Analyzing aggregated linguistic data

Benedikt Szmrecsanyi
University of Freiburg
bszm@frias.uni-freiburg.de

1 Introduction

AGGREGATE DATA ANALYSIS – also known as DATA SYNTHESIS, MASS DATA ANALYSIS, or, especially in biology, (NUMERICAL) TAXONOMY – is concerned not with the distribution of individual features, properties, or measurements, but with the joint analysis of multiple characteristics. Aggregate data analysis is a methodical cornerstone in many academic disciplines: taxonomists, for instance, typically categorize species not on the basis of a single morphological or genetic criterion, but of many; economists assess macroeconomic changes not on the basis of individual macroeconomic indicators (unemployment, say), but also consider inflation, GDP per capita, interest rates, and so on. Outside academia, aggregate data analysis is quite customary in fields such as marketing research and consumer creditworthiness modeling.

By contrast, in the realm of linguistics and particularly in variationist linguistics, there is a long and strongly entrenched tradition of looking at individual features in isolation rather than at feature aggregates. This is why we find an abundance of what Nerbonne (2008) has referred to as 'single-feature-based studies' in the pertinent literature. Browse through the discipline's flagship journals and you will find a great many studies with dauntingly specific titles such as of *The glottal stop in language A*, *Auxiliary contraction in variety B*, *The use of abstract nouns in register C*, or *Quotatives in sociolect D*. It is, essentially, only in three linguistic subfields that we find aggregate data analysis employed on a regular basis: in the study of cross-linguistic typology and language universals (for instance, Greenberg, 1963), in dialectometry (Goebl, 1982; Nerbonne, Heeringa & Kleiweg, 1999; Séguy, 1971), and multidimensional register studies in the spirit of Biber (1988; see also Biber & Gray, this volume).

The aim of this chapter is to sketch ways of analyzing aggregated linguistic data. Rather than providing a step-by-step manual, it endeavors to inspire readers to think and work holistically. Thus in section 2, I discuss the rationale behind aggregate data analysis, its range of applications, and its limitations. Section 3 offers a concise cooking recipe for aggregate linguistic analysis. Subsequently, I present three case studies to exemplify the methodology: in section 4, I show how text frequencies of grammatical markers in naturalistic corpus data can be aggregated to establish a register typology of analyticity-syntheticity profiles. In section 5, I present an aggregate methodology for investigating the role that geography plays in structuring morphosyntactic variability in British English dialects. In section 6, I demonstrate how an aggregate, survey-based approach can help to uncover the network structure of World Englishes. Section 7 offers some concluding remarks.

2 Why aggregate analysis? Range of applications and limitations

Succinctly put, aggregate analysis is appropriate whenever the analyst's attention is turned to the forests, not the trees; this is what I will refer to as the AGGREGATE PERSPECTIVE. Forests, along these lines, may be languages, regional language varieties, stylistic language varieties, or any other multidimensional object. If it is the individual trees (i.e. linguistic phenomena) that matter, the FEATURE-CENTERED PERSPECTIVE is called for.

To illustrate the two perspectives, Tagliamonte & Smith (2005) conduct a feature-centered study that takes an interest in the zero complementizer in dialectal English. The paper affords important insights about this particular feature by studying it in a number of British English dialects. Crucially, the paper does not purport that we can characterize those particular dialects from studying this particular feature alone – and indeed we cannot, because the dialects have many other features that we would also need to consider if our interest was with the dialects as such. By contrast, adopting an aggregate multidimensional perspective on register variation in English, Biber (1988) also studies the zero complementizer (THAT deletion, in his parlance), but as just one of a large number of features that gang up to characterize text types in the aggregate perspective.

It is clear that both research designs have their merits, conditioned on the research questions being asked. Crossing designs and research questions is problematic, however: attempts to characterize multidimensional objects (e.g. regional or stylistic varieties) by just looking at one particular feature, such as the zero complementizer, are flawed. The reason is that picking out just one particular feature is highly subjective (why this feature and not the many other features that could have been studied?), by virtue of there being no guarantee that two varieties A and B exhibit the same distributional behavior in regard to different features.

In short, the aggregate perspective is fairly imperative whenever the analyst formulates a research question about forests (languages, varieties, and the like). The inherent) limitation is that by studying aggregates (forests), the analyst looses sight of particular features (trees). A practical limitation of the aggregate perspective is feasibility: ideally, the aggregate analysis would seek to include all available data. Because there is typically a large to infinite number of linguistic features that could be used to characterize a language or language variety, often a choice has to be made when defining a feature portfolio. A certain degree of subjectivity is therefore, alas, inevitable.

3 Aggregate linguistic analysis: a cooking recipe

In the most general terms, four steps are necessary to conduct an aggregate linguistic analysis:

1. *Define the list of features on which to base the aggregate analysis.* The name of the game is to consider as many features as possible to ensure comprehensiveness and to avoid subjectivity.

2. *Create a feature matrix with* N $\times$ p *dimensions* (*N*: number of objects, i.e. languages, varieties, or texts; *p*: number of features). When tapping into pre-existing data (e.g.

surveys, dialect atlases, etc.), the dimensionality of the dataset is usually dictated by the data source. If the analyst draws, e.g., on naturalistic corpus data, this step might entail compiling the corpus and extracting the features (or their frequencies) from the corpus.

3. *Aggregate*. As a rule, $N \times p$ feature matrices are unwieldy (especially if p is large). This fact of life calls for the application of some sort of aggregation or dimension reduction technique. In this spirit, the analyst may generate an $N \times N$ distance matrix (which abstracts away from features and specifies pairwise linguistic distances) via some distance measure, or draw on aprioristic categorization and grouping schemes.
4. *Analyze, visualize, and interpret.*

4 Aggregating part-of-speech frequencies: analyticity vs. syntheticity in British English text types

The first case study is a loose paraphrase of some of the research reported in Szmrecsanyi (2009), a study that is interested in aggregate intralingual variability in terms of OVERT GRAMMATICAL ANALYTICITY (i.e. the text frequency of free grammatical markers) and OVERT GRAMMATICAL SYNTHETICITY (the text frequency of bound grammatical markers). Among other things, the paper investigates text type variability along these parameters in the *British National Corpus* (BNC), and it is this line of stylistic variability that will exemplarily concern us in this section.[1]

4.1 Defining the list of features

The empirical basis for the investigation is the BNC's part-of-speech annotation. The catalogue of features to be considered was therefore a function of the design of the BNC tag set (cf. Aston & Burnard, 1998), which spans 55 major part-of-speech tags: for example, adjectives (tag `AJ0`), plural common nouns (tag `NN2`), and the past tense form of the verb DO (tag `VDD`). These 55 tags – more specifically, their text frequencies in the corpus texts – are the features on which the aggregate analysis is based.

4.2 Creating the feature matrix

The BNC samples 34 spoken and written macro registers, such as spontaneous face-to-face conversation (genre classification code `S_conv`) or fiction (genre classification code `W_fict`). Custom-made scripts written in Perl (*Practical Extraction and Report Language*) (cf. Schwartz, Phoenix & Foy, 2008) took on the heavy lifting and queried all 4,052 individual BNC texts (each annotated for membership in one of the macro registers) in regard to the text frequencies, normalized to a frequency per 1,000 words of running text (*ptw*), of every one of the 55 part-of-speech tags. The scripts subsequently calculated normalized text frequencies on the macro register level and generated, as

[1] For various extensions of this line of research see Kortmann & Szmrecsanyi (2009, 2011) and Szmrecsanyi & Kortmann (2009c, 2011).

output, a .csv (comma separated values) spreadsheet. This spreadsheet details a 34 × 55 feature matrix: 34 macro registers, every one of them characterized by a vector of 55 normalized part-of-speech frequencies.

4.3 Aggregation

Subsequently, the 55 part-of-speech tags were classified into three categories (analytic tokens, synthetic tokens, purely lexical tokens). Summing up – and thus, aggregating – tag frequencies per category then yielded a set of two Greenberg-inspired indices (cf. Greenberg, 1960), an *analyticity index* and a *syntheticity index*. Observe, along these lines, that the terms 'analytic' and 'synthetic' have a long and distinguished tradition in linguistic inquiry, a fact which makes an ad hoc classification of grammatical markers relatively unproblematic. In this spirit, *grammatical analyticity* was defined as comprising all those coding strategies where grammatical information is conveyed by free grammatical markers. Free grammatical markers, in turn, were defined as synsemantic word tokens that have no independent lexical meaning. Formal *grammatical syntheticity* was defined as comprising all those coding strategies where grammatical information is signaled by bound grammatical markers. These definitions give rise to the following category/tag matches:

- *analytic tags or tokens*: conjunctions, subjunctions, and prepositions (tags `CJ*, PRF, PRP`); determiners, articles, and *wh*-words (`D*, AT0, AVQ, PNQ`); existential *there* (`EX0`); pronouns (`PNI, PNP, PNX`); the tokens *more* and *most*; the infinitive marker *to* (`TO0`); modals (`VM0`); the negator *not* (`XX0`), auxiliary BE (`VB*+V*, VB*+*+V*, VB*+XX0`), auxiliary DO (`VD*+V*, VD*+*+V*, VD*+XX0`), and auxiliary HAVE (`VH*+V*, VH*+*+V*, VH*+XX0`)
- *synthetic tags or tokens*: the *s*-genitive (`POS`); comparative and superlative adjectives (`AJC, AJS`); plural nouns (`NN2`); plural reflexive pronouns (`PNX` + word token ending in **ves*); inflected verbs (`V*D, V*G, V+N, V*Z`)

Subsequently, the analyst sums up tag frequencies, thus obtaining index scores which are normalized to a sample size of 1,000 words of running text. Hence the analyticity index measures text frequencies of conjunctions, subjunctions, and prepositions plus text frequencies of determiners, articles, and *wh*-words plus the text frequency of existential *there*, and so on; the syntheticity index calculates the text frequency of the *s*-genitive plus text frequencies of comparative and superlative adjectives plus text frequencies of plural nouns, and so on. Note that this step reduces the original 34 × 55 feature matrix to a 34 × 2 index matrix (34 registers, each characterized by an analyticity and a syntheticity index score).

4.4 Analyzing, visualizing, and interpreting

[*** insert Figure 1 here ***]

Figure 1[2] is a so-called scatter plot that visualizes the 34 × 2 index matrix, plotting analyticity index scores (vertical axis) against syntheticity index scores (horizontal axis). A closer look at the extreme cases along the two dimensions in the diagram is instructive. In the syntheticity dimension, with index scores beyond 190, we find institutional documents and news texts as the most synthetic genres in the BNC. At the other end of the spectrum, it is public debate and demonstrations that turn out to be the least synthetic text types in the BNC. The extreme data points in the analyticity dimension are sermons and advertisements. Sermons, for one thing, are extremely analytic (analyticity index: 548). Example (3) exemplifies this genre:

(3) Why not have the light within you so you don't have to go and get it outside but it's there dwelling within you, day by day, moment by moment? And he longs to meet this woman's need. And we can try all sorts of things. And there's, there's things are not necessarily wrong, there's the legitimate things, erm, wi within our work, th there's a, there's job satisfaction, but there's more to that than, in life than just job satisfaction. <BNC text KN8>

We are dealing in (3) with a relatively high degree of reference tracking via pronouns (*you, it, he, we*), many prepositions (e.g. *within, by, in*), and much repetition of analytic material (for instance, multiple repetition of existential/dummy *there*). Compare this to (4), an advertisement illustrating the BNC's least analytic text type (analyticity index: 379):

(4) Build up a total heating system room by room Interested? USE THE POST-FREE COUPON OVERLEAF. Total Heating. Forget fuel deliveries, dust, dirt, smells, noise, fetching, carrying, tending the boiler. Get a new electric boiler and forget it – all of it! <BNC text HT1>

In (4), it is obvious that all non-essential material is dispensed with, thanks to a genre-specific pressure for output economy that can be quantified, as it were, in monetary terms. This pressure appears to affect analytic material in particular.

Particularities of individual registers aside, Figure 1 indicates a couple of interesting generalizations. First, we note that there are significant correlations between the index levels for individual text types and some of the dimensions of register variation identified by Douglas Biber (see, e.g., Biber, 1988). The relevant Biberian dimensions are *involved* vs. *informational production* and *abstract* vs. *non-abstract information*. The technicalities need not concern us here (cf. Szmrecsanyi, 2009 for a discussion); suffice it to point out that increased analyticity correlates with involved production whereas increased syntheticity correlates with abstract informational content. Observe, however, how Figure 1 suggests that these correlations ultimately boil down to the following very robust differences between spoken and written text types:

1. Spoken texts are significantly more analytic than written texts. The average spoken text exhibits 50 more analytic markers per 1,000 words of running text than the typical written text.

[2] All plots in this contribution were created using the software package SPSS. Note the open-source statistical analysis package R would have been equally suitable.

2. Written texts are significantly more synthetic than spoken texts, in that the former exhibit, on average, approximately 30 more synthetic markers per 1,000 words of running text than the latter.
3. As for the scope of variability, variability among written texts is more sizable than variability among spoken texts: in Figure 1, the cloud embedding spoken genres is substantially more compact than its written counterpart.

By way of an interim summary, we have seen in this section how an aggregation of part-of-speech frequencies in the BNC – informed by two parameters, analyticity and syntheticity, well-known from the cross-linguistic classification of languages – can reveal important differences between spoken and written text types. Crucially, the approach offered here is more encompassing and informative than gauging text type variability on the basis of individual features or variants (say, the distribution of the analytic *of*-genitive versus the synthetic *s*-genitive by medium). The limitation is that the methodology utilized in this section does not *per se* tell us which grammatical markers are most robustly implicated in overall analyticity-syntheticity variability, but note that it is always possible to deconstruct the indices, as it were, to get a hold on which grammatical markers are especially variable. Szmrecsanyi (2009) demonstrates that it is primarily frequency fluctuations in pronouns, negators, auxiliary DO/HAVE, and modals which cause the bulk of variability.

5 Aggregating text frequencies of dialect features: determinants of morphosyntactic variability in British English dialects

In our second case study, we set out to explore determinants of aggregate morphosyntactic variability in traditional British English dialects, centering on factors such as geographic distance, travel time, and Peter Trudgill's notion of 'linguistic gravity'. Unlike in the previous case study, we will rely not on an aprioristic (i.e. analytic vs. synthetic) aggregation method, but will make use of a theory-neutral, statistical distance measure (Euclidean distance) to calculate aggregate dialect distances. As for the general methodological orientation of this research, this section is an exercise in CORPUS-BASED DIALECTOMETRY (see also Szmrecsanyi, 2008, 2011; Szmrecsanyi & Wolk, 2011). Dialectometry (for seminal work, see Séguy, 1971; Goebl, 1982; Nerbonne et al., 1999) is the branch of geolinguistics concerned with measuring, visualizing, and analyzing aggregate dialect similarities or distances:

> Dialectometry is not concerned with the analysis or the discussion of single or a very few dialect features. Instead it offers a methodology to discern general, seemingly hidden structures from a larger amount of features. (Goebl & Schiltz, 1997: 13)

Crucially, orthodox dialectometry draws on linguistic atlas data (typically describing accent differences) as its primary source of information (see also Kretzschmar, this volume). By contrast, I shall seek in this section to combine the philologically responsible corpus-based study of morphosyntactic variability in British English dialects with aggregational-dialectometrical analysis techniques. In this spirit, I will tap the *Freiburg English Dialect Corpus* (henceforth: FRED) (Hernández, 2006; Szmrecsanyi &

Hernández, 2007). FRED spans 2.5 million words of running text, consisting of samples (mainly transcribed so-called 'oral history' material) of dialectal speech from a variety of sources. Typically, a fieldworker interviews an informant about life, work etc. in former days. The 431 informants sampled in the corpus are typically elderly people with a working-class background. The interviews were conducted in 162 different locations (that is, villages and towns) in 38 different pre-1974 counties in Great Britain plus the Isle of Man and the Hebrides. The level of areal granularity investigated in the present study will be the county level. From the 38 counties sampled in FRED, I removed four counties with comparatively thin coverage (< 5,000 words of running text), leaving us with a geographical network of 34 counties subject to analysis in this section. Note that longitude/latitude information is available for each of the locations sampled in FRED.

5.1 Defining the list of features

True to the spirit of dialectometrical analysis, the overarching aim was to include as many phenomena as possible, the rationale being that a "large number of variables, even though they will contain a great deal of variation irrelevant to questions of geographic or social conditioning, will nonetheless provide the most accurate picture of the relations among the varieties examined" (Nerbonne, 2006: 464). To this purpose, I canvassed the dialectological, variationist, and corpus-linguistic literature, and identified suitable phenomena. The criteria for inclusion of a candidate feature in the catalogue were the following:

1. To ensure statistical robustness of text frequencies, the feature had to have a raw frequency of at least 100 hits in FRED as a whole (this rules out interesting but infrequent dialect phenomena such as double modals).
2. The feature also had to be extractable subject to a reasonable input of labor resources by a human coder. This is why, for example, many hard-to-retrieve null phenomena such as zero relativization are not considered in the catalogue.
3. In the case of the particular dataset analyzed in this section, the feature had to be a deeply vernacular and broad dialect feature, defined as a feature that has a text frequency of < 1 per ten thousand words (*pttw*) in a corpus of standard colloquial English (our reference data source was the conversational section `[s1a]` of the British component of the *International Corpus of English*).

I thus arrived at a list of 17 non-standard morphosyntactic dialect features, which are listed in Appendix A, along with linguistic examples.

5.2 Creating the feature matrix

The next step involved extracting the relevant feature frequencies from FRED. Some features in the catalogue are sufficiently ‘surfacy' to be extractable without human intervention (for instance, feature [10]: the negator *ain’t*). In such cases, retrieval scripts written in Perl established the relevant text frequencies automatically, generating a .csv spreadsheet detailing feature frequencies (normalized to frequency *pttw*) per FRED county. A number of features in the catalogue (for example, feature [13]: *don’t* with 3rd person singular subjects) required manual disambiguation prior to extraction via Perl scripts (Szmrecsanyi, 2010b spells out the coding guidelines). Subsequently, the resulting text frequencies were *log* transformed (a customary procedure to de-emphasize large frequency differentials and to alleviate the effect of frequency outliers) and arranged in a 34 ×17 dimensional frequency matrix (34 counties, each characterized by a vector of 17 discrete text frequencies).

5.3 Aggregation

By way of aggregation, the 34 × 17 frequency matrix was transformed into a 34 × 34 distance matrix (similar to distance tables available in, e.g., road atlases), which abstracts away from individual feature frequencies and specifies pair-wise distances between the dialects considered. The measure used to calculate these distances was the well-known EUCLIDEAN DISTANCE MEASURE (see, for instance, Aldenderfer & Blashfield, 1984: 25-26), where the distance between two dialects is defined as the square root of the sum of all 17 squared frequency differentials.[3] I emphasize here that the Euclidean distance measure is maximally straightforward computationally and theory-neutral in that all features receive the same weight in the distance calculation. The mean distance in the 34 × 34 matrix is 3.6 Euclidean distance points (minimum: .9 points, maximum: 6.3 points, standard deviation: .9 points).

5.4 Analyzing, visualizing, and interpreting

[*** insert Map 1 here ***]

Map 1 projects the 34 × 34 Euclidean distance matrix to geography. As a so-called LINK MAP, the dialectometrical projection connects counties that are close morphosyntactically by darker lines, and morphosyntactically more distant counties by lighter lines (for presentational purposes, Map 1 omits links between counties/locations that are more than 250km apart).[4] Visual inspection reveals that the links in England are overall darker than in Scotland. This means that we are dealing with a network of comparatively strong and coherent morphosyntactic links in England, and with a somewhat looser network structure in Scotland.

All in all, Map 1 suggests that there is some geographic structure in dialectal variability. Let us now quantify the correlation between aggregate morphosyntactic distances and the following three language-external distance measures:

- AS-THE-CROW-FLIES DISTANCE. Using a trigonometry formula on the FRED county coordinates, pair-wise as-the-crow flies distances may be calculated.[5] Notice that as-

[3] The distance matrix was calculated using SPSS, but note that any statistical software package, such as, e.g., R, could have been utilized instead.

[4] The link maps were created using the maplink module, which is part of Peter Kleiweg's R*u*G/L04 dialectometry software package (available online and for free at http://www.let.rug.nl/~kleiweg/L04/).

[5] The R*u*G/L04 dialectometry software package comes with a module (ll2dst) that can do this job automatically. Geographic county coordinates (mean longitude and latitude) were

the-crow-flies distance is the most common geographic distance measure in the dialectological and dialectometrical literature.

- LEAST-COST TRAVEL TIME. To calculate this measure, I turned to Google Maps (`http://maps.google.co.uk/`), which has a route finder facility that allows the user to enter longitude/latitude pairings for two coordinate pairs to obtain a least-cost travel route and, crucially, an estimate of the total travel time. Google Maps was queried for all 34 × 33/2 = 561 county/county pairings, thus obtaining pair-wise least-cost-travel time estimates.
- LINGUISTIC GRAVITY. In a (1974) paper, Peter Trudgill suggested a gravity model to account for geographic diffusion. Trudgill conjectured that "the interaction (*M*) of a centre *i* and a centre *j* can be expressed as the population of *i* multiplied by the population of *j* divided by the square of the distance between them" (1974: 233). Using Trudgill's formula on the FRED county coordinates and a standard spreadsheet application, linguistic gravity values were calculated for every one of the 561 county/county pairings in our database, feeding in least-cost travel time as geographic distance measure and early 20th century population figures by county[6] (in thousand) as a proxy for speaker community size.

Every one of theses language-external distance measures yields a 34 × 34 distance matrix which can be quantitatively correlated (utilizing any statistical software package) with the linguistic 34 × 34 Euclidean distance matrix. It is to this task that we turn next.

[*** insert Figure 2 here ***]

[*** insert Table 1 here ***]

In Figure 2, we find three scatter plots that visualize the relationship between aggregate morphosyntactic distances (vertical axis) and the three language-external distance measures (horizontal axes). Table 1 reports the corresponding Pearson correlation coefficients – a measure of the strength of dependence between two variables, ranging between -1 (a maximal negative relationship) to +1 (a maximal positive relationship) – as well as R^2 values, a measure indicating the proportion of variance in the dependent variable (in our case, aggregate morphosyntactic distances) accounted for by the independent variables (in the context of the present study, the language-external distance measures). As can be seen from the slope of the smoother curves in Figure 2 and the sign of the Pearson correlation coefficients in Table 1, the relationship between the language-external variables and aggregate morphosyntactic distances is the theoretically expected one: increased as-the-crow-flies distance and increased least-cost travel time predicts increased morphosyntactic dialectal distance; conversely, increased linguistic gravity implicates decreased dialectal distance. As for the relative strengths of these correlations, it turns out that as-the-crow-flies distance is the weakest predictor, accounting for 11.3% of the morphosyntactic variance; least-cost travel time fares only minimally better, explaining 11.8% of the overall variance; and Trudgill's notion of

calculated by computing the arithmetic mean of all the location coordinates associated with individual interview texts in FRED.

[6] Specifically, I used 1901 figures, as published in the *Census of England and Wales, 1921* and the *Census of Scotland, 1921*. These documents are available online at http://histpop.org/.

linguistic gravity explains 14.4% of the overall variance when modeled logarithmically (see. Szmrecsanyi to appear for a more detailed discussion of this issue).

From aggregating 17 dialect feature frequencies and correlating aggregate morphosyntactic distances with three language-external variables, we have learned that the linguistic distance between two dialects increases with increasing geographic distance, but that this effect is counterbalanced by population size: large speaker communities will tend to interact linguistically more than smaller speaker communities, all other things (and especially geographic distance) being equal. Note now the analysis offered here is empirically fairly robust in that it considers joint variance of *many* dialect features, and not just one.

6 Aggregating survey responses: World Englishes from a bird's eye perspective

In our third and final case study, we leave the comparatively neat and orderly realm of geographically adjacent traditional British English dialects and foray into the somewhat more heterogeneous but exciting universe of World Englishes – be they L1 varieties, indigenized L2 varieties, or English-based pidgin and creole languages. Drawing on analysis and interpretation techniques first presented in Szmrecsanyi & Kortmann (2009a,b), we shall be specifically concerned in this section with large-scale (read: aggregate) patterns and generalizations that emerge when investigating morphosyntactic variation in World Englishes from a bird’s eye perspective. Observe that unlike the two previous case studies, which were corpus-based, the analysis in this section will explore the questionnaire-based morphosyntax survey coming with the *Handbook of Varieties of English* (Kortmann, Schneider, Burridge, Mesthrie & Upton, 2004). On a more methodological note, the measure utilized to derive pair-wise aggregate distances between World Englishes will be the number of discordant feature classifications.

6.1 The list of features

The list of features feeding into the subsequent aggregate analysis is dictated by the design of the *Handbook*'s morphosyntax survey (cf. Kortmann & Szmrecsanyi, 2004 for details). Kortmann and Szmrecsanyi compiled a catalogue of 76 features and sent out this catalogue to the authors of the chapters in the morphosyntax volume of the *Handbook*. For each of these 76 features, the contributors were asked to specify into which of the following three categories the relevant feature falls:

A pervasive (possibly obligatory) or at least very frequent

B exists but a (possibly receding) feature used only rarely, at least not frequently

C does not exist or is not documented

40 *Handbook* authors responded and sent in data on 46 non-standard varieties of English. These varieties are from all seven anglophone world regions (British Isles, America, Caribbean, Australia, Pacific, Asia, Africa) and represent a fair mix of L1 varieties (such as New Zealand English), indigenized L2 varieties (e.g. Butler English), and English-based pidgin and creole languages (for example, Gullah). The survey features are

numbered from 1 to 76 (see Appendix B for the feature catalogue in its entirety) and cover 11 broad areas of morphosyntax.

6.2 Creating the feature matrix

Unlike binary contrasts or continuous variables, tripartite discrete classification systems (such as the survey's original 'A' – 'B' – 'C' scheme) are not trivial to handle statistically. As a first step towards an aggregate analysis, we therefore conflate 'A' responses ('pervasive') and 'B' responses ('exists) into an 'attested' category, to which we assign the numerical value '1'. The 'C' category ('does not exist') is assigned the numerical value '0'. Next, we create a spreadsheet with the binary feature values ('0' vs. '1') in columns and varieties in rows. We thus obtain a 46 × 76 feature matrix: 46 World Englishes, each characterized by 76 binary feature classifications.

6.3 Aggregation

To convert the 46 × 76 feature matrix into a 46 × 46 distance matrix specifying pair-wise aggregate distances, we may utilize the SQUARED EUCLIDEAN DISTANCE measure, defined as the sum of all squared feature differentials.[7] An interpretationally convenient property of this particular distance measure is that when applied to binary data (where contrasts are specified as '0' vs. '1'), pair-wise distances correspond numerically to the number of discordant feature classifications. To illustrate: Scottish English and Irish English share 57 (of 76) feature classifications; with regard to 19 features, their classifications differ. So, in the distance matrix, their distance is 19 squared Euclidean distance points. Observe that in the resulting 46 × 46 distance matrix as a whole, the mean distance is 31.5 squared Euclidean distance points (minimum: 6 points, maximum: 58 points, standard deviation: 8 points).

6.4 Analyzing, visualizing, and interpreting

Applying correlation techniques along the lines of those that were presented in the previous case study, we find that geography (specifically, as-the-crow-flies distance) explains only 3.6 per cent of the overall variance in aggregate morphosyntactic distances between World Englishes. If it is not areal proximity that is important here, then, what other factors are?

To explore this issue, we will now turn to an analysis technique known as MULTIDIMENSIONAL SCALING (MDS) (cf. Kruskal & Wish, 1978 for the technicalities). The fact of the matter is that on the interpretational plane, distance matrices are fairly unwieldy entities – in the present case, every one of the 46 varieties of English considered is characterized by its distance to the other 45 varieties in the dataset. MDS takes as its input the original 46 × 46 distance matrix and seeks to reduce its dimensionality on the condition that the ensuing information loss be minimized. Here, we

[7] This distance measure is available in all standard statistical software packages, such as SPSS and R.

will be interested in a low-dimensional 46 × 2 MDS solution, which can be visualized – in a manner that is more accessible to human cognition – in a two-dimensional plane.[8]

[*** insert Figure 3 here ***]

Figure 3 plots the resulting MDS plot. In the case at hand, the correlation between the 46 × 2 MDS matrix and the original 46 × 46 squared Euclidean distance matrix yields a Pearson correlation coefficient of .86, which is another way of saying that the plot in Figure 3 captures approximately .86 × .86 = 74 per cent of the variance in the original squared Euclidean distance matrix, which is a rather good value. The plot works like a geographic map: the further two data points are apart, the more dissimilar (in geographic terms, distant) they are. If two pairs of points are equally close or distant, the pairs of varieties they represent are equally (dis-)similar. The interesting fact about Figure 3 is, then, that it groups varieties fairly consistently according to variety type: notice that we find native L1 varieties (white dots) towards the top left corner of the diagram, English-based Pidgin and Creole languages (grey diamonds) are situated towards the bottom right corner of the diagram, and indigenized L2 varieties (black dots) are sandwiched, as it were, in between. Outliers are rare, but do exist and are plausible considering their variety genesis (cf. Szmrecsanyi & Kortmann, 2009a for an in-depth discussion).

[*** insert Figure 4 here ***]

The paramount role that variety type plays in structuring aggregate morphosyntactic variability in World Englishes is further highlighted when applying HIERARCHICAL AGGLOMERATIVE CLUSTER ANALYSIS (cf. Aldenderfer & Blashfield, 1984) to the data set.[9] Cluster analysis can group a large number of objects (e.g. varieties of English) into a smaller number of discrete and meaningful clusters on the basis of aggregate distances between those objects. The resulting classification can be visually represented using tree diagrams, also known as DENDROGRAMS, where one finds individual varieties to the left and successively larger clusters as one moves rightwards. Essentially, dendrograms work in much the same way as family trees. The dendrogram visualizing our dataset is shown in Figure 4.[10] Starting at the right and moving leftwards, the most basic split occurs between a cluster spanning predominantly non-L1 varieties of English (in the dendrogram, AbE through WhSafE), on the one hand, and a cluster uniting L1 varieties of English (AppE through WelE), on the other. Next, the non-L1 cluster is split up into a cluster containing primarily pidgin and creole languages (AbE through NigP), and a cluster principally encompassing indigenized L2 varieties of English (Bislama through WhSafE). At this level of granularity, we have arrived at the tripartite division (L1 vs. L2 vs. pidgin/creole languages) already familiar from the MDS plot in Figure 3. This does

[8] MDS can be conducted using standard statistical software packages, such as SPSS and R. It is also implemented in the R*u*G/L04 package (module mds), which was actually utilized here.

[9] Cluster analysis is implemented in all standard statistical software packages. I drew on the R*u*G/L04 package (modules cluster and den) to conduct the analysis.

[10] Observe in this connection that there are quite a few clustering algorithms. While the dendrogram in Figure 4 was created using 'Ward's Minimum Variance Method', it should be noted that other popular algorithms – such as the 'Weighted Pair Group Method using Arithmetic Averages' or the 'Complete Link Method' – generate strikingly similar dendrograms.

not mean that there are no areal effects at all – there clearly are (for instance, in the L1 cluster in Figure 4, all American varieties are grouped together in a sub-cluster). It is just that variety type appears to have primacy over areal effects.

The main implication, then, of the aggregate analysis offered in this section is that morphosyntactic similarities and distances between World Englishes are primarily a function of variety type. It is, I believe, fair to say that the thrust of this large-scale generalization would be fairly hard to come by adopting a single-feature approach. Of course, subject to the limits of the methodology, I have had nothing to say about those individual non-standard features that are prominently involved in making the difference. Therefore, by adopting a feature-centered perspective, the aggregate perspective offered here can be nicely complemented by an analysis of what Szmrecsanyi & Kortmann (2009b) call 'varioversals'. The term refers to features that are highly characteristic of specific varieties; for example, feature [50] (*no* as preverbal negator) turns out to be a highly distinctive pidgin and creole feature.

7 Concluding remarks

Our point of departure was that the single-feature-centered perspective (cf. Nerbonne, 2008) implicit in the bulk of variationist research is woefully inadequate for characterizing multidimensional linguistic objects such as languages, dialects, registers, and so on. The reason is that the next feature down the road may or may not contradict the characterization suggested by the previous feature. Aggregate data analysis mitigates this problem by analyzing joint variance of many features – and in joint variance, noise and feature-specific quirks cancel themselves out. The comprehensiveness and empirical robustness inherent in the aggregate perspective is certainly worth the trouble – having to collect data on many features, and having to deal with numbers and statistics galore – incurred by the methodology.

As we have seen, the aim of aggregate data analysis is to uncover sweeping generalizations. In this spirit, our case studies have suggested that, first, written English is robustly more synthetic and less analytic than spoken English; second, that aggregate morphosyntactic variability in traditional British English dialects is best explained by considering least-cost travel time between dialect localities as well as speaker community sizes; and third, that the crucial factor for predicting grammatical distances between World Englishes is variety type. Generalizations like these come at a cost, however, which is that the aggregate analyst inevitably loses sight of individual features with perhaps interesting distributions. As always, then, the smart thing to do is to aim for methodological pluralism: the aggregate perspective should *complement* the feature-centered perspective without replacing it.

The case studies discussed in this contribution do not exhaust the range of possible applications. Needless to say, the feature portfolios feeding into the aggregate analysis do not have to be morphological or syntactic, as they were in the case studies presented in this contribution. Instead, the phenomena considered may as well be, e.g., phonetic and phonological (cf. Heeringa, 2004), lexical (cf. Viereck, 1986), or even content-analytic (cf. Goldschmidt & Szmrecsanyi, 2007) in nature. In point of fact, the features do not even have to be concrete but can be fairly abstract: Longobardi & Guardiano (2009) aggregate cross-linguistic parameter settings along the lines of Chomsky’s 'Principles and Parameters' framework (cf. Chomsky, 1981), and Szmrecsanyi (2010a) aggregates probabilistic regression weights to elucidate short-term diachronic drifts of factors affecting genitive choices. The fact of the matter is that there are – quite literally – few limits to aggregate analysis.

Further reading

International Journal of Humanities and Arts Computing 2(1-2). Special Issue "Language Variation", ed. by John Nerbonne, Charlotte Gooskens, Sebastian Kürschner, and Renée van Bezooijen. 2008.

Lingua 119(11). Special issue "The Forests behind the Trees", ed. by John Nerbonne and Franz Manni. 2009

Appendix A: Aggregate variability in traditional British dialects – the feature catalogue

Pronouns and determiners

1. non-standard reflexives (e.g. *they didn't go theirself*)
2. archaic *thee/thou/thy* (e.g. *I tell thee a bit more*)
3. archaic *ye* (e.g. *ye'd dancing every week*)

Tense and aspect

4. the present perfect with auxiliary BE (e.g. *I'm come down to pay the rent*)

Verb morphology

5. *a*-prefixing on -*ing*-forms (e.g. *he was a-waiting*)
6. non-standard weak past tense and past participle forms (e.g. *they knowed all about these things*)
7. non-standard past tense *done* (e.g. *you came home and done the home fishing*)
8. non-standard past tense *come* (e.g. *he come down the road one day*)

Negation

9. the negative suffix -*nae* (e.g. *I cannae do it*)
10. the negator *ain't* (e.g. *people ain't got no money*)
11. multiple negation (e.g. *don't you make no damn mistake*)
12. *never* as past tense negator (e.g. *and they never moved no more*)

Agreement

13. *don't* with 3rd person singular subjects (e.g. *if this man don't come up to it*)
14. absence of auxiliary BE in progressive constructions (e.g. *I said, How you doing?*)

Relativization

15. the relative particle *what* (e.g. *the man what read the book*)

Complementation

16. *as what* or *than what* in comparative clauses (e.g. *we done no more than what other kids used to do*)
17. unsplit *for to* (e.g. *it was ready for to go away with the order*)

Appendix B: Aggregate variability in World Englishes – the feature catalogue
NOTE: For a version of the feature catalogue annotated with linguistic examples, see Kortmann and Szmrecsanyi (2004: 1146-1148)

Pronouns, pronoun exchange, and pronominal gender

1. *them* instead of demonstrative *those*
2. *me* instead of possessive *my*
3. special forms or phrases for the second person plural pronoun
4. regularized reflexives-paradigm
5. object pronoun forms serving as base for reflexives
6. lack of number distinction in reflexives
7. *she*/*her* used for inanimate referents
8. generic *he/his* for all genders
9. *myself/meself* in a non-reflexive function
10. *me* instead of *I* in coordinate subjects
11. non-standard use of *us*
12. non-coordinated subject pronoun forms in object function
13. non-coordinated object pronoun forms in subject function

Noun phrase

14. absence of plural marking after measure nouns
15. group plurals
16. group genitives
17. irregular use of articles
18. postnominal *for*-phrases to express possession
19. double comparatives and superlatives
20. regularized comparison strategies

Verb phrase: tense & aspect

21. wider range of uses of the progressive
22. habitual *be*
23. habitual *do*
24. non-standard habitual markers other than *do*
25. levelling of difference between Present Perfect and Simple Past
26. *be* as perfect auxiliary
27. *do* as a tense and aspect marker
28. completive/perfect *done*
29. past tense/anterior marker *been*
30. loosening of sequence of tense rule
31. *would* in *if*-clauses
32. *was sat/stood* with progressive meaning
33. *after*-Perfect

Verb phrase: modal verbs

34. double modals
35. epistemic *mustn't*

Verb phrase: verb morphology

36. levelling of preterite and past participle verb forms: regularization of irregular verb paradigms
37. levelling of preterite and past participle verb forms: unmarked forms
38. levelling of preterite and past participle verb forms: past form replacing the participle
39. levelling of preterite and past participle verb forms: participle replacing the past form
40. zero past tense forms of regular verbs
41. *a*-prefixing on *ing*-forms

Adverbs

42. adverbs (other than degree modifiers) have same form as adjectives
43. degree modifier adverbs lack -*ly*

Negation

44. multiple negation / negative concord
45. *ain't* as the negated form of *be*
46. *ain't* as the negated form of *have*
47. *ain't* as generic negator before a main verb
48. invariant *don't* for all persons in the present tense
49. *never* as preverbal past tense negator
50. *no* as preverbal negator
51. *was–weren't* split
52. invariant non-concord tags

Agreement

53. invariant present tense forms due to zero marking for the third person singular
54. invariant present tense forms due to generalization of third person -*s* to all persons
55. existential / presentational *there's, there is, there was* with plural subjects
56. variant forms of dummy subjects in existential clauses
57. deletion of *be*
58. deletion of auxiliary *have*
59. *was/were* generalization
60. Northern Subject Rule

Relativization

61. relative particle *what*
62. relative particle *that* or *what* in non-restrictive contexts
63. relative particle *as*
64. relative particle *at*
65. use of analytic *that his/that's, what his/what's, at's, as'* instead of *whose*
66. gapping or zero-relativization in subject position
67. resumptive / shadow pronouns

Complementation

68. *say*-based complementizers
69. inverted word order in indirect questions
70. unsplit *for to* in infinitival purpose clauses
71. *as what / than what* in comparative clauses
72. serial verbs

Discourse organization and word order

73. lack of inversion / lack of auxiliaries in *wh*-questions
74. lack of inversion in main clause *yes/no* questions
75. *like* as a focussing device
76. *like* as a quotative particle

References

Aldenderfer, Mark S. & Blashfield, Roger K. (1984). *Cluster Analysis.* Newbury Park, London, New Delhi: Sage Publications.

Aston, Guy & Burnard, Lou (1998). *The BNC Handbook: Exploring the British National Corpus with SARA.* Edinburgh: Edinburgh University Press.

Biber, Douglas (1988). *Variation across Speech and Writing.* Cambridge: Cambridge University Press.

Chomsky, Noam (1981). *Lectures on government and binding.* Dordrecht, Cinnaminson: Foris Publications.

Goebl, Hans (1982). *Dialektometrie: Prinzipien und Methoden des Einsatzes der Numerischen Taxonomie im Bereich der Dialektgeographie.* Wien: Österreichische Akademie der Wissenschaften.

Goebl, Hans & Schiltz, Guillaume (1997). A dialectometrical compilation of CLAE 1 and CLAE 2: Isoglosses and dialect integration. In Viereck, W. & Ramisch, H. (Eds.), *Computer developed linguistic atlas of England (CLAE).* Tübingen: Max Niemeyer Verlag. 13-21.

Goldschmidt, Nils & Szmrecsanyi, Benedikt (2007). What do economists talk about? A linguistic analysis of published writing in economic journals. *The American Journal of Economics and Sociology* 66(2): 335-378.

Greenberg, Joseph H. (1960). A quantitative approach to the morphological typology of language. *International Journal of American Linguistics* 26(3): 178-194.

Greenberg, Joseph H. (1963). Some Universals of Grammar with Particular Reference to the Order of Meaningful Elements. In Greenberg, J. H. (Ed.), *Universals of Language.* Cambridge, Mass.: MIT Press. 58-90.

Heeringa, Wilbert (2004). *Measuring dialect pronunciation differences using Levenshtein distance*. PhD dissertation, University of Groningen.

Hernández, Nuria (2006). *User's Guide to FRED.* Available online: `http://www.freidok.uni-freiburg.de/volltexte/2489/`. University of Freiburg.

Kortmann, Bernd, Schneider, Edgar, Burridge, Kate, Mesthrie, Raj & Upton, Clive (Eds.) (2004). *A Handbook of Varieties of English.* Berlin/New York: Mouton de Gruyter.

Kortmann, Bernd & Szmrecsanyi, Benedikt (2004). Global synopsis: morphological and syntactic variation in English. In Kortmann, B., Schneider, E., Burridge, K., Mesthrie, R. & Upton, C. (Eds.), *A Handbook of Varieties of English.* Berlin/New York: Mouton de Gruyter. 1142-1202.

Kortmann, Bernd & Szmrecsanyi, Benedikt (2009). World Englishes between simplification and complexification. In Siebers, L. & Hoffmann, T. (Eds), *World Englishes -- Problems, Properties and Prospects: selected papers from the 13th IAWE conference*. Amsterdam: Benjamins, 265-285.

Kortmann, Bernd & Szmrecsanyi, Benedikt (2011). Parameters of morphosyntactic variation in World Englishes: prospects and limitations of searching for universals. In Siemund, P. (Ed), *Linguistic Universals and Language Variation*. Berlin/New York: De Gruyter Mouton, 264-290.

Kruskal, Joseph B. & Wish, Myron (1978). *Multidimensional Scaling.* Newbury Park, London, New Delhi: Sage Publications.

Longobardi, Giuseppe & Guardiano, Cristina (2009). Evidence for syntax as a signal of historical relatedness. *Lingua* 119(11): 1679-1706.

Nerbonne, John (2006). Identifying Linguistic Structure in Aggregate Comparison. *Literary and Linguistic Computing* 21(4): 463-475.

Nerbonne, John (2008). Variation in the aggregate: an alternative perspective for variationist linguistics. In Dekker, K., MacDonald, A. & Niebaum, H. (Eds.), *Northern Voices: Essays on Old Germanic and Related Topics offered to Professor Tette Hofstra.* Leuven: Peeters. 365-382.

Nerbonne, John, Heeringa, Wilbert & Kleiweg, Peter (1999). Edit Distance and Dialect Proximity. In Sankoff, D. & Kruskal, J. (Eds.), *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison.* Stanford: CSLI Press. v-xv.

Schwartz, Randal L., Phoenix, Tom & Foy, Brian D. (2008). *Learning Perl.* Beijing, Sebastopol: O'Reilly.

Séguy, Jean (1971). La relation entre la distance spatiale et la distance lexicale. *Revue de Linguistique Romane* 35: 335-357.

Szmrecsanyi, Benedikt (2008). Corpus-based dialectometry: aggregate morphosyntactic variability in British English dialects. *International Journal of Humanities and Arts Computing* 2(1-2): 279-296.

Szmrecsanyi, Benedikt (2009). Typological parameters of intralingual variability: grammatical analyticity vs. syntheticity in varieties of English. *Language Variation and Change* 21(3): 319–353.

Szmrecsanyi, Benedikt (2010a). The English genitive alternation in a cognitive sociolinguistics perspective. In Geeraerts, D., Kristiansen, G. & Peirsman, Y. (Eds), *Advances in Cognitive Sociolinguistics*. Berlin/New York: De Gruyter Mouton. 141-166.

Szmrecsanyi, Benedikt (2010b). *The morphosyntax of BrE dialects in a corpus-based dialectometrical perspective: feature extraction, coding protocols, projections to geography, summary statistics*. Available online: `http://www.freidok.uni-freiburg.de/volltexte/7320/`. University of Freiburg.

Szmrecsanyi, Benedikt (2011). Corpus-based dialectometry: a methodological sketch. *Corpora* 6(1): 45-76.

Szmrecsanyi, Benedikt (to appear). Geography is overrated. In Hansen, S., Schwarz, C., Stoeckle, P. & Streck, T. (Eds.), *Dialectological and folk dialectological concepts of space.* Berlin, New York: Walter de Gruyter.

Szmrecsanyi, Benedikt & Hernández, Nuria (2007). *Manual of Information to accompany the Freiburg Corpus of English Dialects Sampler ("FRED-S").* Available online: `http://www.freidok.uni-freiburg.de/volltexte/2859/`. University of Freiburg.

Szmrecsanyi, Benedikt & Kortmann, Bernd (2009a). The morphosyntax of varieties of English worldwide: a quantitative perspective. *Lingua* 119(11): 1643-1663.

Szmrecsanyi, Benedikt & Kortmann, Bernd (2009b). Vernacular universals and angloversals in a typological perspective. In Filppula, M., Klemola, J. & Paulasto, H. (Eds.), *Vernacular Universals and Language Contacts: Evidence from Varieties of English and Beyond.* London, New York: Routledge. 33-53.

Szmrecsanyi, Benedikt & Kortmann, Bernd (2009c). Between simplification and complexification: non-standard varieties of English around the world. In: Sampson, G., Gil, D. & Trudgill, P. (Eds.), *Language Complexity as an Evolving Variable*. Oxford: Oxford University Press, 64-79.

Szmrecsanyi, Benedikt & Kortmannm Bernd (2011). Typological profiling: learner Englishes versus indigenized L2 varieties of English. In: Mukherjee, J. & Hundt, M. (Eds), *Exploring Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap*. Amsterdam: Benjamins, 167-187.

Szmrecsanyi, Benedikt & Wolk, Christoph (2011). Holistic corpus-based dialectology. *Brazilian Journal of Applied Linguistics/Revista Brasileira de Linguística Aplicada*. Available online.

Tagliamonte, Sali & Smith, Jennifer (2005). *No momentary fancy!* The *zero* 'complementizer' in English dialects. *English Language and Linguistics* 9(2): 289-309.

Trudgill, Peter (1974). Linguistic change and diffusion: description and explanation in sociolinguistic dialect geography. *Language in Society* 2: 215-246.

Viereck, Wolfgang (1986). Dialectal speech areas in England: Orton's lexical evidence. In Kastovsky, D. & Szwedek, A. (Eds.), *Linguistics across Historical and Geographical Boundaries.* Berlin, New York: Mouton de Gruyter. 725-740.

	Pearson correlation coefficient (*r*)	variance explained, in % ($R^2 \times 100$)
as-the-crow-flies distance (linear estimate)	.336	11.3
least-cost travel time (linear estimate)	.344	11.8
Trudgill's linguistic gravity index (logarithmic estimate)	-.379	14.4

Table 1: Correlation coefficients and R^2 values – morphosyntactic distances versus as-the-crow-flies distance, least-cost travel time, and Trudgill's linguistic gravity index (note: all correlation coefficients are significant at $p < .001$)

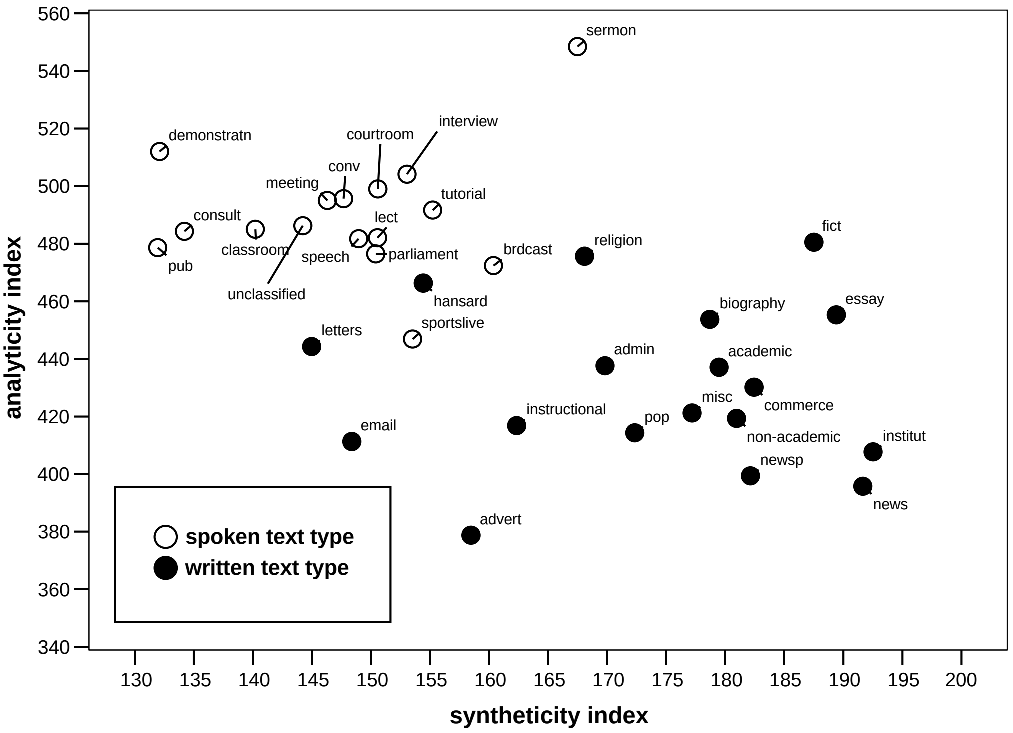


Figure 1: Visualization of the 34 × 2 index matrix: BNC macro registers – analyticity by syntheticity (in index points, *ptw*). Black dots indicate written registers, white dots indicate spoken registers

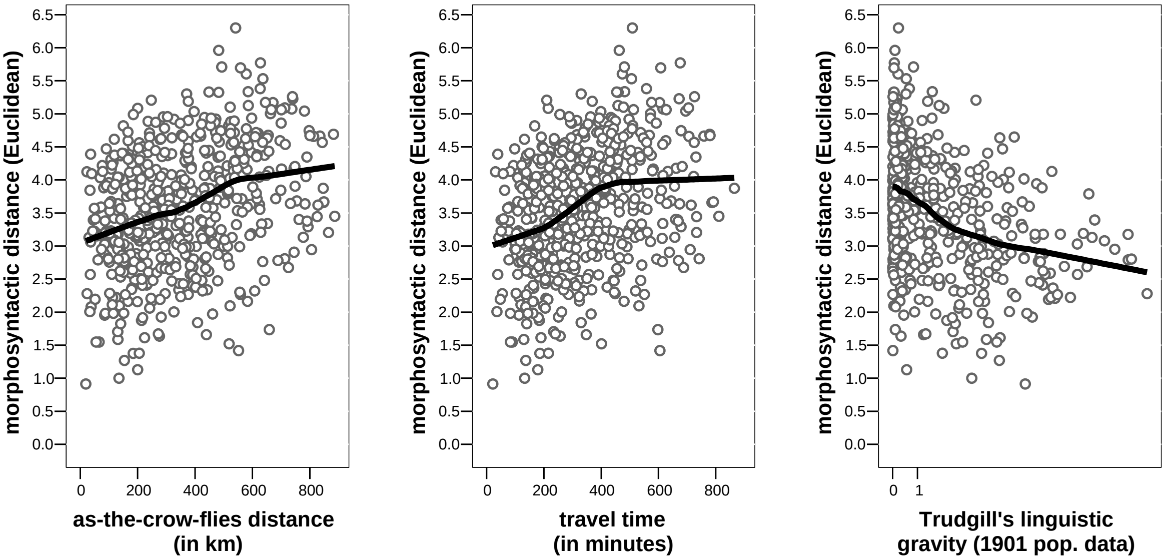


Figure 2: Scatterplots – morphosyntactic distance versus as-the-crow-flies distance (left), least-cost travel time (middle), and Trudgill's linguistic gravity index (right; *log* scale). Solid lines are non-parametric smoothers estimating the overall nature of the relationship

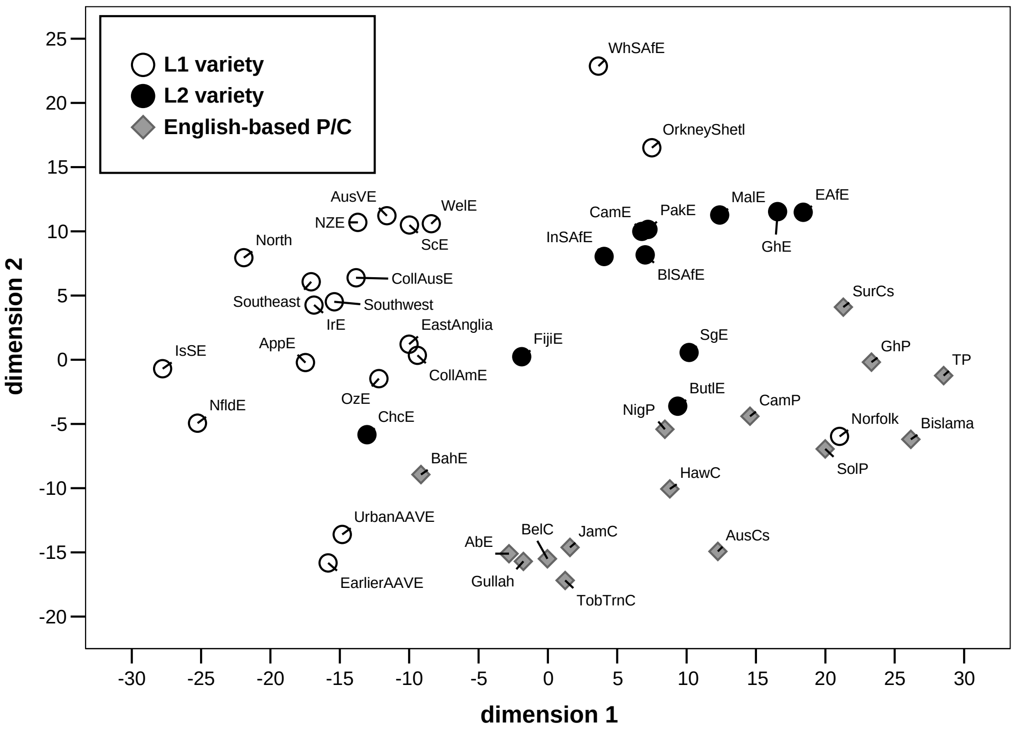


Figure 3: Two-dimensional multidimensional scaling plot – varieties of English worldwide. Input: shared feature classification matrix (squared Euclidean distance). Correlation with original squared Euclidean distances: *r* = .86. White dots indicate L1 varieties, black dots indicate indigenized L2 varieties of English, grey diamonds indicate English-based pidgin and creole languages. Figure 4 below spells out the abbreviations used in the diagram.

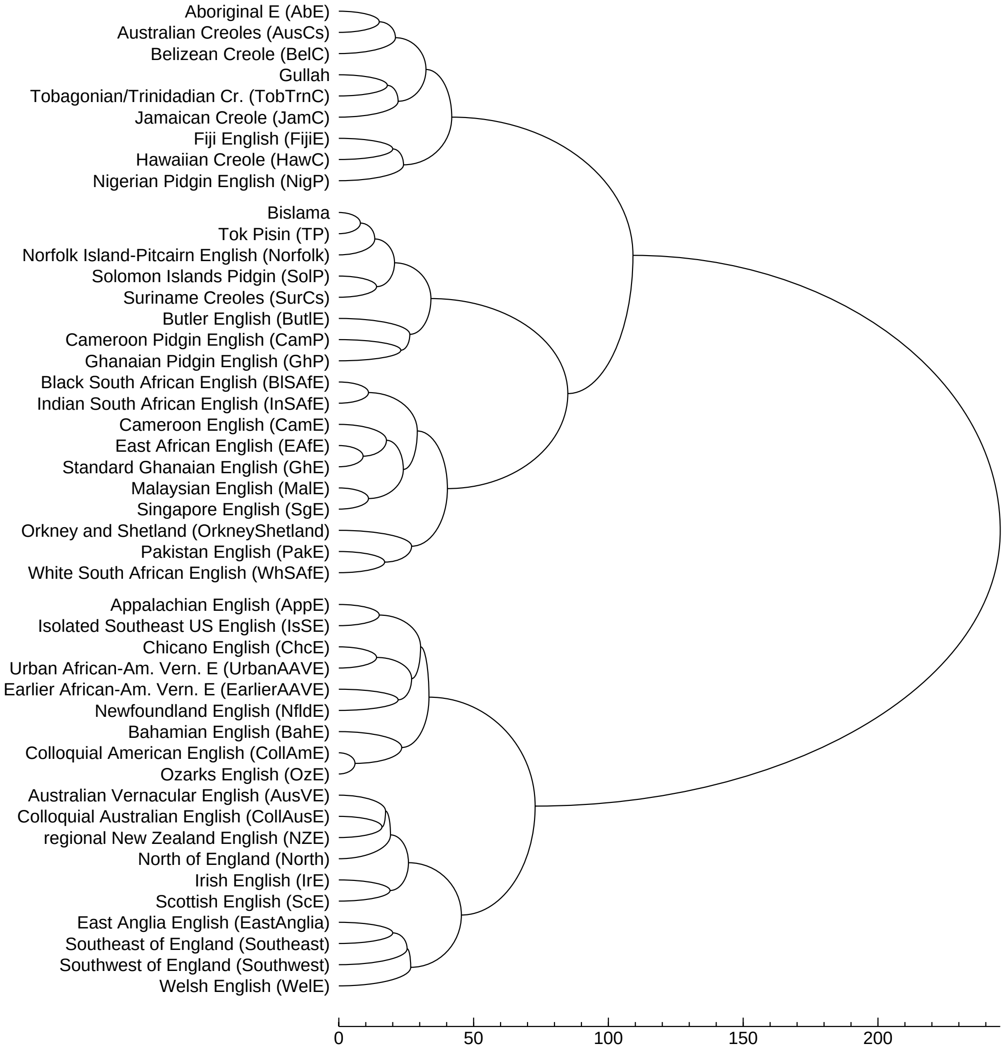
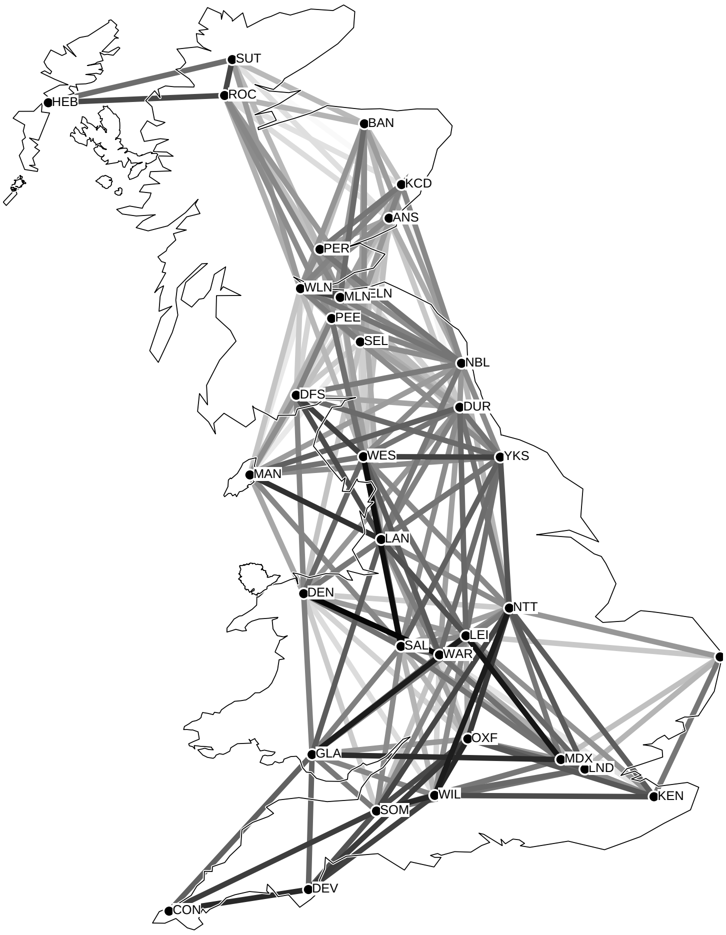


Figure 4: Dendrogram deriving from hierarchical agglomerative cluster analysis – varieties of English world-wide (clustering algorithm: Ward). Input: shared feature classification matrix (squared Euclidean distance).



ANS	Angus
BAN	Banffshire
CON	Cornwall
DEN	Denbighshire
DEV	Devon
DFS	Dumfriesshire
DUR	Durham
ELN	East Lothian
GLA	Glamorganshire
HEB	Hebrides
MAN	Isle of Man
KCD	Kincardineshire
KEN	Kent
LAN	Lancashire
LEI	Leicestershire
LND	London
MDX	Middlesex
MLN	Midlothian
NBL	Northumberland
NTT	Nottinghamshire
OXF	Oxfordshire
PEE	Peebleshire
PER	Perthshire
ROC	Ross and Cromarty
SAL	Shropshire
SEL	Selkirkshire
SFK	Suffolk
SOM	Somerset
SUT	Sutherland
WAR	Warwickshire
WES	Westmorland
WIL	Wiltshire
WLN	West Lothian
YKS	Yorkshire

Map 1. Link map – traditional British English dialects. Morphosyntactically more distant counties are connected by lighter lines, counties that are close morphosyntactically are connected by darker lines (distance limit: 250 km).