dimensionality of the Semantic Vector Space is set to 4,900

dimensions. This dataset is obtained by only considering the 5,000 most frequent contexts, minus the first 100 most frequent words, which are typically function words. Function words are generally considered to be noise in a Semantic Vector Space model, and are therefore always removed.[2]

The output of the Semantic Vector Space model is a search space consisting of the approximately 8,000 unique nouns that occur in the Brown corpora as rows, and the exactly 4,900 contexts as columns. This lexeme-by-context matrix is next transformed into a square lexeme-by-lexeme similarity matrix by means of the cosine similarity metric. The resulting similarity matrix is then used as input for a clustering algorithm which bears resemblance to the 'Clustering by Committee method' (Pantel 2003). The cluster algorithm calculates the 100 nearest neighbors for every target word. Calculating more than 100 of the nearest neighbors typically yields unrelated words and takes up more computation time. In the set of nearest neighbors of each target word w_t, all possible pairs of neighbors (excluding pairs of identical words) are formed, and for every pair, a similarity score is calculated. Specifically, we calculate the similarity score by multiplying the average similarity of the pair — this is the sum of the similarities of both neighbors w_1 and w_2 to the target word, divided by 2 — with the square of the difference in similarity of both neighbors to the target word. The first term catches the relevance of the pair to the target word; the second term, which is considered more important and is therefore squared, catches the intrinsic tightness of the pair. This similarity score is mathematically presented in the following equation:

$$\frac{sim(w_1,w_t) + sim(w_2,w_t)}{2} \times (sim(w_1,w_t) - sim(w_2,w_t))^2$$

For every target word, the pairs that have a similarity score greater than 0.4 are combined so that for every target word a cluster of words emerges. A cutoff at 0.4 makes sure that not the whole search space is clustered, i.e. it is acceptable when not all lexemes are subsumed in a cluster. The cut-off at 0.4 was set after some experimentation, and appeared to give the most intuitive results. As a result, a list of highly similar lexemes per target word is found. In our study, the cluster algorithm yielded 2,138 clusters of highly similar words. Our analysis does not necessarily fully comply with the Principle of Accountability (Labov 1969, fn. 20), as the list of variables in the Appendix may miss some potential variants. Note here, however, that only those variants that actually occur in the Brown corpora are being considered in this study. Admittedly, low-frequency lexical items may

2. Co-occurrence frequencies were transformed to Pointwise Mutual Information scores (Church & Hanks 1990), with negative scores set to zero. These settings are considered to be optimal default settings (Peirsman 2010).

be underrepresented, since Semantic Vector Space models are to a certain extent biased towards high-frequency items.

Before extracting the frequencies of the lexical variants from the Brown family, however, the clusters were subjected to manual verification for the sake of keeping recall as high as possible: human coders judged clusters of highly similar words on the type level. If a cluster contained words that are potentially interchangeable — even in a restricted context — the cluster was retained. *Milk* and *cream*, for example, are usually not considered interchangeable, yet in certain contexts such as in *Do you take milk/cream in your coffee?*, they are. Thus, such clusters were retained. Only if no context at all could be found in which all the words in the cluster were interchangeable, the cluster was removed. Of the 2,138 automatically detected clusters, only 303 clusters were retained after manual inspection. We identified the text frequencies of the lexemes contained in the 303 retained clusters from the 8 subcorpora of the Brown family. Since a further manual verification step would follow later, the manual inspection described here was performed by the lead author only, in contrast to the upcoming manual step, where the judgments of multiple annotators were compared.

As the variables were retained in all potentially interchangeable contexts, a further (semasiological) control of the lexemes on the token level was necessary. Hence a subset of lexemes which were deemed particularly problematic by the authors was manually verified for interchangeability in their context in the corpus data. Thus, the second stage of manual verification warrants precision. Only occurrences in which all the words of a given cluster could be interchanged were retained. Example (1) illustrates the manual coding of interchangeability in the *cream*/*milk* cluster. While *milk* and *cream* are interchangeable in Examples (1a) and (1b), they are not in Examples (1c) and (1d). Thus, the first occurrence was retained while the latter two were excluded.

(1) a. The cat which had just licked its saucer of **milk** clean of every final scrap curled up into a fluffy ball of ginger fur […]. (FLOB, F03-TAG)
b. **Cream** is not always offered with tea, as once it was: it was usually handed separately and added to the tea in the cup.(LOB, E26, BrE60 Inf)
c. If it's true that contented cows give more **milk** why shouldn't happy ball players produce more base hits? (Brown, A15, AmE60 Inf)
d. […] Alemagna a delightful though moderately expensive restaurant which is particularly noted for its exceptional selection of ice **creams** and patisseries. (Brown, F41, AmE60 Inf)

Given the semantic, and therefore difficult nature of the manual annotation, we evaluated inter-annotator agreement to assess the trustworthiness of the judgments in the second stage of manual verification. A subset ($n = 679$) of the

observations subject to manual verification was annotated by two annotators. Cohen's unweighted kappa for two raters is 0.882, which indicates an excellent (almost perfect) agreement.

The complete dataset consists of 160,842 lines of KWIC concordance, based on 303 lexical variables. Of the more than 160,000 observations, 32,008 (20%) were manually verified; of the 303 variables, 64 (21%) were manually verified.

Admittedly, the application of the Semantic Vector Spaces requires a great deal of manual verification and a critical reader may wonder why such heavy machinery was needed in the first place. In response to this criticism, we argue that the proposed automatic methodology is actively being developed and improved, e.g. in the domain of Token-based Semantic Vector Spaces (Heylen et al. 2012, Navigli 2012, Dinu et al. 2012), and we are confident that in the near future, the amount of manual verification that is needed for the current research will decrease substantially. Moreover, the automatic approach optimizes the recall of potential variables, which is an objective that cannot be accomplished in a top-down approach. The amount of manual work that is required to bring precision to an acceptable level is therefore justified.

3.3 Lectometric analysis

Semi-automatic retrieval of lexical variables in the Brown corpora with subsequent semantic verification in context yielded a dataset containing 140,681 observations, of which 11,847 were considered to be truly interchangeable and 20,161 were considered not to be interchangeable after manual screening (the remaining 128,834 observations were not manually verified and considered interchangeable by default[3]). This dataset served as input for the lectometric analysis consisting of the following three steps. The first step reshaped this dataset into a series of contingency tables per variable, with variants as columns and the intersections of the subcorpora as rows. An example of such a contingency table is shown in Table 2. The absolute frequencies of these contingency tables are transformed into relative frequencies per row (added in parentheses in Table 2), so that the preference for a certain variant can be observed for each intersection of subcorpora, independent of the amount of observations.

The second step measured the lexical distances between all possible pairs of language-external intersections, utilizing the City-Block distance metric (with the relative frequencies of the contingency table as input) as explained in Speelman, Grondelaers, and Geeraerts (2003). The output was stored in distance matrices. An

3. Default interchangeability is therefore nothing more than accepting the outcome of the Semantic Vector Space clustering without manual intervention.

Table 2. Contingency table for the variable *tv* versus *television*

	television	*tv*
Imaginative BrE 1960	11 (79%)	3 (11%)
Informative BrE 1960	55 (49%)	58 (51%)
Imaginative AmE 1960	6 (50%)	6 (50%)
Informative AmE 1960	44 (58%)	32 (42%)
Imaginative BrE 1990	14 (52%)	13 (48%)
Informative BrE 1990	101 (66%)	52 (34%)
Imaginative AmE 1990	11 (30%)	26 (70%)
Informative AmE 1990	100 (60%)	65 (40%)

example of such a distance matrix, which shows the distances between the intersections of subcorpora on the basis of the *tv* versus *television* variable, is shown in Table 3. The distances in matrices like these were further filtered by using the Log Likelihood Ratio test (Dunning 1993). If the Log Likelihood Ratio test indicated that the difference between two intersections was not significant at $p < 0.05$, we concluded that there was not enough evidence to consider both intersections to be statistically different. Therefore, the calculated distance between the intersections was set to zero. If, however, the Log Likelihood Ratio test indicated a statistically significant difference, the distance was retained. This safeguards against relying on frequencies that are, statistically speaking, not very trustworthy.

Table 3. Distance matrix on the basis of the City-Block distance for the variable *tv* versus *television*

	Imaginative BrE 1960	**Informative BrE 1960**	**Imaginative AmE 1960**	**Informative AmE 1960**	**Imaginative BrE 1990**	**Informative BrE 1990**	**Imaginative AmE 1990**	**Informative AmE 1990**
Imaginative BrE 1960	0.00	0.30	0.29	0.21	0.27	0.13	0.49	0.18
Informative BrE 1960	0.30	0.00	0.01	0.09	0.03	0.17	0.19	0.12
Imaginative AmE 1960	0.29	0.01	0.00	0.08	0.02	0.16	0.20	0.11
Informative AmE 1960	0.21	0.09	0.08	0.00	0.06	0.08	0.28	0.03
Imaginative BrE 1990	0.27	0.03	0.02	0.06	0.00	0.14	0.22	0.09
Informative BrE 1990	0.13	0.17	0.16	0.08	0.14	0.00	0.36	0.05
Imaginative AmE 1990	0.49	0.19	0.20	0.28	0.22	0.36	0.00	0.31
Informative AmE 1990	0.18	0.12	0.11	0.03	0.09	0.05	0.31	0.00

However, note that this approach does not guarantee that the calculated distances between the subcorpora are in themselves statistically significant. Indeed, the Log Likelihood Ratio test merely verifies that the absolute frequencies on which the distance metric is based come from different populations. A methodology to test the statistical significance of linguistic distances is presented in Delaere et al. (2012).

The third step applies INDSCAL. Some variables yielded very sparse distance matrices in step two, and these were discarded if all distances in the distance matrix were zero. This left us with 232 distance matrices as input for the INDSCAL analysis. The INDSCAL analysis was conducted in *R* (R Core Team 2012) by using the SmacofIndDif method in the SMACOF package by de Leeuw & Mair (2009). The dataset and the *R*-code used for the analysis are available from the first-named author.

4. Three-dimensional analysis results

We first discuss INDSCAL's Group Stimulus Space, which offers a bird's eye perspective on the aggregate behavior of all input variables. As the Brown corpora cover three lectal dimensions (standard variety, real time period, discourse type), INDSCAL was instructed to calculate a three-dimensional solution. The resulting three-dimensional Group Stimulus Space is depicted in Figure 1.

Inspection of the cube in Figure 1 shows that the three lectal dimensions are indeed captured by the INDSCAL analysis. The first dimension distinguishes between informative texts (on the left) and imaginative texts (on the right). The second dimension distinguishes between the American subcorpora (in the front of the cube) and the British subcorpora (in the back). The third dimension puts the subcorpora from the 1960s at the top and subcorpora from the 1990s at the bottom of the cube. The fact that our analysis picks up the lectal dimensions that it should serves to enhance confidence in our method. Notice that an analysis of mean distances by dimension indicates that lower-numbered dimensions in the INDSCAL solution depicted in Figure 1 capture more variation than the higher-numbered dimensions. In plain English, the three lectal dimensions may be ranked in terms of their overall importance in the following way: discourse type > standard variety > real time period.

We now move on to investigate how individual lexical variables relate to the three dimensions discussed in the foregoing discussion; recall that this sort of transparency is the added value of INDSCAL in contrast to traditional two-way MDS. By-variable Configuration Weights are reported in Table A (Appendix), where the variable labels are brought into correspondence with the numbers in

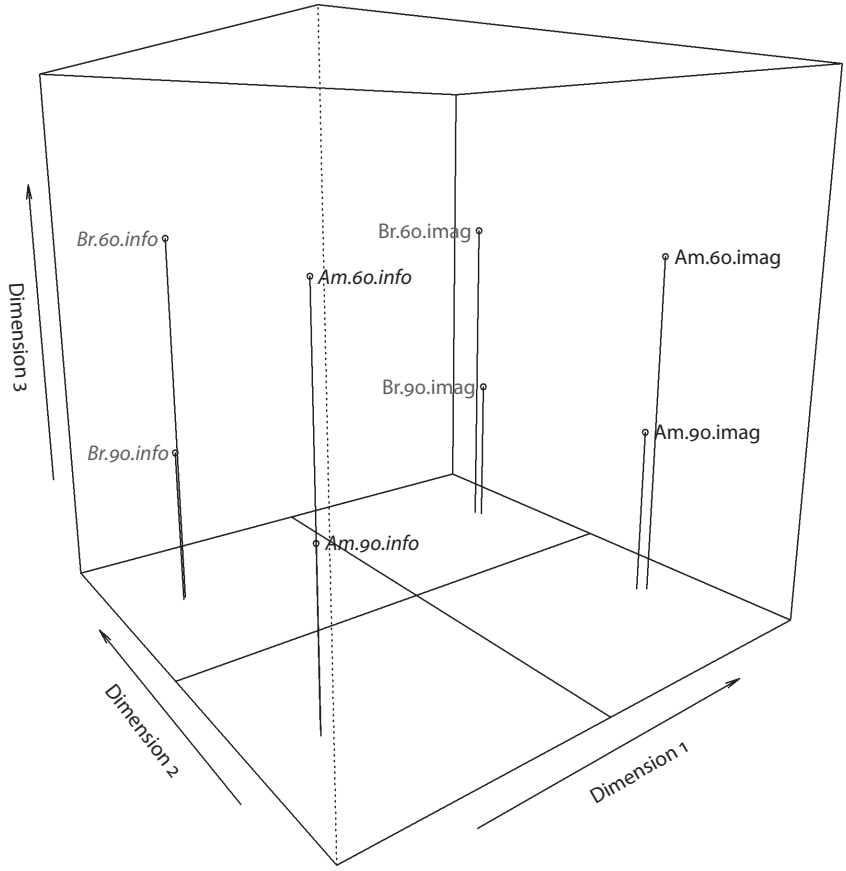

Figure 1. Group Stimulus Space showing the position of the subcorpora in three dimensions

Figure 2. The weights are visually depicted in Figure 3. The origin of the cube (at (1, 1, 1)) is located in the middle of the cluster of variables, just above variable 98 (see arrow). The scale of the dimensions is omitted, because the actual values are only relative with respect to 1. A value of 1, i.e. close to the origin near variable 98, indicates that the variable agrees with the overall structure of the Group Stimulus Space in Figure 1. Observe, first, that the variables are not scattered randomly in the cube but form clusters. This suggests that the lexical variables are differentially sensitive to language-external lectal dimensions. We performed a *k*-means cluster analysis (with the *kmeans* function in the *stats* package in *R*) on the Configuration Weights data that revealed, for several values of *k*, recurring groupings of variables. We will discuss three of these clusters in more detail below, but a full interpretation is beyond the scope of the current paper.

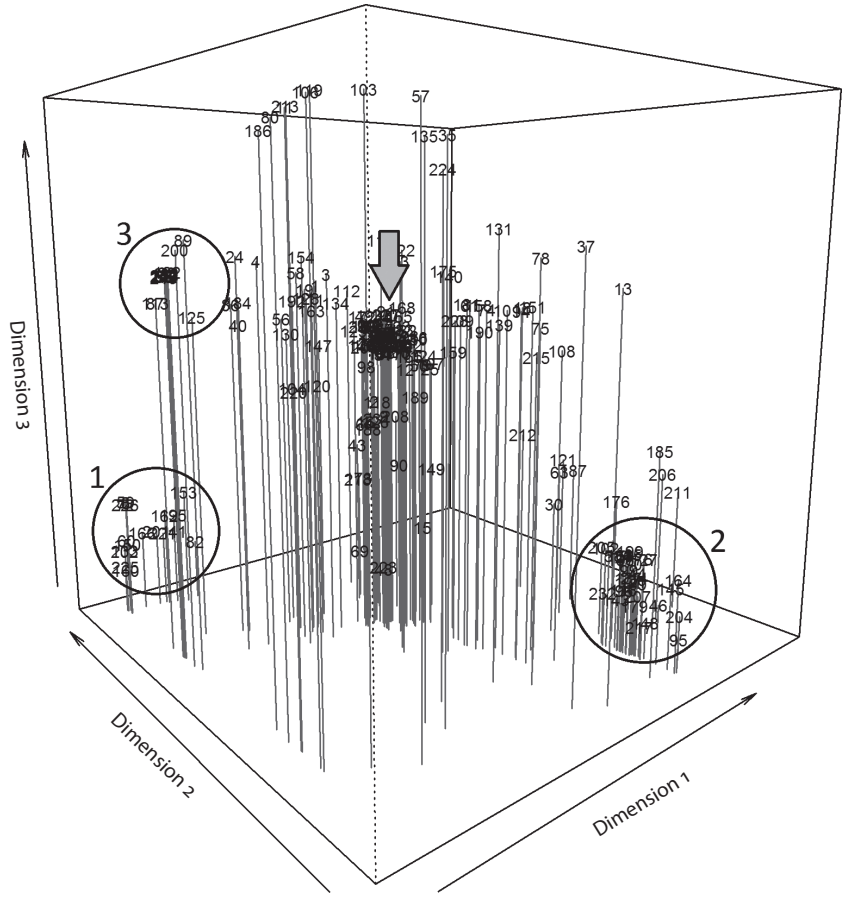

Figure 2. Configuration Weights of the individual lexical variables for the three-dimensional Group Stimulus Space

Consider, for example, the cluster of variables in the bottom-left-back corner ("1") of Figure 2. These variables all have high Configuration Weights for Dimension 2, but low Configuration Weights for Dimensions 1 and 3. Since Dimension 2 distinguishes between the American and British subcorpora, these variables are more sensitive than other variables to BrE/AmE differences. The cluster includes some of the usual suspects, such as *holiday*/*trip*, as in Example (2), but also alternations such as *ocean*/*sea*, as in Example (3), or *computer*/*pc*, as in Example (4).

The relative occurrence ratios reported in Table 4 show how the three variables differ considerably across the standard varieties. *Holiday* is clearly preferred in BrE, whereas *trip* is the AmE option. *Sea* is always preferred over *ocean*, but

Table 4. Frequency ratios for some variables from the first group of variables

	AmE	**BrE**
holiday/trip	48/192 = 0.3	197/122 = 1.6
sea/ocean	182/67 = 2.7	279/24 = 11.6
computer/pc	171/2 = 85.5	126/8 = 15.8

relatively speaking, *ocean* is more popular in AmE than in BrE. Similarly, *computer* is preferred over *pc* in both varieties, but *pc* is more frequently used in the UK.

(2) a. It was summer in England and our reunited family went down to Cooden for a *holiday* with Mummy brother his wife and our two girl cousins. (FLOB, G07, BrE90 Inf)
b. However Mr. Parichy and his bride will go to Vero Beach on their wedding *trip* and will stay in the John G. Beadles' beach house. (Brown, A16, AmE60 Inf)

(3) a. A picture flashes to mind: a graceful old three-arched bridge, a river flowing through a rocky valley to the *sea*. (LOB, E09, BrE60 Inf)
b. Jack arrived in California in late July, but with the exception of the composition of a Joycean poem written about the sounds of the *ocean*, his trip was a disaster. (Frown, G60, AmE90 Inf)

(4) a. But industrial computer companies have been slow to make such PLC modules for the *computer* industry perhaps because they lack familiarity with the PLC controls world. (Frown, E32, AmE90 Inf)
b. That the *pc* industry is in trouble is not in question. (FLOB, H28, BrE60 Inf)

Another cluster of variables can be found in the bottom-right-front corner of Figure 2 ("2"), where the Configuration Weights are high for Dimension 1, but not for Dimensions 2 and 3. In other words, these are variables which are above-average sensitive to the distinction between informative and imaginative texts, but are stable in terms of the other dimensions. Examples include *fog/mist*, as in Example (5), *floor/ground*, as in Example (6), and *cost/expense*, as in Example (7).

Table 5. Frequency ratios for some variables from the second group of variables

	Informative	**Imaginative**
fog/mist	36/12 = 3.0	39/24 = 1.6
ground/floor	421/279 = 1.5	192/308 = 0.6
cost/expense	1032/233 = 4.4	24/33 = 0.7

Table 5 provides the ratios for the relative occurrence of these variables in the informative and imaginative text type, respectively. *Fog* is the preferred option in both informative and imaginative texts, but in imaginative texts, *mist* is more likely to appear. The preferences for *ground* versus *floor* flip according to the text type: *ground* is preferred in informative texts, and *floor* is preferred in imaginative texts. Similarly, *cost* is preferred in informative texts and *expense* is more frequent in imaginative texts.

(5) a. The Midlands and the South were hit by black ice and freezing *fog*.
(FLOB, A24, BrE90 Inf)
b. His fatigues made a streak of almost phosphorescent green in the *mist*.
(Brown, N25, AmE60 Im)

(6) a. Mr. Parrillo dropped the gun which fired as it struck the *ground*. (Brown, A24, AmE60 Inf)
b. She would drop her clothes on the *floor* and join him.
(FLOB, K23, BrE90 Im)

(7) a. He vows to control *costs* with vague notions of insurance reform and the elimination of administrative waste and billing fraud.
(Frown, A14, AmE90 Inf)
b. But lacking money from commercial sponsors the stations have had difficulties meeting *expenses* or improving their service.
(Brown, B02, AmE60 Inf)

In the top-left-back corner of Figure 2's Configuration Weights cube we find variables that have a high Configuration Weight for both Dimension 2 and 3, but not for Dimension 1 ("3"). These variables are sensitive to an interaction between standard variety and real time, in that these variables have changed over time, but differently so in the two varieties. Examples include *therapy*/*treatment*, as in Example (8), *earnings*/*income*, as in Example (9), and *concentration*/*density*, as in Example (10).

Table 6 illustrates the interaction between period and variety. To illustrate, the ratio of *therapy*/*treatment* increases by a factor of 1.8 from 0.09 in the 1960s to

Table 6. Frequencies of some variables in the third group of variables

	1960s		**1990s**	
	AmE	**BrE**	**AmE**	**BrE**
therapy/*treatment*	13/133 = 0.1	3/120 = 0.0	25/153 = 0.2	14/140 = 0.1
earnings/*income*	19/96 = 0.2	49/127 = 0.4	37/174 = 0.2	45/118 = 0.4
density/*concentration*	3/1 = 3.0	0/4 = 0.0	1/3 = 0.3	9/0 = NaN

0.16 in the 1990s in AmE. In BrE, however, the ratio increases by a factor of four, from 0.025 in the 1960s to 0.1 in the 1990s. Similarly, in the case of *earnings* versus *income*, the ratio remains stable from the 1960s to the 1990s in BrE, but increases in AmE. The last example, *density* versus *concentration* is an extreme case. In both periods, AmE and BrE have inverse preferences; additionally, the two varieties change ends between the 1960s and the 1990s.

(8) a. In children it appears that *therapy* can safely be stopped after 3 years of relapse maintenance therapy. (Frown, J14, AmE90 Im)
b. The Duchess took a real interest in how Brenda was feeling and what *treatment* she had received here and how she had benefited. (FLOB, A42, BrE90 Inf)

(9) a. The value of the policy would be worked out entirely according to your husband's *earnings*. (LOB, R03, BrE60 Im)
b. Farmers spend more of their *income* on tractors and implements than on any other group of products. (Brown, A28, AmE60 Inf)

(10) a. Boats are operated in every state in the Union with the heaviest *concentrations* along both coasts and in the Middle West. (Brown, E06, AmE60 Inf)
b. These studies illustrate the important role of host population *density* in the response of a parasite transmission rate to thermal stress […]. (Frown, J08, AmE90 Inf)

There is also a cloud of lexical variables located right in the center of Figure 2, around the origin of the cube. These variables behave inconspicuously, that is to say, just as depicted in the Group Stimulus Space plot in Figure 1. In other words, these variables tend to have sensitivity to the three lectal dimensions under consideration that aligns with the average sensitivity, i.e. discourse type > standard variety > real time period. Examples include *despair*/*frustration*, as in Example (11), *dirt*/*mud*, as in Example (12), and *meal*/*snack*, as in Example (13).

(11) a. That word was withheld when the need of it seemed the measure of his *despair*. (Brown, G40, AmE60 Inf)
b. […] and then supports this with survey data reporting conscious experienced *frustrations*. (Brown, J63, AmE60 Inf)

(12) a. People took turns with the palas heaping *dirt* onto the coffin. (Frown, P13, AmE90 Im)
b. The stocky man knelt and dug his fingers into leather, *mud* and slush. (FLOB, M01, BrE90 Im)

(13) a. You can be having just anybody in for a *meal*. (FLOB, L02, BrE90 Im)
b. The two tall brothers waited silently while their mother handed Gran her cold *snack* and water jug […] (Brown, N13, AmE60 Im)

5. Conclusion

In this paper, we have offered the first-ever lectometrical analysis of lexical variation in written standard English as sampled in the Brown family of corpora. Our analysis has contributed to the literature both in terms of methodology and in terms of our knowledge about (lexical) variation in an aggregate perspective.

The main methodological achievement of our analysis is that we have applied Semantic Vector Space modeling to large corpora in order to generate — semi-automatically and in a bottom-up, yet theoretically responsible fashion (motto: "You shall know a word by the company it keeps"; Firth 1957: 11) — a representative, unbiased sample of lexical variables. The distribution of these variables was subsequently explored in the Brown corpora by transparently aggregating frequency patterns with the help of Individual Differences Scaling (INDSCAL). Whereas each of these methods is fairly well-known on their own, we believe that their combined power benefits aggregation studies. Granted, Semantic Vector Space modeling cannot yet idenitify lexical variables in a completely unsupervised fashion (at least not as reliably as corpus linguists would expect them to), and substantial manual verification was necessary at the time of writing this paper. Nonetheless, our analysis has captured those lectal dimensions — real time (1960s vs. 1990s), standard variety (AmE vs. BrE), and discourse function (imaginative vs. informative) — that it should have, given the design of the Brown corpora. What is interesting is how these dimensions are ranked: discourse type is the overall most important lectal dimension, while real time is the least important lectal dimension. Moreover, thanks to our transparent aggregation methodology, we saw that not all lexical variables are created equal: for example, *computer*/*pc* is particularly sensitive to the standard variety dimension, whereas *cost*/*expense* is particularly variable along the discourse type dimension. That we captured patterns like these highlights one of the strengths of our analysis technique, namely that it moves beyond a discussion of the usual suspects (*truck*/*lorry*, *pants*/*trousers*, and so on), instead generating variables in a data-driven fashion. We are confident that Semantic Vector Space research will progress towards overcoming current limitations, and that the data-driven spirit of the method holds great potential.

Our work has some interesting theoretical implications. We investigated over 300 lexical variables, and found that most variables are sensitive to at least one of the three language-external dimensions (variables insensitive to any

language-external dimension would have appeared in the front-left-bottom corner of Figure 3, which is empty). But, as it turns out, not every dimension has dedicated lexical variables: the real time dimension is not clearly associated with distinctive lexical variables. Adopting the definition of a language variety as "a set of linguistic items with similar distribution" (Hudson 1996: 22), this observation could be interpreted as an indication that informative and imaginative texts, as well as AmE and BrE, are distinct language varieties — registers and national lects, respectively — while 1960s and 1990s written English are not. We hasten to add that this may very well be due to the fact that the extensive corpora (the British National Corpus, the American National Corpus, and the Blog Authorship Attribution Corpus) on which we ran the Semantic Vector Space models do not sufficiently represent 1960s English usage. But be that as it may, the fact that short-term diachronic lexical variation is harder to capture, vis-à-vis functional and regional variation, by means of large corpus database does tell us something, we believe, about the possibly frailer nature of this particular lectal dimension. In a bird's eye perspective, lexical change in late 20th century written Standard English is just not that pronounced.

Acknowledgements

Funding by a KU Leuven Special Research Fund (BOF) grant [ZKB5610 OT/06/08], a travel grant by the Flemish Research Council (FWO) [V4.508.11N], and financial support by the Freiburg Institute for Advanced Studies (FRIAS) is gratefully acknowledged. This research is part of a PhD dissertation that was conducted under the supervision of Dirk Speelman and Dirk Geeraerts. We thank Lars Hinrichs, Koen Plevoets, Melanie Röthlisberger, and two anonymous reviewers for helpful comments and suggestions. All remaining errors are, of course, our own.

References

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. New York, NY: ACM Press.

Biber, D. (1988). *Variation acros Speech and Writing*. Cambridge, UK: Cambridge University Press. doi: 10.1017/CBO9780511621024

Biber, D. (1989). A typology of English texts. *Linguistics*, *27*(1), 3–42. doi: 10.1515/ling.1989.27.1.3

Bickel, B. (2007). Typology in the 21st century: Major current developments. *Linguistic Typology*, *11*(1), 239–251. doi: 10.1515/LINGTY.2007.018

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.

Borin, L., & Saxena, A. (2013). *Approaches to Measuring Linguistic Differences*. Berlin, Germany: Mouton de Gruyter. doi: 10.1515/9783110305258

Church, K. W., & Hanks, P. (1990). Word association, mutual information and lexicography. *Computational Linguistics*, *16*(1), 22–29.

Cysouw, M. (2005). Quantitative methods in typology. In G. Altmann, R. Köhler, & R. Piotrowski (Eds.), *Quantitative Linguistics: An International Handbook* (pp. 554–578) Berlin, Germany: Mouton de Gruyter.

de Leeuw, J., & Mair, P. (2009). Multidimensional scaling using Majorization: SMACOF in R. *Journal of Statistical Software*, *31*(3), 1–30.

Delaere, I., De Sutter, G., & Plevoets, K. (2012). Is translated language more standardized than non-translated language? Using profile-based correspondence analysis for measuring linguistic distances between language varieties. *Target: An International Journal of Translation Studies*, *24*(2), 203–224.

Dinu, G., Thater, S., Laue, S. (2012). A comparison of models of word meaning in context. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 611–615). Montréal, Canada: Association for Computational Linguistics.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, *19*(1), 61–74.

Firth, J. (1957). A synopsis of linguistic theory 1930-1955. In J. R. Firth (Ed.), *Studies in Linguistic Analysis* (pp. 1–32). Oxford, UK: Philological Society.

Francis, W., & Kucera, H. (1964). *Manual of Information to Accompany a Standard Corpus of Present-Day Edited American English, for Use with Digital Computers*. Providence, RI: Department of Linguistics, Brown University.

Geeraerts, D. (2010). *Theories of Lexical Semantics*. Oxford, UK: Oxford University Press.

Geeraerts, D., Grondelaers, S., & Bakema, P. (1994). *The Structure of Lexical Variation. Meaning, Naming, and Context*. Berlin, Germany: Mouton de Gruyter. doi:10.1515/9783110873061

Geeraerts, D., Grondelaers, S., & Speelman, D. (1999). *Convergentie en divergentie in de Nederlandse woordenschat. Een onderzoek naar kleding- en voetbaltermen*. Amsterdam, Netherlands: Meertens Instituut.

Goebl, H. (1984). *Dialektometrische Studien: Anhand italoromanischer, raetoromanischer und galloromanischer Sprachmaterialien aus AIS und ALF*. Tübingen, Germany: Max Niemeyer.

Grieve, J. (2007). Quantitative authorship attribution: An evaluation of techniques. *Literary and Linguistic Computing*, *22*(3), 251–270. doi:10.1093/llc/fqm020

Grieve, J., Speelman, D., & Geeraerts, D. (2011). A statistical method for the identification and aggregation of regional linguistic variation. *Language Variation and Change*, *23*(2), 193–221. doi:10.1017/S095439451100007X

Heeringa, W. (2004). *Measuring Dialect Pronunciation Differences using Levenshtein Distance*. (Unpublished doctoral dissertation). Groningen, Netherlands: Rijksuniversiteit Groningen.

Heylen, K., & Ruette, T. (2013). Degrees of semantic control in measuring aggregated lexical distances. In L. Borin & A. Saxena (Eds.), *Approaches to Measuring Linguistic Differences* (pp. 353–374). Berlin, Germany: Mouton de Gruyter.

Heylen, K., Speelman, D., & Geeraerts, D. (2012). Looking at word meaning. An interactive visualization of Semantic Vector Spaces for Dutch synsets. *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH* (pp. 16–26). Avignon, France: Association for Computational Linguistics.

Hinrichs, L., Smith, N., & Waibel, B. (2010). A manual of information for the part-of-speech-tagged 'Brown' corpora. *ICAME Journal*, *34*, 189–230.

Horan, C. (1969). Multidimensional Scaling: Combining observations when individuals have different perceptual structures. *Psychometrica*, *34*(2), 139–165. doi:10.1007/BF02289341

Hudson, R. (1996). *Sociolinguistics*. Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9781139166843

Hundt, M., Sand, A., & Siemund, R. (1999). *Manual of Information to Accompany the Freiburg-LOB Corpus of British English ("FLOB")*. Freiburg, Germany: Albert-Ludwigs-Universität Freiburg.

Johansson, S., Leech, G., & Goodluck, H. (1978). *Manual of Information to Accompany the Lancaster-Oslo/Bergen Corpus of British English, for use with Digital Computers*. Oslo, Norway: University of Oslo.

Labov, W. (1969). Contraction, deletion, and inherent variability of the English Copula. *Language 45*(4), 715–62. doi:10.2307/412333

Labov, W. (1972). *Sociolinguistic Patterns*. Oxford, UK: Blackwell.

Labov, W., Ash, S., & Boberg, C. (2006). *The Atlas of North American English. Phonetics, Phonology and Sound Change*. Berlin, Germany: Mouton de Gruyter.

Lavandera, B. (1978). Where does the sociolinguistic variable stop? *Language in Society 7*(2), 171–183. doi:10.1017/S0047404500005510

Navigli, R. (2012). A Quick tour of word sense disambiguation, induction and related approaches. In *Proceedings of the 38th Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM)* (pp. 115–129). Heidelberg, Germany: Springer-Verlag.

Nerbonne, J. (2009). Data-driven dialectology. *Language and Linguistics Compass 3*(1), 175–198. doi:10.1111/j.1749-818X.2008.00114.x

Nerbonne, J., & Kretzschmar, W. (2003). Introducing computational techniques in dialectometry. *Computers and the Humanities*, *37*(3), 245–255. doi:10.1023/A:1025064105053

Pado, S., & Lapata, M. (2007). Dependency-based construction of semantic space models. *Computational Linguistics*, *33*(2), 161–199. doi:10.1162/coli.2007.33.2.161

Pantel, P. (2003). *Clustering by committee*. (Unpublished doctoral dissertation). Alberta, Canada: University of Alberta.

Peirsman, Y. (2008). Word space models of semantic similarity and relatedness. In *Proceedings of the ESSLLI-2008 Student Session* (pp. 143–152). Hamburg, Germany.

Peirsman, Y. (2010). *Crossing corpora.* (Unpublished doctoral dissertation). Leuven, Belgium: University of Leuven.

Peirsman, Y., Deyne, S. D., Heylen, K., & Geeraerts, D. (2008). The construction and evaluation of word space models. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2008)*. Marrakech, Morocco.

Peirsman, Y., Geeraerts, D., & Speelman, D. (2015). The corpus-based identification of cross-lectal synonyms in pluricentric languages. *International Journal of Corpus Linguistics*, *20*(1), 54–80. doi:10.1075/ijcl.20.1.03pei

Plevoets, K., Speelman, D., & Geeraerts, D. (2008). The distribution of T/V pronouns in Netherlandic and Belgian Dutch. In K. Schneider & A. Barron (Eds.), *Variational Pragmatics: A Focus on Regional Varieties in Pluricentric Languages* (pp. 181–210). Amsterdam, Netherlands: John Benjamins Publishing Company. doi:10.1075/pbns.178.09ple

Quine, W. V. O. (1951). Two dogmas of empiricism. *The Philosophical Review*, *60*, 20–43. doi:10.2307/2181906

R Core Team. (2012). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Reppen, R., Ide, N., & Suderman, K. (2005). *American National Corpus (ANC)*. Philadelphia, PA: Linguistic Data Consortium.

Ruette, T. (2012). *Aggregating Lexical Variation: Towards large-scale lexical lectometry*. (Unpublished doctoral dissertation). Leuven, Belgium: University of Leuven.

Ruette, T., Geeraerts, D., Peirsman, Y., & Speelman, D. (2014). Semantic weighting mechanisms in scalable lexical sociolectometry. In B. Szmrecsanyi & B. Wälchli (Eds.), *Aggregating Dialectology and Typology: Linguistic Variation in Text and Speech, within and across Languages* (205–230). Berlin, Germany: Mouton de Gruyter.

Ruette, T., & Speelman, D. (2014). Transparent aggregation of variables with individual differences scaling. *Literary and Linguistic Computing*, *29*(1), 89–106. doi:10.1093/llc/fqt011

Schler, J., Koppel, M., Argamon, S., & Pennebaker, J. (2006). Effects of age and gender on blogging. In *Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*. Palo Alto, California.

Schneider, E. (1988). Qualitative vs. quantitative methods of area delimitation in dialectology: A comparison based on lexical data from Georgia and Alabama. *Journal of English Linguistics 21*(1), 175–212.

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

Sinclair, J. (1991). *Corpus, Concordance, Collocations*. Oxford, UK: Oxford University Press.

Speelman, D., Grondelaers, S., & Geeraerts, D. (2003). Profile-based linguistic uniformity as a generic method for comparing language varieties. *Computers and the Humanities*, *37*, 317–337. doi:10.1023/A:1025019216574

Szmrecsanyi, B. (2011). Corpus-based dialectometry: A methodological sketch. *Corpora*, *6*(1), 45–76. doi:10.3366/cor.2011.0004

Szmrecsanyi, B. (2013). *Grammatical Variation in British English Dialects: A Study in Corpus-Based Dialectometry*. Cambridge, UK: Cambridge University Press.

Takane, Y., Young, F., & de Leeuw, J. (1977). Nonmetric individual differences multidimensional scaling: An alternating least squares method with optimal scaling features. *Psychometrika*, *42*(1), 7–67. doi:10.1007/BF02293745

Turney, P., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, *37*, 141–188.

Wälchli, B., & Szmrecsanyi, B. (2014). Introduction: The text-feature-aggregation pipeline in variation studies. In B. Szmrecsanyi & B. Wälchli (Eds), *Aggregating Dialectology, Typology, and Register Analysis: Linguistic Variation in Text and Speech* (1–25). Berlin, Germany: Mouton de Gruyter. doi:10.1515/9783110317558.1

Wieling, M., & Nerbonne, J. (2011). Bipartite spectral graph partitioning for clustering dialect varieties and detecting their linguistic features. *Computer Speech and Language*, *25*(3), 700–715. doi:10.1016/j.csl.2010.05.004

Wieling, M., Nerbonne, J., & Baayen, H. (2011). Quantitative social dialectology: Explaining linguistic variation geographically and socially. *PLoS ONE*, *6*(9), e23613. doi:10.1371/journal.pone.0023613

Woolhiser, C. (2005). Political borders and dialect divergence/convergence in Europe. In P. Auer & F. Kerswill (Eds.), *Dialect Change. Convergence and Divergence in European Languages* (pp. 236–262). Cambridge, UK: Cambridge University Press.

Zauner, A. (1902). *Die romanischen Namen der Körperteile: Eine onomasiologische Studie.* (Unpublished doctoral dissertation). Erlangen, Germany: Universität Erlangen.

Appendix

By-variable Configuration Weights

Lexical variable	**Index**	**Dim. 1**	**Dim. 2**	**Dim. 3**
acquisition, *purchase*	1	1.02	2.36	3.52
ad, *advertisement*	2	1.64	2.49	2.17
advertisement, *advertising*	3	1.04	2.24	3.64
adviser, *consultant*	4	0.59	2.47	3.78
agreement, *consent*	5	2.55	0.69	0.74
agreement, *treaty*	6	1.40	2.26	3.06
aid, *assistance*	7	1.52	2.18	2.97
aim, *objective*	8	0.01	3.75	1.05
alliance, *cooperation*	9	1.72	2.46	1.99
alliance, *union*	10	1.63	2.12	2.78
alteration, *change*	11	0.02	1.03	5.04
analysis, *examination*	12	1.70	2.14	2.58
anger, *rage*	13	1.88	0.10	3.62
animal, *creature*	14	1.55	2.26	2.76
anybody, *anyone*	15	2.03	2.34	0.62
apartment, *flat*	16	1.62	2.60	1.91
approach, *method*, *procedure*	17	1.43	2.22	3.07
approach, *strategy*	18	1.87	1.00	3.36
area, *district*	19	0.97	2.43	3.48
area, *domain*	20	1.63	2.08	2.86
army, *troops*	21	2.58	0.44	0.55
aspect, *attribute*, *characteristic*, *feature*	22	1.58	2.20	2.78
ass, *asshole*, *bastard*, *idiot*	23	1.34	1.73	3.78
assault, *attack*	24	0.44	2.50	3.83
assault, *raid*	25	1.81	1.97	2.60
assay, *assessment*, *evaluation*, *measurement*	26	1.46	2.33	2.87
assembly, *committee*	27	2.56	0.55	0.77
assembly, *council*	28	1.50	2.18	2.99
asset, *capital*, *fund*	29	1.39	2.24	3.09

Lexical variable	Index	Dim. 1	Dim. 2	Dim. 3
asset, *equity*	30	2.42	1.16	1.18
assistant, *secretary*	31	1.48	2.18	2.98
association, *union*	32	1.50	2.18	2.99
atom, *particle*	33	0.00	2.68	3.69
attorney, *counsel*, *solicitor*	34	1.54	2.16	2.95
audience, *viewer*	35	0.86	0.62	4.87
author, *writer*	36	1.63	2.08	2.84
autonomy, *freedom*	37	1.67	0.23	3.98
avenue, *lane*, *street*	38	1.59	2.09	2.90
back, *rear*	39	1.50	2.17	3.01
backyard, *yard*	40	0.66	2.91	3.08
bathroom, *toilet*	41	0.44	3.75	0.57
beach, *coast*, *shore*	42	1.67	2.00	2.87
beef, *steak*	43	1.62	2.73	1.61
belief, *doctrine*	44	2.58	0.43	0.63
belief, *ideology*, *principle*	45	1.46	2.09	3.19
bloke, *chap*, *lad*	46	2.60	0.25	0.36
boat, *ship*	47	1.45	2.12	3.11
bowl, *pot*	48	1.85	2.67	0.03
boy, *guy*	49	1.35	2.21	3.19
breeze, *wind*	50	1.74	2.02	2.65
buck, *pound*	51	1.56	2.25	2.75
buddy, *pal*	52	1.45	2.25	2.99
business, *company*, *firm*	53	1.55	2.18	2.85
business, *corporation*	54	2.56	0.53	0.79
cab, *taxi*	55	1.53	2.17	2.96
cafe, *pub*	56	0.92	2.71	3.14
campaign, *rally*	57	0.50	0.41	5.10
cancer, *tumor*	58	0.87	2.39	3.66
capability, *potential*	59	0.01	3.75	1.06
cdna, *chromosome*, *gene*	60	0.02	3.80	0.58
cdna, *genome*	61	1.74	1.42	3.35
championship, *tournament*	62	1.54	2.15	2.96
change, *modification*	63	2.42	1.16	1.18
chapel, *church*	64	1.62	2.61	1.88
characteristic, *feature*, *trait*	65	2.56	0.47	0.83
chart, *map*	66	1.50	2.17	2.97
chick, *gal*, *girl*	67	1.61	2.10	2.87

Lexical variable	Index	Dim. 1	Dim. 2	Dim. 3
chief, *director*, *president*	68	1.55	2.23	2.78
chunk, *slice*	69	1.73	2.87	0.23
circumstance, *context*	70	1.63	2.10	2.77
city, *town*	71	1.69	2.05	2.74
class, *lesson*	72	0.01	3.75	1.02
client, *customer*	73	1.54	2.24	2.79
cloth, *fabric*	74	1.79	1.29	3.31
clothes, *clothing*	75	1.97	0.91	3.16
coat, *jacket*	76	2.57	0.37	0.81
combination, *mix*	77	1.50	2.15	3.04
commitment, *involvement*	78	1.69	0.65	3.86
community, *group*	79	1.76	2.00	2.65
competitor, *opponent*	80	0.01	1.17	4.98
component, *element*	81	2.56	0.56	0.78
computer, *pc*	82	0.58	3.72	0.41
concentration, *density*	83	0.00	2.72	3.65
concept, *notion*	84	1.59	2.09	2.90
conference, *congress*	85	1.52	2.17	2.97
conflict, *dispute*	86	0.55	2.82	3.32
consequence, *result*	87	0.00	2.90	3.38
constitution, *legislation*	88	0.00	2.72	3.66
constraint, *limit*, *restriction*	89	0.02	2.43	4.01
consumer, *customer*	90	1.86	2.45	1.38
cord, *rope*	91	1.44	2.05	3.25
correspondent, *reporter*	92	1.40	2.25	3.08
corridor, *hall*	93	1.56	2.00	3.08
corridor, *hallway*	94	1.88	1.02	3.32
cost, *expense*	95	2.61	0.06	0.08
cost, *fee*	96	2.58	0.56	0.40
couch, *sofa*	97	1.49	2.23	2.95
country, *nation*	98	1.51	2.41	2.59
county, *province*	99	2.57	0.51	0.64
court, *tribunal*	100	1.46	2.08	3.20
cousin, *nephew*	101	1.85	1.13	3.32
crap, *damn*, *fuck*, *hell*	102	2.54	0.71	0.81
crap, *shit*	103	0.25	0.58	5.14
crime, *offence*	104	1.15	2.97	2.31
criminal, *offender*	105	1.48	2.26	2.87

Lexical variable	Index	Dim. 1	Dim. 2	Dim. 3
criterion, *requirement*	106	0.02	0.79	5.12
culture, *tradition*	107	1.39	2.25	3.07
dad, *father*	108	2.07	0.79	2.94
danger, *hazard*	109	1.76	1.54	3.18
delight, *pleasure*	110	0.00	2.70	3.68
demonstration, *protest*	111	1.04	2.47	3.36
despair, *frustration*	112	1.59	2.09	2.90
development, *evolution*	113	0.00	2.90	3.38
device, *tool*	114	1.46	2.13	3.08
diagram, *graph*	115	1.54	2.16	2.95
difference, *variation*	116	2.58	0.56	0.40
dinner, *supper*	117	1.54	2.17	2.93
dirt, *mud*	118	1.65	2.47	2.16
discipline, *subject*	119	0.01	0.75	5.13
disease, *illness*	120	1.30	2.82	2.34
diversity, *variation*	121	2.42	1.16	1.18
door, *doorway*	122	0.01	2.68	3.69
drive, *ride*	123	1.60	2.09	2.88
duty, *responsibility*	124	1.74	1.96	2.75
earnings, *income*	125	0.30	2.96	3.20
earnings, *revenue*	126	1.68	2.53	1.92
effectiveness, *efficiency*	127	2.57	0.37	0.81
emotion, *feeling*	128	1.01	2.49	3.38
employee, *staff*	129	1.34	2.34	3.02
enquiry, *investigation*	130	0.99	2.76	2.97
entrance, *gate*	131	1.51	0.86	4.11
enzyme, *protein*	132	0.00	2.70	3.68
estate, *land*	133	0.00	3.81	0.45
everybody, *everyone*	134	1.18	2.34	3.32
exam, *quiz*, *testing*	135	0.81	0.76	4.86
exam, *test*	136	1.39	2.26	3.08
expansion, *growth*	137	0.00	2.72	3.66
expenditure, *investment*	138	1.46	2.33	2.84
expert, *specialist*	139	1.88	1.24	3.16
expertise, *skill*	140	1.57	1.45	3.64
extent, *range*	141	1.53	2.18	2.95
federation, *union*	142	0.00	2.72	3.65
film, *flick*	143	2.58	0.63	0.26

Lexical variable	Index	Dim. 1	Dim. 2	Dim. 3
film, *movie*	144	1.51	2.17	2.98
floor, *ground*	145	2.60	0.14	0.59
fog, *mist*	146	1.43	2.36	2.85
football, *soccer*	147	1.22	2.66	2.83
force, *strength*	148	2.60	0.36	0.12
forest, *wood*	149	2.01	2.19	1.37
format, *template*	150	0.02	3.80	0.53
fuck, *shit*	151	1.88	0.90	3.37
fuel, *gas*	152	1.46	2.26	2.91
fuel, *petrol*	153	0.48	3.68	1.06
funding, *grant*	154	0.84	2.26	3.82
funding, *subsidy*	155	1.69	2.08	2.74
game, *match*	156	1.57	2.16	2.84
garden, *park*	157	1.50	2.19	2.93
gig, *show*	158	1.76	1.33	3.36
giggle, *laugh*	159	1.85	1.73	2.81
glass, *jar*	160	0.01	3.82	0.22
growth, *increase*	161	2.55	0.58	0.78
ha, *hmm*, *huh*	162	0.33	3.75	0.80
happiness, *joy*	163	1.07	2.52	3.24
hey, *yo*	164	2.59	0.07	0.71
hill, *mountain*	165	1.51	1.96	3.20
holiday, *trip*	166	0.15	3.79	0.63
hotel, *inn*	167	1.54	2.25	2.76
human, *man*	168	1.50	1.92	3.30
hurt, *pain*	169	2.58	0.47	0.50
idiot, *jerk*, *moron*	170	2.57	0.51	0.71
idiot, *loser*	171	1.51	2.17	3.00
impact, *influence*, *role*	172	1.35	2.24	3.18
importance, *significance*	173	1.67	2.79	1.18
improvement, *progress*	174	1.62	2.05	2.91
income, *pay*, *salary*	175	1.53	1.46	3.69
income, *wage*	176	2.58	0.56	0.40
institute, *institution*	177	2.58	0.51	0.55
investor, *shareholder*	178	0.00	2.72	3.65
involvement, *participation*	179	2.59	0.46	0.26
issue, *problem*	180	1.65	1.96	2.94
killing, *murder*	181	1.73	1.43	3.36

Lexical variable	Index	Dim. 1	Dim. 2	Dim. 3
km, *mile*	182	1.58	2.01	3.04
lady, *woman*	183	1.56	2.03	3.04
lake, *pond*	184	0.60	2.79	3.34
laugh, *laughter*	185	2.41	0.15	2.03
launch, *release*	186	0.03	1.34	4.90
law, *regulation*	187	2.58	0.56	0.40
lorry, *truck*	188	1.65	2.60	1.82
meal, *snack*	189	1.82	2.17	2.25
merchant, *trader*	190	1.86	1.43	3.07
metre, *yard*	191	2.58	0.52	0.50
missile, *rocket*	192	0.93	2.55	3.35
mom, *mother*	193	1.54	2.25	2.77
observation, *finding*, *outcome*	194	1.51	2.21	2.94
ocean, *sea*	195	0.40	3.73	0.79
offence, *violation*	196	2.56	0.58	0.73
offer, *proposal*	197	1.79	1.93	2.67
officer, *policeman*	198	1.59	2.00	2.97
organisation, *planning*	199	2.56	0.49	0.85
output, *product*	200	0.00	2.51	3.91
pants, *trousers*	201	0.27	3.78	0.62
parameter, *variable*	202	0.00	3.81	0.45
percentage, *proportion*, *ratio*	203	1.46	2.35	2.83
person, *someone*	204	2.61	0.06	0.33
perspective, *view*	205	2.59	0.46	0.26
perspective, *viewpoint*	206	2.58	0.51	0.55
prize, *trophy*	207	2.59	0.43	0.38
profit, *surplus*	208	1.77	2.37	2.00
pupil, *student*	209	1.36	2.25	3.07
region, *territory*	210	0.00	2.72	3.65
region, *zone*	211	2.59	0.46	0.26
representative, *spokesman*	212	2.42	1.16	1.18
researcher, *scholar*	213	0.02	1.01	5.04
revenue, *turnover*	214	2.58	0.46	0.58
river, *stream*	215	2.06	1.03	2.84
road, *route*	216	1.64	1.95	2.97
sadness, *sorrow*	217	2.60	0.41	0.05
sculpture, *statue*	218	1.67	2.79	1.16
seller, *vendor*	219	0.00	2.73	3.64

Lexical variable	Index	Dim. 1	Dim. 2	Dim. 3
share, *stock*	220	1.17	2.97	2.26
shop, *store*	221	0.32	3.77	0.58
something, *stuff*	222	1.30	1.65	3.90
song, *tune*	223	1.85	2.68	0.05
success, *triumph*, *win*	224	1.06	0.90	4.62
supporter, *voter*	225	0.00	3.82	0.26
tax, *taxation*	226	0.01	3.75	1.02
teacher, *tutor*	227	1.46	2.06	3.21
television, *tv*	228	1.74	1.59	3.17
therapy, *treatment*	229	0.00	2.72	3.66
timber, *wood*	230	1.55	2.20	2.86
trip, *vacation*	231	1.68	2.52	1.95
wow, *yea*, *yeah*	232	2.56	0.75	0.29

Authors' addresses

Tom Ruette
Faculty of Arts
KU Leuven
Blijde Inkomststraat 21
3000 Leuven
Belgium

tom.ruette@kuleuven.be

Benedikt Szmrecsanyi
Faculty of Arts
KU Leuven
Blijde Inkomststraat 21
3000 Leuven
Belgium

benedikt.szmrecsanyi@kuleuven.be

Katharina Ehret
Hermann Paul School of Linguistic
Universität Basel und Freiburg (i. Br.)
Belfortstr. 18
D-79098 Freiburg
Germany

katharina.ehret@hpsl.uni-freiburg.de